Leistungsbeurteilung in der Schule

Herausgegeben von

Kurt Heller

Mit Beiträgen von

Peter Büscher, Mannheim
Walter Fingerhut, Marburg
Anne-Katrin Gaedike, Bonn
Kurt Heller, Bonn
Ralf Horn, Weinheim
Hans-Peter Langfeldt, Heidelberg
Erich Langhorst, Bonn
Horst Nickel, Düsseldorf
Bernhard Rosemann, Bonn
Jürgen Wendeler, Frankfurt/M.
Wilhelm Wieczerkowski, Hamburg

(1974)

Quelle & Meyer Heidelberg

Das in der Umschlaggraphik verwendete
Faktorenmodell der Schulleistung
wurde Kapitel 2.1. in diesem Buch (Seite 42 f.) entnommen.

Printed in Germany.

Satz und Druck: Pilger-Druckerei, GmbH, Speyer.

Umschlagentwurf: Dieter Hoffmann, Heidelberg.

ISBN 3-494-00798-5

Vorwort

Der vorliegende Sammelband mit 12 zum Teil größeren Originalbeiträgen versucht einen Überblick über die wichtigsten Fragen und Probleme schulischer Leistungsbeurteilung zu geben. Dabei werden sowohl Informationen zum gegenwärtigen Stand der Forschung auf diesem Gebiet vermittelt als auch konkrete Hilfen für die praktische Arbeit der Schülerbeurteilung angeboten, so daß Schulpädagogen und Lehrer aller Schularten bzw. Studierende der Erziehungswissenschaft(en), aber auch Psychologen und Soziologen sowie interessierte Laien und in der Bildungspolitik Engagierte angesprochen sind.

Die Einzelbeiträge zum Generalthema „Leistungsbeurteilung" sind in vier große Problembereiche gruppiert, denen der Herausgeber ein Einleitungs- und Übersichtsreferat vorangestellt hat. Die zusammenfassende Darstellung soll dem Leser eine erste Orientierung über die verschiedenen Themenkomplexe erlauben, nicht jedoch die Lektüre der Originalarbeiten dieses Readers ersetzen. Weiterhin dienen die jedem Hauptkapitel beigefügten Kurzkommentare sowie ein ausführliches Personen- und Sachregister am Ende des Bandes der schnelleren und gründlichen Einarbeitung in die einschlägigen Begriffe, Probleme und methodischen Möglichkeiten schulischer Leistungsbeurteilung.

Die beiden ersten Originalbeiträge sind dem Beurteilungsgegenstand gewidmet. Während zunächst eine operationale Bestimmung dessen, was man gemeinhin als *Schulleistung* bezeichnet, vorgeschlagen wird, informiert das anschließende Sammelreferat ausführlich über die intellektuellen und außerintellektuellen Determinanten der Schulleistung. Die Kenntnis dieser Zusammenhänge ist gleichermaßen im Hinblick auf die Förderung und Bewertung von Schülerleistungen bedeutsam.

Die zentrale Thematik des Buches findet ihren Niederschlag in den folgenden Kapiteln über *testtheoretische* Fragen sowie *objektive* und *subjektive Verfahren* der Schülerbeurteilung. Eingehend werden die klassische Testtheorie wie auch moderne testtheoretische Ansätze dargestellt und auf ihre Relevanz für die Leistungsbeurteilung hin kritisch untersucht. Weitere Beiträge beschäftigen sich mit Definitionsproblemen von Schultests und einschlägigen Fragen der Lernzieldefinition. Besonderes Gewicht wird im folgenden auf die Beschreibung der *objektiven* Verfahren zur Leistungsbeurteilung, der sog. standardisierten (formellen) und kriteriums- oder lernzielbezogenen (informellen) Schulleistungstests, gelegt. Erklärtes Ziel der beteiligten Autoren ist es, dem Lehrer nicht nur brauchbare Hinweise für die Anwendung und eventuelle Konstruktion solcher Verfahren zu geben; die mit zahlreichen Beispielen versehenen Anleitungen sollen ihm zugleich die praktische Arbeit in der Schule erleichtern, indem sie ihn zur objektiven,

zuverlässigen und gültigen — somit vielleicht gerechteren — Urteilsfindung befähigen.

Aus dem gleichen Anliegen heraus wird den *subjektiven* Verfahren der Schülerbeurteilung durch Lehrer eine breite Diskussion gewidmet. Neben methodischen Problemen der Beobachtung und Beurteilung des Schülerverhaltens im Unterricht kommen hier u. a. neue varianzanalytische Befunde zur Notengebung sowie umfangreiche experimentelle Untersuchungsergebnisse über die verschiedenen Einflüsse der Aufsatzbeurteilung zur Sprache. Im letzten Beitrag werden schließlich Wege zur Vereinheitlichung der Zensierung von Schüleraufsätzen aufgezeigt — ein Anliegen, das besonders in der Praxis stehende Schulpädagogen interessieren dürfte.

Dieser thematisch weitgesteckte Rahmen in *einem* Band wäre nicht zu realisieren gewesen ohne die bereitwillige Mitarbeit aller angesprochenen Autoren, denen der Herausgeber darüberhinaus wertvolle Anregungen verdankt. Dank schulde ich ferner den Autoren des Beitrags zu Kap. 2.1. für die freundliche Genehmigung, das dort abgebildete Faktorenmodell der Schulleistung als Einbandgrafik zu verwenden, meinen wissenschaftlichen Mitarbeitern Dipl.-Psych. Renate BONN, Dipl.-Psych. Anne-Katrin GAEDIKE und Dipl.-Psych. Manfred SCHNEIDER für die Erstellung der Register und ihre Mithilfe beim Korrekturlesen sowie dem Quelle & Meyer Verlag für die ansprechende Gestaltung und verlegerische Besorgung des Buches.

Bonn, im September 1973 K. H.

Inhaltsverzeichnis

1.	**Einleitung und Übersichtsreferat**	13
1.1.	Gegenstand schulischer Leistungsbeurteilungen	14
1.1.1.	Zur Problematik des Leistungsbegriffs in der Pädagogik	14
1.1.2.	Hauptdimensionen der Schulleistung	15
1.1.3.	Bedingungskomponenten der Schulleistung	16
1.2.	Aufgaben und Ziele schulischer Leistungsbeurteilung	18
1.2.1.	Leistungsbeurteilung im Dienste der Unterrichtsorganisation und Bildungsreform	18
1.2.2.	Leistungsbeurteilung als Funktion individueller Beratung	19
1.2.3.	Leistungsbeurteilung als Funktion der Schullaufbahn- bzw. Systemberatung	20
1.3.	Formen und Methoden der Leistungsbeurteilung im Bildungswesen	21
1.3.1.	Testtheoretische Grundlagen	21
1.3.1.1.	Die klassische Testtheorie und ihre Kritik	21
1.3.1.2.	Neuere Modellansätze und ihre Problematik	25
1.3.1.3.	Vorschläge zur Klassifikation von Schultests	26
1.3.2.	Objektive Verfahren schulischer Leistungsbeurteilung	27
1.3.2.1.	Lernzieldefinition als Voraussetzung der Leistungsmessung	27
1.3.2.2.	Informelle Tests (Lernkontrolltests)	28
1.3.2.3.	Standardisierte Schulleistungstests	29
1.3.3.	Subjektive Verfahren schulischer Leistungsbeurteilung	30
1.3.3.1.	Verhaltensbeobachtung und Schülerbeurteilung	30
1.3.3.2.	Leistungsbeurteilung durch Notengebung	31
1.3.3.3.	Probleme der Aufsatzbeurteilung	33
1.4.	Ausblick	34
1.5.	Literaturverzeichnis	35
2.	**Schulleistung als pädagogisch-psychologisches Problem**	37
2.1.	Empirische Ansätze zur Aufklärung des Konstruktes „Schulleistung" *(H. P. Langfeldt* und *W. Fingerhut)*	38
2.2.	Determinanten der Schulleistung *(A.-K. Gaedike)*	46
2.2.1.	Kognitive Faktoren der Schulleistung	47
2.2.1.1.	Korrelative Beziehungen zwischen verschiedenen Intelligenztests und der Schulleistung	47
2.2.1.2.	Korrelative Beziehungen zwischen verschiedenen Intelligenzfaktoren und der Schulleistung	54
2.2.2.	Nicht-kognitive Faktoren der Schulleistung	60
2.2.2.1.	Motivation und Arbeitshaltung	60
2.2.2.2.	Persönlichkeitsvariablen	65
2.2.2.2.1.	Angst	66
2.2.2.2.2.	Selbstachtung	69
2.2.2.2.3.	Extraversion	70

2.2.2.3. Lehrerverhalten 72
2.2.2.3.1. Setzung „sachfremder" Leistungsmotivation 72
2.2.2.3.1. Kognitive Stile 74
2.2.2.3.3. Unterrichtsatmosphäre 75
2.2.2.3.4. Direktives Verhalten 77
2.2.2.3.5. Werthaltungen 79
2.2.2.4. Sozio-kulturelles Milieu (Familie, Peergroups) 80
2.2.2.4.1. Äußere, objektive Bedingungen 81
2.2.2.4.2. Werthaltungen, Normen 82
2.2.3. Schlußbemerkungen 86
2.2.4. Literaturverzeichnis 87

3. Testtheoretische Ansätze der Schulleistungsmessung . . . 94

3.1. Die klassische Testtheorie als Grundlage standardisierter Schulleistungstests *(H.-P. Langfeldt)* 96

3.1.0. Vorbemerkung 96
3.1.1. Messen und Testen 97
3.1.1.1. Definition von Messen 97
3.1.1.2. Skalen für Meßwerte 98
3.1.1.3. Gütekriterien von Messungen 101
3.1.1.4. Zusammenfassung 101
3.1.2. Grundlagen der klassischen Testtheorie 102
3.1.2.1. Fragestellung 102
3.1.2.2. Axiome der klassischen Testtheorie 102
3.1.2.3. Das Reliabilitätskonzept 104
3.1.2.4. Validitätskonzepte 105
3.1.2.5. Objektivitätsarten 107
3.1.2.6. Zusammenhänge zwischen den Gütekriterien 108
3.1.2.7. Zusammenfassung 109
3.1.3. Reliabilität von Schultests 109
3.1.3.1. Retest-Reliabilität 109
3.1.3.2. Paralleltest-Reliabilität 110
3.1.3.3. Halbierungs-Reliabilität 111
3.1.3.4. Die Konsistenz-Reliabilität 111
3.1.3.5. Das Konzept des Standardmeßfehlers 112
3.1.3.6. Zusammenfassung 113
3.1.4. Validität von Schultests 114
3.1.4.1. Curriculare Validität 114
3.1.4.2. Kriterienbezogene Validität 115
3.1.4.3. Konstruktvalidität 116
3.1.4.4. Zusammenfassung 117
3.1.5. Objektivität von Schultests 117
3.1.5.1. Durchführungs- und Auswertungsobjektivität 117
3.1.5.2. Interpretationsobjektivität 118
3.1.5.3. Normierung von Testergebnissen 119
3.1.5.4. Kriterien zur Anwendung unterschiedlicher Normen 122
3.1.5.5. Zusammenfassung 122
3.1.6. Formaler Aufbau von Schultests 123
3.1.6.1. Aufgaben und Aufgabenanalyse 123
3.1.6.2. Untertests und Testbatterien 125

3.1.6.3. Ein praktisches Beispiel: Der AST 4 125
3.1.6.4. Abschließende Definition eines Schultests 128
3.1.7. Diskussion 129
3.1.7.1. Kritik an der klassischen Testtheorie 129
3.1.7.2. Kritik an der Anwendung klassischer Tests 131
3.1.8. Zusammenfassung 134
3.1.9. Literaturverzeichnis 135

3.2. Einige testtheoretische Aspekte kriterienbezogener Leistungsmessung *(P. Büscher)* 137
3.2.1. Terminologische Probleme 137
3.2.1.1. Lernzielorientierte und standardisierte Tests 137
3.2.1.2. Lernzielorientierte und lehrzielorientierte Tests 137
3.2.1.3. Normbezogene und kriterienbezogene Tests 138
3.2.1.4. Kriterien 139
3.2.2. Testtheoretische Probleme 141
3.2.2.1. Unzulänglichkeit der klassischen Testtheorie bei kriterienbezogenen Tests 142
3.2.2.1.1. Analyse der Aufgabenschwierigkeit 142
3.2.2.1.2. Analyse der Aufgabentrennschärfe 143
3.2.2.1.3. Analyse der Aufgabengültigkeit 144
3.2.2.2. Verfahren zur Test- bzw. Itemanalyse kriterienbezogener Meßinstrumente 144
3.2.2.2.1. Der Vortest-Nachtest-Differenzindex D_{pp} nach Cox und Vargas 145
3.2.2.2.2. Die Itemanalyse von Popham 146
3.2.2.2.3. Der U-Koeffizient nach Fricke 148
3.2.2.2.4. Das Rasch-Modell 149
3.2.2.2.5. Reliabilitätsbestimmung nach Carver 151
3.2.2.2.6. Reliabilitätsbestimmung nach Livingston 151
3.2.2.2.7. Reliabilitätsschätzung durch Skalogrammanalyse 152
3.2.2.2.8. Reliabilitätsschätzung nach Cronbach 152
3.2.2.2.9. Reliabilitätsbestimmung nach Jackson 153
3.2.2.2.10. Reliabilitätsbestimmung nach Fricke 153
3.2.2.2.11. Klassische Reliabilitätsschätzung 153
3.2.2.2.12. Validitätsbestimmung 153
3.2.3. Zusammenfassung 155
3.2.4. Literaturverzeichnis 156

3.3. Zur Problematik der Klassifikation von Schultests *(B. Rosemann)* 158
3.3.1. Untersuchung der bisherigen Terminologie 158
3.3.1.1. Standardisierte und nichtstandardisierte Tests 158
3.3.1.2. Der Begriff des Kriteriums 160
3.3.1.3. Normbezogene versus lernzielorientierte Tests 161
3.3.1.4. Normbezogene versus kriteriumsbezogene Tests 161
3.3.2. Gedanken zu einer pädagogisch begründeten Klassifikation der Schultests 163
3.3.2.1. Die Unterscheidung zwischen Leistungsfeststellung und Leistungsbewertung 163
3.3.2.2. Lernsteuerungstests und Lernkontrolltests 165
3.3.3. Literaturverzeichnis 166

4.	**Objektive Verfahren der Leistungsbeurteilung in der Schule**	167
4.1.	Leistungsmessung und Lernzieldefinition *(R. Horn)*	169
4.1.1.	Voraussetzungen der Leistungsmessung	169
4.1.2.	Operationalisierte Lernziele	170
4.1.3.	Erstellung von Prüfungsaufgaben	176
4.1.4.	Zusammenfassung	180
4.1.5.	Literaturverzeichnis	181
4.2.	Konstruktion und Einsatz von Informellen Tests zur Leistungsbeurteilung (Lernkontrolltests) *(B. Rosemann)*	182
4.2.1.	Einleitung	182
4.2.2.	Die Formulierung der Items	183
4.2.2.1.	Die Spezifikationstabelle	183
4.2.2.2.	Itemtypen und ihre Konstruktion	186
4.2.2.2.1.	Aufgaben mit gebundenen Antworten	187
	Auswahlantworten	187
	Ordnungsantworten	194
4.2.2.2.2.	Aufgaben mit nicht-gebundenen Antworten	196
	Ergänzungs- bzw. Kurzantworten	196
	Kurzaufsatzantwort (Essay-Test)	198
4.2.2.2.3.	Aufgaben zur Messung komplexer Leistungen	198
	Interpretationsübungen	199
4.2.3.	Entwicklung der Test„vorform“	201
4.2.3.1.	Gruppierung der Items und Erstellung des Testheftes	201
4.2.3.2.	Aufgabenbewertung und Ermittlung der Gesamtleistung	205
4.2.4.	Die Itemanalyse	205
4.2.4.1.	Berechnung des Schwierigkeitsgrades einer Aufgabe	206
4.2.4.2.	Berechnung der Trennschärfe einer Aufgabe	206
4.2.4.3.	Distraktorenanalyse	208
4.2.5.	Itemselektion und -revision	209
4.2.6.	Die Reliabilität des Tests	212
4.2.7.	Die Validität des Tests	215
4.2.8.	Die Normierung	215
4.2.9.	Literaturverzeichnis	220
4.3.	Einsatz standardisierter Schulleistungstests *(R. Horn)*	222
4.3.0.	Vorbemerkung	222
4.3.1.	Lesetests	223
4.3.2.	Rechtschreibtests	224
4.3.3.	Rechentests	225
4.3.4.	Allgemeine Schulleistungstests	225
4.3.5.	Fremdsprachentests	226
4.3.6.	Tests für verschiedene Fächer	
4.3.7.	Übersicht über die zur Zeit verfügbaren Tests nach Testkategorien und Klassenstufen	228 229
5.	**Subjektive Verfahren der Leistungsbeurteilung in der Schule**	230
5.1.	Beobachtung und Beurteilung des Schülerverhaltens im Unterricht *(E. Langhorst)*	233

5.1.1. Beobachtungsnotwendigkeit und Beurteilungspraxis 233
5.1.2. Beobachtung 235
5.1.3. Beschreibung 237
5.1.4. Beurteilung 241
5.1.5. Beurteilungsfehler 246
5.1.6. Literaturverzeichnis 251

5.2. Leistungsbeurteilung durch Notengebung (*W. Fingerhut* und *H.-P. Langfeldt*) 253

5.2.1. Schulnoten als Urteile 253
5.2.2. Schulnoten als Meßwerte 254
5.2.2.1. Verteilungsform von Schulnoten 255
5.2.2.2. Objektivität von Schulnoten 257
5.2.2.3. Reliabilität von Schulnoten 259
5.2.2.4. Validität von Schulnoten 260
5.2.3. Schulnoten als Variablen 262
5.2.3.1. Methodische Fragen 262
5.2.3.2. Schulnoten und Intelligenz 263
5.2.3.3. Schulnoten und soziale Herkunft 264
5.2.3.4. Schulnoten und Persönlichkeit 265
5.2.4. Zusammenfassung 266
5.2.5. Literaturverzeichnis 267

5.3. Einflüsse auf die Beurteilung von Schüleraufsätzen — Ergebnisse einer quasi-experimentellen Versuchsreihe (*E. Nickel* und *W. Wieczerkowski*) 271

5.3.1. Einleitung und Problemstellung 271
5.3.2. Die erste Untersuchung 272
5.3.2.1. Ausgangsproblem und Fragestellungen 272
5.3.2.2. Versuchsablauf 273
5.3.2.2.1. Das Beurteilungsmaterial 273
5.3.2.2.2. Der Versuchsplan 273
5.3.2.2.3. Beurteiler und Bewertungsvorgang 275
5.3.2.3. Ergebnisse 276
5.3.2.3.1. Beurteilerübereinstimmung und Geschlechtsunterschiede 276
5.3.2.3.2. Verteilungsform und Streuung der Beurteilungen 277
5.3.2.3.3. Der Einfluß von Informationsart und Vergleichsreizen auf die Urteile 277
5.3.2.3.4. Der Einfluß verschiedener Sprachmerkmale 280
5.3.3. Die zweite Untersuchung 285
5.3.3.1. Fragestellung 285
5.3.3.2. Versuchsablauf 286
5.3.3.2.1. Das Beurteilungsmaterial 286
5.3.3.2.2. Die Beurteiler und ihre experimentelle Beeinflussung 286
5.3.3.2.3. Der Versuchsplan 287
5.3.3.3. Ergebnisse 287
5.3.3.3.1. Globalanalyse und Reduktion des Versuchsplans 287
5.3.3.3.2. Effekte der Informationsart auf die Aufsatzbewertung 289
5.3.3.3.3. Einflüsse der Stichprobenzugehörigkeit 291
5.3.3.3.4. Wechselwirkungen von Informationsart und Erfahrungshintergrund der Beurteiler 292

5.3.4. Die dritte Untersuchung 293
5.3.4.1. Fragestellung 293
5.3.4.2. Versuchsablauf 294
5.3.4.2.1. Das Beurteilungsmaterial 294
5.3.4.2.2. Beurteilungsstichprobe und Bewertungsablauf 294
3.5.4.3. Ergebnisse 294
5.3.4.3.1. Der Einfluß der schulpraktischen Erfahrung 294
5.3.4.3.2. Der Einfluß verschiedener Sprachkriterien 296
5.3.5. Diskussion und Ergebnisse 297
5.3.5.1. Allgemeine Befunde zur experimentellen Situation 297
5.3.5.2. Diskussion der ersten Untersuchung 299
5.3.5.2.1. Zur Beeinflussung der Beurteilung durch Vergleichsserien . . . 299
5.3.5.2.2. Zum Einfluß unterschiedlicher Information über die Ausgangssituation 299
5.3.5.2.3. Zum Einfluß verschiedener Sprachvariablen 300
5.3.5.3. Diskussion der zweiten Untersuchung 301
5.3.5.3.1. Zum Einfluß unterschiedlicher Informationen über die Leistungen der Schüler 301
5.3.5.3.2. Die Beurteilung durch Schüler, Studenten und Junglehrer . . . 302
5.3.5.3.3. Geschlechtsspezifische Einflüsse 303
5.3.5.4. Diskussion der dritten Untersuchung 303
5.3.5.4.1. Zum Einfluß unterschiedlicher Information über das Leistungsverhalten der Schüler 303
5.3.5.4.2. Zur unterschiedlichen Beurteilung durch Referendare und Lehrer 303
5.3.5.4.3. Zum Einfluß verschiedener Sprachkriterien 304
5.3.6. Zusammenfassung und Schlußfolgerungen 304
5.3.7. Literaturverzeichnis 306

5.4. Bemühungen um Vereinheitlichung der Aufsatzbeurteilung *(J. Wendeler)* 309
5.4.1. Die mangelnde Übereinstimmung von Aufsatzbeurteilungen . . 309
5.4.2. Leistungsermittlung und Zensierung 312
5.4.3. Kriterien der Aufsatzbeurteilung 313
5.4.4. Erfassung der Urteilsobjektivität 316
5.4.5. Versuche zur Erhöhung der Urteilsobjektivität 317
5.4.6. Literaturverzeichnis 322

6. Autorenverzeichnis 324

7. Personenregister 327

8. Sachregister 331

1. Einleitung und Übersichtsreferat

Die schulische Leistungsbeurteilung steckt voller Probleme, wie die Ausführungen in diesem Buch noch verdeutlichen werden. Dies gilt sowohl hinsichtlich der theoretischen Grundlegung als auch im Hinblick auf das Verfahrensangebot (Beurteilungsmethoden). Sofern man Leistungsforderung und Leistungsbewertung als berechtigte Anliegen der Schule ansieht, womit jedoch keiner Verabsolutierung des Leistungsprinzips das Wort geredet wird, muß man sich auch den Problemen, die sich hierbei ergeben, stellen. Andernfalls praktiziert man ‚Vogel-Strauß-Politik' und huldigt bekannten Grundsätzen wie „Was ich nicht weiß, macht mich nicht heiß!" oder „Wasch' mir den Pelz, aber mach' mich nicht naß!", was im Grunde — mit Blick auf die Hauptbetroffenen: Schüler und Lehrer (einschließlich Eltern) — das Dilemma nur noch verschärft. Aber selbst, wenn man Modeparolen folgend glaubt, dem Problem der Leistung in unserer Gesellschaft überhaupt aus dem Wege gehen zu können, begibt man sich meines Erachtens nur in neue Schwierigkeiten. Das Terrain, auf dem entsprechende Konflikte dann ausgetragen werden, mag zwar wechseln, etwa von der Leistungsebene auf das Feld ideologischer Auseinandersetzung (statt Leistungserweise werden dann ‚Glaubens'-Bekenntnisse, ‚Überzeugungs'-Taten und ähnliche Merkmale die Kriterien abgeben), der Problematik selbst, d. h. dem Individuum und der Gesellschaft gleichermaßen Rechnung tragende Verpflichtungen zu postulieren und daraus resultierende Urteilskriterien zu finden, wird man kaum aus dem Wege gehen können. Siehe ergänzend dazu z. B. die Überlegungen von HENTIGS (1970, S. 140 ff.).

Somit wäre unsere Position abgesteckt: Wir erblicken in der schulischen Forderung nach Lernleistungen eine legitime Forderung. Diese stellt freilich nicht das einzige und vielleicht nicht einmal das wichtigste, gleichwohl ein unentbehrliches, Bildungsziel dar. Erkennt man diese Aufgabenfunktion — neben anderen — an, so muß man auch die bei der Verwirklichung dieses Auftrags in Erscheinung tretenden Probleme artikulieren und wissenschaftlich begründete Wege (Methoden) zu ihrer Lösung aufzeigen. Zweifellos gilt ein solches Postulat zunächst für die pädagogische Funktion der Aktivierung von Schülerleistungen, erst danach erhält es seine volle Berechtigung im Hinblick auf die verschiedenen Ziele der Leistungs*beurteilung* (z. B. im Rahmen unterrichtlicher Differenzierungsansätze, schulischer Effizienzkontrollen, der Schullaufbahnberatung usw.), die ja letzten Endes wiederum im Kontext schulischer Bildungsbemühungen zu sehen sind.

Die Beurteilung des Schülerverhaltens unter dem Gesichtspunkt der Lernleistung erweist sich somit als ein Problem, das nicht nur die Schule, sondern auch die Familie, die engagierten Wissenschaftler und Bildungspolitiker

und noch viele andere angeht. Aktualität und Virulenz dieser Problematik wurden jüngst beim Streit um die Zulassungsbestimmungen und Kriterien für die Auswahl der Studienbewerber besonders gefragter Studienfächer erneut offenbar. Bedeutet dies nicht eine Herausforderung an uns alle?

1.1. Gegenstand schulischer Leistungsbeurteilungen

1.1.1. Zur Problematik des Leistungsbegriffs in der Pädagogik

Mit dem Begriff der *Schulleistung* sei hier das gesamte Leistungsverhalten, soweit es im Kontext schulischer Bildungsbemühungen virulent wird, angesprochen. Dabei verdienen der dynamische Aspekt (Lernprozeß) und der statische Aspekt (Lernprodukt) gleichermaßen Beachtung — entgegen manchen Begriffsvorstellungen, die nur das Ergebnis der Schülerleistung und nicht auch deren Bedingungsgefüge im Auge haben. Diesen mehr oder weniger operational definierbaren Kategorien können schließlich ethische (Leistungspflicht) und ökonomische Überlegungen (Leistungsnotwendigkeit) — als Sinndimensionen menschlicher Leistung überhaupt — hinzugefügt werden.

Daraus resultiert nach Furck (1964, S. 118 ff.) eine doppelte Aufgabenfunktion der Schulleistung, nämlich die Persönlichkeitsbildung des Schülers und die Sicherung des volkswirtschaftlichen Leistungspotentials. „Das Problem der Leistung in der Schule spitzt sich so zu der Frage nach dem rechten Verhältnis von individueller Bildsamkeit und ihr angemessener Anforderung zu" (loc. cit.). Damit ist im weiteren Sinne das pädagogisch-didaktische Anliegen der Leistungsförderung angesprochen. Sicherlich lassen sich noch andere Aspekte und möglicherweise ganz neue Perspektiven zum Schulleistungsproblem aufweisen, die bislang artikulierten, realisierbaren Ansätze (z. B. von Furck, v. Hentig oder Gaude & Teschner u. a.) kulminieren letzten Endes doch immer wieder um die hier skizzierten Problembereiche. Im Hinblick auf die zentrale Thematik dieses Buches erübrigt sich vorläufig eine weiterführende Diskussion; die eine oder andere Grundsatzfrage wird im Rahmen der einzelnen Beiträge erneut aufgegriffen und im entsprechenden Kontext zu diskutieren sein.

Unsere Definitionsformel „Schulleistung meint das gesamte Leistungsverhalten im Kontext der Schule" schließt prinzipiell das Schüler- (bzw. Eltern-) *und* Lehrerverhalten ein, wenngleich die Vokabel ‚Schulleistung' häufig nur die Leistung des Schülers indiziert. Ein solchermaßen eingeengter Leistungsbegriff ist jedoch nur akzentuierend, d. h. im Hinblick auf jeweils thematisierte Fragestellungen, angebracht. Die Gefahr, hierbei den Gesamtkomplex interdependenter Zusammenhänge aus dem Auge zu verlieren, ist freilich nicht von der Hand zu weisen. Nicht zufällig haben wir den Beitrag von A.-K. Gaedike zum Thema *Determinanten* der Schulleistung nebst

dem Diskussionsbeitrag von LANGFELDT & FINGERHUT zur *Faktorenstruktur* der Schulleistung an den Anfang gestellt. Möglichen Mißverständnissen sollte hiermit von vornherein begegnet werden. Wir müssen davon ausgehen, daß Schulleistungen mehrdimensional bedingt sind und in hohem Maße von persönlichkeits- und sozialpsychologischen Faktoren abhängige Variablen(bündel) darstellen. Vor diesem Hintergrund sind praktisch alle Beiträge dieses Sammelbandes zu sehen, auch dann, wenn im einen oder anderen Fall diese Implikationen nicht expressis verbis zum Ausdruck kommen.

1.1.2. Hauptdimensionen der Schulleistung

Ausgehend von den üblichen Indikatoren der Schulleistung versuchen LANGFELDT & FINGERHUT in ihrem Beitrag eine empirische Beschreibung dieses Phänomens. Die zu diesem Zweck faktorenanalysierten Schulzeugnisse teils eigener, teils fremder Provenienz deuten übereinstimmend an, „daß die Schulleistung — so wie sie durch Schulnoten erfaßt wird — bei weitem nicht so differenziert ist, wie es Zeugnisse mit zehn und mehr Einzelnoten nahelegen". Für die Schülerleistungen auf der Sekundarstufe konnten lediglich drei Faktoren nachgewiesen werden: ein Fremdsprachenfaktor, ein mathematisch-naturwissenschaftlicher Faktor und ein sachkundlicher Faktor (sensu DENIG & WEIS 1970). Für die Primarstufe (Grundschüler und lernbehinderte Sonderschüler) ergaben sich sogar nur zwei Faktoren: ein allgemeiner Schulleistungsfaktor (vor allem auf die Fachzensuren in Deutsch und Rechnen bezogen) und ein Faktor der schulischen Disziplin (mit entsprechenden Ladungen auf den sog. Kopfnoten). Die Ergebnisse von FUNKE (1972) und ZIMMERMANN (1968) bestätigen — unabhängig voneinander — diese gegenüber dem wesentlich differenzierteren Zeugnisbild stark reduzierte Faktorenstruktur der tatsächlichen Schülerleistung. Im Hinblick auf die Schulleistungsbeurteilung herkömmlicher Art (Zensurierung) ergäbe sich hieraus die Konsequenz, künftig nicht mehr als zwei oder drei Schulnoten zu vergeben. Zumindest unter dem Gesichtspunkt streng leistungsbezogener Beurteilung sind mehr als drei Einzelnoten irreführend und überflüssig zugleich.

Sofern man neben den Lehrerurteilen in Form von Zensuren noch weitere Daten, z. B. Schulleistungstests, Intelligenztests, biographische Informationen usw., in die Faktorenanalyse miteinbezieht, gewinnt man ein höher strukturiertes Modell der Schulleistung. Wie FINGERHUT & LANGFELDT (1971) unter Bezug auf ein entsprechendes Datenmaterial von 1756 Viertklaßkindern nachweisen konnten, ergeben sich hierbei folgende vier Faktoren: 1. „Schulische Leistungsfähigkeit" mit Ladungen auf den Zensuren der Kernfächer (Deutsch, Rechnen, Heimatkunde), sämtlichen Skalen des AST 4 (Allgemeiner Schulleistungstest von FIPPINGER) und den Untertests

3 + 4 des LPS (Leistungsprüfsystem von W. HORN); 2. „Genereller Notenfaktor" mit Ladungen auf den Fachleistungs- und Kopfnoten sowie den Variablen Alter und Geschlecht; 3. „Test-Intelligenz" mit Ladungen auf sämtlichen Skalen des LPS; 4. „Rechenleistungen" mit Ladungen auf der Rechennote, den Rechenskalen des AST 4 und den LPS-Subtests 3 + 4, 14 und 15 (Arbeitsprobe).

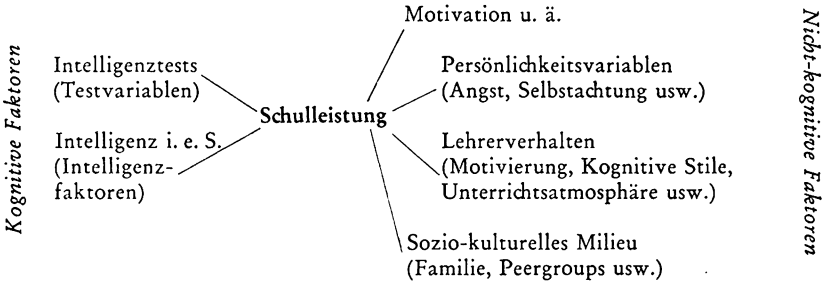
Diese Ergebnisse lassen einen gewissen Zusammenhang zwischen einigen Intelligenzfaktoren (Reasoning, Accuracy, Number) und der (Grund-) Schulleistung, besonders hinsichtlich des allgemeinen Leistungsstandards und der Rechenleistung, erkennen. Die gerade in der jüngeren Literatur häufig vertretene Auffassung, wonach zwischen Intelligenz und Schulleistung keine oder nur schwache Zusammenhänge bestehen, muß demnach korrigiert bzw. folgendermaßen präzisiert werden. „Es kann ... nur gesagt werden, daß Schul*noten* und Intelligenz relativ unabhängig voneinander sind", nicht aber Schul*leistung* und Intelligenz. Da Schulnoten allein nach den Befunden von LANGFELDT & FINGERHUT kein adäquates Abbild der Schulleistung geben, können diese „nur in mäßigem Umfange etwas über die *tatsächlichen* Bedingungen der Schulleistung aussagen". Dieses Ergebnis erhellt indirekt die Bedeutung kognitiver Faktoren im Hinblick auf die Konstituierung von Schulleistungen, ohne hiermit den Einfluß nicht-kognitiver Determinanten als gering einzuschätzen.

Eine modifizierte Faktorenanalyse der Daten von FINGERHUT & LANGFELDT ergab schließlich fünf abgrenzbare Cluster: 1. Intelligenztestleistungen (LPS), 2. Schultestleistungen (AST 4), 3. Fachleistungsnoten (Deutsch, Rechnen, Heimatkunde), 4. Kopfnoten (Betragen, Fleiß, Aufmerksamkeit, Ordnung), 5. Merkmale biographischer Art (Alter, Geschlecht, soziale Herkunft). Aus der Distanz der Cluster zueinander postulieren die Autoren jetzt ein zweidimensionales Konstrukt „Schulleistung", demzufolge sich Schulleistungen im gegenwärtigen Bildungssystem als Funktion von „tatsächlicher Leistung aufgrund von Begabung" (Faktor I) und „Anpassung an dieses System" (Faktor II) erweisen. Siehe dazu Abb. 1 auf Seite 42 ff. Zur Stützung ihres Interpretationsmodells, das die Autoren als — vorläufigen — Diskussionsbeitrag gewertet wissen wollen, können sie auf faktorenanalytische Untersuchungsbefunde von SEITZ und LÖSER aus den Jahren 1969 bis 1971 verweisen.

1.1.3. Bedingungskomponenten der Schulleistung

Der folgende Beitrag von A.-K. GAEDIKE beschäftigt sich nun mit den einzelnen Determinanten der Schulleistung, wozu die Autorin die wichtigsten Untersuchungen der letzten zwei Dezennien gesammelt hat. Nachstehendes Schema, das wir in Anlehnung an eine frühere Publikation (vgl. HELLER

1970, S. 67) bringen, möge dem Leser die Orientierung durch den Dschungel einer kaum mehr zu sichtenden Zahl von Einzelbefunden zu diesem Themenkomplex erleichtern. Zugleich sind hiermit die Hauptpunkte des Sammelreferats skizziert.



Wenn im folgenden von *Determinanten* die Rede ist, so gilt es zu beachten, daß sich dieser Begriff auf *korrelative* Beziehungen beschränkt, d. h. nicht ohne weiteres etwas über Ursache-Wirkungs-Zusammenhänge aussagt. Streng genommen kann bei Beziehungen dieser Art nicht exakt „zwischen ‚verursachenden' Determinanten und ‚enthaltenen' Faktoren" unterschieden werden. Während der zweite Aspekt bereits Gegenstand der vorhergehenden Betrachtungen war, sollen im Mittelpunkt des GAEDIKEschen Beitrags vorrangig die *Bedingungs*komponenten (erster Bedeutungsaspekt von ‚Determinante') der Schulleistung untersucht werden. Für die Interpretation der hier einschlägigen Befunde gelten freilich die angeführten (methodologischen) Kautelen ungemindert.

Zunächst werden die *kognitiven* Determinanten der Schulleistung untersucht, also die Frage, „ob ‚gute' Schüler auch zugleich ‚intelligente' Schüler sind und umgekehrt". Die meisten der von GAEDIKE gesammelten Untersuchungsergebnisse zur Korrelation zwischen (allgemeinen) Intelligenztest- und Schulleistungswerten liegen im Bereich von $r = 0.40$ und $r = 0.70$, gelegentlich auch darüber oder darunter. Dabei korrelieren „schulnahe" Intelligenztests (Verbaltests) im allgemeinen mit der Schulleistung höher als „schulferne" (Handlungstests). Ferner korrelieren Intelligenztests höher mit Schulleistung*stests* als mit anderen Indikatoren, z. B. Schulzensuren. Entsprechende Zusammenhänge fallen bei jüngeren Schülern (Grundschule) deutlicher aus als bei älteren Schülern (Sekundarstufe) und dort wiederum deutlicher bei Hauptschülern als bei Realschülern und Gymnasiasten, was analog auch — durchgängig — für den Vergleich von Mädchen und Jungen gilt.

Allerdings tragen nicht alle kognitiven Faktoren in gleichem Maße zum Schulerfolg bei. So bestimmen vor allem verbale Fähigkeiten, Faktoren des

logischen und schlußfolgernden Denkens (Reasoning) sowie Zahlenverständnis, teilweise auch technische Fähigkeiten (Space) und Wahrnehmungsfaktoren (Perceptual Speed, Accuracy) den Bildungserfolg, gemessen an den traditionellen Formen der Schulleistung (siehe noch HELLER 1970, S. 127 ff.). „Es kommt also nicht (nur) darauf an, intelligent oder kreativ zu sein, sondern (von den Bildungsinstitutionen) bevorzugt werden Schüler, die in ganz bestimmter Weise intelligent sind."

Ausführlicher werden dann die nicht-kognitiven Determinanten der Schulleistung erörtert. Hierbei gewinnen die Lern- und Leistungsmotivation sowie Faktoren der Arbeitshaltung des Schülers, Persönlichkeitsvariablen wie Ängstlichkeit, Selbstachtung, Extraversion vs. Introversion, aber auch Merkmale des Lehrerverhaltens (z. B. Setzung „sachfremder" Leistungsmotivation, Entwicklung und Pflege kognitiver Stile, Unterrichtsatmosphäre, direktives Verhalten, Werthaltungen und Einstellungen) sowie äußere und innere Bedingungen des sozio-kulturellen Hintergrundes (Familie, Peergroups) mehr oder weniger starken Einfluß auf das Leistungsverhalten des Schülers. Die aufgezählten Determinanten bilden ein Bezugsgeflecht vielseitiger und mehrschichtiger Einflußgrößen. Dabei scheinen die Bezugsmuster einem Variationsspielraum ausgesetzt zu sein, der größer ist als der, dem die Phänomene der Schulleistung selbst unterliegen.

Somit liegt nach dem resümierenden Urteil der Autorin „noch ein weites Feld für entsprechende Forschungen vor uns", bis alle Beziehungen aufgedeckt und deren Kenntnis für eine optimale Schulleistungsförderung eingebracht werden können. Immerhin dürfte eine Reihe konkreter Hinweise am Ende jedes Abschnittes diese Aufgabe des Schulpädagogen schon jetzt spürbar erleichtern.

1.2. Aufgaben und Ziele schulischer Leistungsbeurteilung

1.2.1. Leistungsbeurteilung im Dienste der Unterrichtsorganisation und Bildungsreform

Diese Funktion der Leistungsbeurteilung steht in unmittelbarem Zusammenhang mit schulpädagogischen bzw. unterrichtsdidaktischen Problemen. „Nicht die Beurteilung und Benotung eines Schülers, die fast schon zum Selbstzweck geworden ist, sondern die Bewertung des Unterrichtserfolges und die Erfassung von Lernschwierigkeiten ist u. E. die vornehmste Aufgabe pädagogisch-psychologischer Prüfmethoden, ausgerichtet auf das Ziel, rechtzeitiges und wirkungsvolles pädagogisches Handeln zu ermöglichen und zu initiieren" (ZIELINSKI 1973, S. 58). Unter dieser Prämisse dienen schuli-

sche Leistungskontrollen vorab zur Überprüfung aufgestellter Lernziele und zur Diagnose des Lehr-Lernprozesses. Die so gewonnenen Informationen bieten unverzichtbare Hilfen für die Curriculumentwicklung und -revision, die Lernzielbestimmung, die didaktische Planung und Analyse des Unterrichtsverlaufs u. dgl. m. Die Leistungsbeurteilung ist somit notwendiger Bestandteil einer optimalen Unterrichtsgestaltung.

Unter dem Aspekt der (inneren) Bildungsreform wäre hier die Wechselwirkung von Curriculumentwicklung und Leistungsbeurteilung hervorzuheben (vgl. Symposium des Europarates in Berlin 1971). So erfordert die Überprüfung festgelegter Curricula, z. B. im Hinblick auf schulpädagogische und lernpsychologische Möglichkeiten, jeweils bestimmte Methoden ihrer Bewertung. Theoretisch bedeutet dies, daß *alle* kognitiven und nicht-kognitiven Determinanten der konkret geforderten Schülerleistung in einem solchen Bewertungssystem Berücksichtigung finden müssen. Praktisch ist man dieser Idealforderung bislang allenfalls bezüglich der Kontrolle der kognitiven Variablen nahegekommen, wohingegen nicht-kognitive Variablen nur gelegentlich Beachtung fanden (vgl. Kap. 2.2. in diesem Buch). Andererseits bewirkt die Entwicklung bestimmter Beurteilungsmethoden oft auch Modifikationen und Ergänzungen schulischer Curricula, deren Beschränkung auf kognitive Inhalte keine Notwendigkeit darstellt. Daraus läßt sich die Forderung ableiten, die Diskussion über schultestmethodische und curriculare Neuerungen nicht isoliert, sondern in wechselseitiger Abstimmung voranzutreiben. Nur so kann man der Gefahr unerwünschter Stabilisierungseffekte wirksam begegnen.

1.2.2. Leistungsbeurteilung als Funktion individueller Beratung

Dieser Aspekt der Leistungsbeurteilung zielt auf die „Auto-Evaluation", die Orientierung des Schülers (bzw. seiner Eltern) über seine eigenen Schulleistungen, seine Lernfortschritte (intraindividueller Vergleich) und seine Position innerhalb der Klassen- oder Altersgruppe (interindividueller Vergleich). Dazu ist zweierlei notwendig: erstens die genaue Kenntnis der Lernanforderungen, was eindeutig definierte Lernziele erfordert, und zweitens die Transparenz des Beurteilungssystems, dessen unerläßliche Bestandteile objektive, zuverlässige und gültige Bewertungskriterien bilden. Ferner müssen dem Schüler Notwendigkeit, Zweck und Formen der Leistungsbeurteilung einsichtig gemacht werden.

Während im vorigen Abschnitt unterrichtsdidaktische und somit lehrerzentrierte bzw. gruppenimmanente Interessen (z. B. der Schulklasse bzw. Lerngruppe) im Vordergrund der Betrachtung standen, kommt unter dem Gesichtspunkt schülerzentrierter Aufgabenfunktionen der Individualdiagnose besondere Bedeutung zu. Hier sind diagnostizierte Begabungs-Leistungs-

diskrepanzen (z. B. Phänomene des sog. Underachievement versus Overachievement, d. h. unter- versus übererwartungsgemäße Schulleistungen), vorübergehende oder andauernde Lern- und Leistungshemmungen (z. B. Konzentrationsstörungen, sozio-kulturelle oder sensorische Deprivationserscheinungen wie Sprachbarrieren oder Hör- und Sehfehler), aber auch partielle versus allgemeinere Leistungsschwächen (z. B. Legasthenie, Rechenschwäche usw. versus Lernbehinderung[en] im Sinne der Sonderschulbedürftigkeit) sehr oft unerläßliche Voraussetzung für die Bestimmung angemessener (schul)pädagogischer Maßnahmen bzw. therapeutischer Behandlungsansätze. Darüberhinaus will — und sollte — jeder Schüler, auch der ‚unauffällige', kontinuierlich über den persönlichen Stand und Fortschritt und erst recht über eventuelle Rückschritte im Lernprozeß informiert werden. Ob diese Information durch Tests, sog. Diagnosebögen oder Zensuren erfolgt, ist prinzipiell von zweitrangiger Bedeutung, sofern die verwendeten Methoden der Leistungsbeurteilung hinreichend objektiv, verläßlich und gültig sind. Daß diese Anforderungskriterien in der Praxis der Schülerbeurteilung — besonders bei den subjektiven Verfahren (z. B. der Notengebung) — häufig nicht erfüllt sind, betrifft eine Reihe von Problemen, die im folgenden Kapitel 1.3. bzw. 1.3.3. zur Sprache kommen.

1.2.3. Leistungsbeurteilung als Funktion der Schullaufbahn- bzw. Systemberatung

Die bisher erörterten Funktionen schulischer Leistungsbeurteilung sind nicht unabhängig von der Struktur des jeweiligen Schul- und Bildungssystems, wie umgekehrt eine gewisse feed-back-Wirkung auf die Bildungsinnovation zu erwarten ist. Von unterschiedlichen Aufgabenschwerpunkten und entsprechend modifiziert eingesetzten Beurteilungstechniken, etwa im dreigliedrigen versus Gesamtschulsystem, abgesehen kulminieren sämtliche Beurteilungs- und Beratungsansätze letztlich um die Differenzierungsproblematik. An dieser Frage und ihrer Lösung führt m. E. kein Weg vorbei, wenigstens ist bislang kein funktionierendes Schulsystem, das diese Problematik ohne Schaden ausklammern könnte, bekannt.

Die Schullaufbahnberatung im traditionellen (vertikal gegliederten) Schulsystem orientiert sich vornehmlich an den Anforderungskriterien der einzelnen Schularten (Hauptschule, Realschule, Gymnasium), für die — über kürzere oder längere Perioden — in etwa konsistente Lern-Leistungsbedingungen bzw. invariante Schülermerkmale postuliert werden. Der Gefahr systemstabilisierender Tendenzen versucht man hier durch regelmäßige Leistungskontrollen mit entsprechenden Übergangsmöglichkeiten (Prinzip der Durchlässigkeit) zu begegnen. Eine solche Laufbahnberatung müßte sich

allerdings weniger am Selektions- als vielmehr am Klassifikationsmodell orientieren (vgl. HELLER 1970).

In der Gesamtschule verlagern sich analoge Wahl- und Entscheidungsprozesse mehr nach ‚innen', d. h. in den Unterricht und das didaktische Handeln. Hier werden dann Kriterien benötigt, die angemessene Gruppierungen im Niveauunterricht, Wechsel zwischen Fachleistungskursen u. ä. erlauben bzw. sinnvolle Wahlpflichtkombinationen im Hinblick auf die Berechtigungsfunktion der Schulabschlußqualifikation garantieren. Dazu bedarf es wiederum objektivierter Lernleistungskontrollen.

Die angeführten Beispiele zeigen, daß es bislang und auch wohl in absehbarer Zukunft kein Bildungssystem gibt, das auf die Funktion der Leistungsbeurteilung in dieser oder jener Form verzichten könnte. Die unterschiedlichen Methoden schulischer Leistungsbeurteilung und ihre Einsatzmöglichkeiten im Dienste der hier getrennt aufgewiesenen, in Wirklichkeit jedoch verzahnten, Aufgabenfunktionen stehen deshalb im Mittelpunkt der folgenden Ausführungen. Dabei sollen theoretische und praktische Probleme gleichermaßen Beachtung finden.

1.3. Formen und Methoden der Leistungsbeurteilung im Bildungswesen

1.3.1. Testtheoretische Grundlagen

1.3.1.1. Die klassische Testtheorie und ihre Kritik

Die klassische Testtheorie, auch Meßfehlertheorie genannt, bildet nach wie vor die Hauptgrundlage standardisierter (formeller und informeller) Schulleistungstests. Jeder, der sich solcher Meßverfahren im Rahmen der Schülerbeurteilung bedient, sollte deshalb — schon um unkritischer Anwendung und Fehleinschätzungen vorzubeugen — die theoretischen Voraussetzungen testdiagnostischen Vorgehens in etwa kennen. Da die klassische Testtheorie unabhängig vom Inhalt der einzelnen Verfahren (z. B. Intelligenztest, Leistungstest usw.) gilt, maß man diesem Konzept in der pädagogisch-psychologischen Diagnostik bis in die jüngste Vergangenheit praktisch uneingeschränkte Bedeutung bei. Obwohl heute die Relativierung dieses Anspruchs unbestritten ist, finden sich zunehmend Tendenzen, die Theorie als solche überhaupt abzulehnen. Abgesehen davon, daß man damit das Kind mit dem Bad ausschüttet, entarten solche Versuche nicht selten in absurde ideologische Verstrickungen, wie ein kürzlich erschienener Artikel in der Deutschen Schule eindrucksvoll demonstriert (vgl. NEANDER 1973). Damit ist m. E. weder der theoretischen Neubesinnung noch den praktischen Bedürfnissen in irgendeiner Weise gedient.

Das Wort „Test" bedeutet ursprünglich ‚Prüfung', ‚Stichprobe' oder ‚Zeugnis', wobei sich der Begriff in der Testdiagnostik vor allem auf standardisierte Verhaltensstichproben bezieht, d. h. auf Prüfungen, die unter genau festgelegten Bedingungen durchgeführt werden. Solche Prüfverfahren oder *Tests* messen bestimmte Merkmalsausprägungen (z. B. Intelligenz, Schulleistung, Ängstlichkeit usw.), indem sie *inter*individuelle Unterschiede (Unterschiede zwischen den einzelnen Individuen) oder *intra*individuelle Unterschiede (Unterschiede in bezug auf dieselbe Person, z. B. im Verlauf der Ontogenese) erfassen. Bei den sog. standardisierten Schulleistungstests steht der erste Aspekt im Vordergrund, d. h. der Vergleich eines individuellen Meßwertes im Test mit der Gruppennorm, die als Testleistungsmaßstab in Form von Alters- und/oder Schul- bzw. Klassennormen operational definiert ist.

In dem testtheoretischen Beitrag von LANGFELDT (s. Kap. 3.1.) werden zunächst die Begriffe „Messen" (Testen) und „Meßskala" sowie die wichtigsten Anforderungskriterien in bezug auf Messungen, auch „Gütekriterien" genannt, erläutert. *Messen* (Testen) wird hier als ein Vorgang des Vergleichens anhand eines Maßstabs (Testnormen) aufgefaßt, wobei die Meßwerte durch unterschiedliche Skalenniveaus repräsentiert sein können. Die wichtigsten Skalentypen sind: 1. die *Nominal-* oder *Klassifikationsskala* (unterste Ebene des Messens), die nur einem Kriterium, dem der Äquivalenz, genügt; 2. die *Ordinal-* oder *Rangskala* (nächsthöhere Ebene), die zwei Kriterien, dem der Äquivalenz und dem der rangmäßigen Beziehung, genügt; 3. die *Intervallskala*, die drei Kriterien, der Äquivalenz, der Rangbeziehung und der Intervallkonstanz, genügt; 4. die *Rational-* oder *Verhältnisskala* (oberste Ebene des Messens), die zusätzlich zu den aufgeführten Kriterien einen absoluten Nullpunkt aufweist, also vier Kriterien genügt. Je höher das repräsentierte Skalenniveau ist, desto besser und vielfältiger sind die Verarbeitungsmöglichkeiten entsprechender Kennwerte und damit ihr Informationsgehalt. Schulleistungstests messen in der Regel auf Rangskalenniveau, selten auf Intervallskalenniveau (z. B. Intelligenztests).

Unabhängig vom jeweiligen Skalenniveau wird der Interpretationsspielraum noch durch die sog. *Testgütekriterien* (Objektivität, Reliabilität, Validität) beeinflußt. *Objektivität* meint hier die Unabhängigkeit der Testergebnisse von der Person des Testleiters sowohl in bezug auf die Testanweisung (Instruktion) als auch in bezug auf die Testauswertung und Testinterpretation. Die Durchführungsbestimmungen eines Tests müssen so präzise festgelegt sein, daß Intersubjektivität gewährleistet ist. Diese Forderung ist in der Praxis der Testdurchführung nicht immer leicht zu erfüllen. Die *Reliabilität* oder Zuverlässigkeit bezieht sich auf die Meßgenauigkeit. Ein Schulleistungstest ist beispielsweise dann reliabel (zuverlässig), wenn die Ergebnisse unabhängig vom Zeitpunkt der Messung zustandekommen, d. h.

wenn die wiederholte Anwendung des betr. Tests bei derselben Klasse oder einzelnen Individuen in etwa zum gleichen Resultat führt. Es gibt verschiedene Aspekte der Reliabilität und dementsprechend unterschiedliche Methoden ihrer Kontrolle, die alle von LANGFELDT besprochen werden. Die *Validität* oder Gültigkeit bezieht sich auf die Genauigkeit, mit der ein Test das mißt, was er messen soll. So sagt die Validität eines Rechentests etwas darüber aus, wie genau tatsächlich Rechenkenntnisse (und nicht etwa Konzentrationsfähigkeit oder andere Merkmale) erfaßt werden. Während sich also die Reliabilität auf die Zuverlässigkeit des Meßinstrumentes (Tests) selbst und damit auf die formale Meßgenauigkeit bezieht, informiert uns die Validität darüber, „welche psychodiagnostischen Schlußfolgerungen die numerischen Resultate eines Tests zulassen und welchen Grad an Sicherheit solche Schlußfolgerungen aufweisen" (Michel 1964, S. 47). Damit ist die diagnostische Valenz oder Treffsicherheit eines Testverfahrens angesprochen. Was endlich den Zusammenhang zwischen Reliabilität und Validität betrifft, so gilt: Die Reliabilität ist die notwendige, aber keine hinreichende, Voraussetzung für die Validität eines Tests. Sehr oft erweist sich die Validierung eines Schultests als das schwierigste Unterfangen überhaupt, wobei man sich nicht selten mit der Bestimmung der curricularen Gültigkeit begnügt. Seltener kommen hier die Methoden der Kriterienvalidierung (z. B. zur Ermittlung der Übereinstimmungsvalidität von Testleistung und Schulnoten als „Außenkriterien") oder der Konstruktvalidierung (analog zur Validierung „neuer" Intelligenztests; vgl. HELLER 1973, S. 75) zum Einsatz. Bei der Konstruktion eines Schulleistungstests kommt man den Anforderungskriterien der Reliabilität und Validität dadurch entgegen, daß man schon frühzeitig — in der sog. Aufgaben- oder Itemanalyse — *Schwierigkeit* und *Trennschärfe* jeder einzelnen Testaufgabe ermittelt, um von vornherein die unbrauchbaren Aufgaben (Items) auszusondern.

Damit nun die Ergebnisse eines Schulleistungstests vernünftig interpretiert werden können, sind zwei weitere Voraussetzungen unerläßlich: die Normierung des Tests und die Ermittlung des sog. (Standard-)Meßfehlers. In den *Testnormen* (z. B. PR = Prozentränge auf Ordinalskalenniveau versus Z, C, Abweichungs-IQ u. ä. Standardwert-Normen auf Intervallskalenniveau bzw. flächentransformierte T-Standard-Äquivalent-Normen) liegen hierfür relativierte Testwerte, d. h. auf die jeweilige Alters- oder Klassengruppe bezogene Vergleichsmaßstäbe vor. So besagt eine Schülerleistung auf dem 75. PR, daß der betreffende Schüler in dem untersuchten Schulfach bessere Leistungen erzielt als 75 % seiner Bezugsgruppe (und schlechtere Leistungen als 25 %), während eine Leistung auf dem 50. PR durchschnittliche Leistungsfähigkeit indiziert. Nun wäre es naiv anzunehmen, daß die ermittelte (beobachtete) Testleistung von PR = 75 exakt die tatsächliche (wahre) Leistungsfähigkeit des Schülers wiedergibt; vielmehr muß davon

ausgegangen werden, daß in diesem Testergebnis — wie in jedem Meßwert — eine zunächst unbekannte Fehlerquote steckt, die zu Lasten der Irreliabilität bzw. mangelnden Objektivität einer Untersuchung geht. Diese Annahme trifft im Kern das *Meßfehler*konzept, das wichtigste Axiom der klassischen Testtheorie. Danach setzt sich jeder *beobachtete Wert* (Meßwert) aus dem „wahren Wert" (einem zeitlich konstanten Parameter) und einem „Fehlerwert" (Meßfehler) zusammen. Beide werden als unkorrelierte Größen postuliert, wobei der Meßfehler als Zufallsvariable auftreten soll. Mit dessen Hilfe werden dann — auf der Basis wahrscheinlichkeitstheoretischer Überlegungen — die sog. *Vertrauensintervalle* oder *Konfidenzintervalle* ermittelt, d. h. jene Bereiche, in denen mit einer bestimmten Wahrscheinlichkeit der wahre Wert erwartet wird. Auf diese Weise läßt sich die Interpretation der durch Tests erfaßten Schülerleistungen auf eine (meßtheoretisch) gesicherte Grundlage stellen — ein Vorteil gegenüber anderen Formen der Schülerbeurteilung (z. B. Zensurierung), der sowohl dem interindividuellen Vergleich als auch intraindividuellen Untersuchungszielen (z. B. der Erfassung von Lernleistungsfortschritten oder sog. Profilanalysen) zugute kommt (s. noch HELLER 1973, S. 66 ff. u. 157 ff.).

Die *Kritik der klassischen Testtheorie* setzt an der Axiomatik an, d. h. an Widersprüchen zwischen einigen testtheoretischen Voraussetzungen und empirischen Befunden dazu. Die wichtigsten Voraussetzungen der klassischen Testtheorie lassen sich in folgenden drei Axiomen zusammenfassen (nach HELLER 1973, loc. cit.). 1. *Existenzaxiom:* Zu jedem beobachteten (gemessenen) Wert existiert ein „wahrer" Wert im Sinne einer bestimmten individuellen Merkmalsausprägung (z. B. Höhe der Schulleistung). Diese wird als Konstante — wenigstens über einen gewissen Zeitraum hinweg — angenommen. 2. *Fehleraxiom:* Der Meßfehler einer Messung ist eine Zufallsvariable. Für diese gilt, daß die Summe bzw. das arithmetische Mittel der Fehlerwerte den Wert Null ergibt. 3. *Verknüpfungsaxiom:* Der beobachtete Wert (Meßwert) setzt sich additiv aus wahrem Wert und Fehlerwert zusammen. Daraus kann eine Reihe von Sätzen abgeleitet werden, worauf wir nicht mehr eingehen wollen.

Die auf der Grundlage der klassischen Meßfehlertheorie konstruierten Schulleistungstests lassen vor allem folgende Phänomene ungeklärt: erstens die Beobachtung, daß — entgegen obiger Annahme — die Fehlerwerte nicht unabhängig von den wahren Werten auftreten, d. h. „daß unterschiedliche wahre Werte auch unterschiedliche Fehlerwerte bedingen" können und diese nicht allein von der (mangelnden) Reliabilität des Tests, sondern auch von der jeweiligen Untersuchungspopulation beeinflußt werden; zweitens die (prinzipielle) Variabilität der wahren Werte, also die Tatsache, daß Merkmalsschwankungen (z. B. Intelligenz- oder Schulleistungszuwachs durch gezielte Fördermaßnahmen) auftreten können; drit-

tens die damit verbundene Schwierigkeit, meßtheoretischen Anforderungen (vor allem der Reliabilität und Validität) *und* berechtigten schulpädagogischen Anliegen (Lernleistungsförderung) gleichermaßen zu genügen. Inwieweit diese Probleme durch die „modernen" testtheoretischen Ansätze einer Lösung zugeführt werden können, soll im folgenden behandelt werden.

1.3.1.2. Neuere Modellansätze und ihre Problematik

Nach einigen terminologischen Vorklärungen befaßt sich BÜSCHER (s. Kap. 3.2.) mit der durch das RASCH-Modell eingeleiteten neuen Entwicklung einer *psychologischen* Meßtheorie, die vor allem am Problem der Populationsabhängigkeit klassischer Leistungsmessung ansetzt. „Für gewöhnlich werden die Eigenschaften eines psychologischen Tests durch individuelle Unterschiede innerhalb einer bestimmten Population definiert, und die Beurteilung jeder einzelnen Person ist mit der Referenzpopulation verknüpft. In vielen Fällen ist man jedoch daran interessiert, Individuen *per se* zu vergleichen, ohne sich dabei auf eine Population beziehen zu müssen. Die Frage etwa, ob sich ein Kind während eines oder mehrerer Jahre verbessert hat, kann unmöglich mit Hilfe verschiedener Tests beantwortet werden, die an Populationen unterschiedlich alter Kinder geeicht wurden. Im Rahmen eines neuen testtheoretischen Ansatzes versuchte RASCH (1960) derartige Probleme zu lösen, um das Studium einzelner Personen oder einzelner Testitems (Testaufgaben) unabhängig von Referenzpopulationen zu ermöglichen" (STENE 1968, S. 229).

Am Beispiel *kriterienbezogener* Leistungstests behandelt nun BÜSCHER alle einschlägigen Probleme moderner testtheoretischer Provenienz. Ausführlich werden hierbei neue Verfahrensansätze zur Test- bzw. Itemanalyse erörtert. Deren praktische Relevanz im Hinblick auf die Schülerbeurteilung ist jedoch — vorerst noch — ungeklärt. Symptomatisch hierfür ist die abschließende Bewertung des RASCH-Modells durch BÜSCHER. „Die praktische Anwendung des RASCH-Modells auf lehrzielorientierte Tests ist problematisch, da zum einen eine große Anzahl von Probanden erforderlich ist, zum andern der Konstruktions- und Rechenaufwand enorm ist und schließlich das Modell nicht mehr angewandt werden kann, wenn alle Probanden das Lehrziel erreicht (bzw. nicht erreicht) haben. Das bedeutet, wenn überhaupt, dann ist das RASCH-Modell nur für normbezogene Messung sinnvoll."

Damit, so könnte es scheinen, sind wir wieder an den Anfang unserer Problemdiskussion verwiesen. Dem ist nicht so. Vielmehr stehen wir mitten in einer Umorientierungsphase, in der neue Denkmodelle entwickelt (vgl. FISCHER 1968, FRICKE 1972) und die klassische Testtheorie relativiert (nicht verworfen) werden. Soviel steht jetzt schon fest, daß kriterien-

bezogene Messungen genauso wie normbezogene „ihre eigene bedeutsame pädagogische Funktion" haben. „Keineswegs wird das eine Meßverfahren das andere verdrängen."

1.3.1.3. Vorschläge zur Klassifikation von Schultests

So unterschiedlich wie die testtheoretischen Ansätze ist die in der Literatur verwendete Terminologie im Bereich der Schulleistungsmessung. Nach einer kritischen Analyse in sich oft widersprüchlicher Konzepte unterbreitet ROSEMANN (s. Kap. 3.3.) eigene Vorschläge zu einer *pädagogisch* begründeten Klassifikation von Schultests. Während sich die konventionellen Bezeichnungen vorwiegend an meßtheoretischen Kriterien orientieren, begründet ROSEMANN sein Klassifikationskonzept auf den *Funktionen* schulischer Leistungsbeurteilung. Dabei ist die Unterscheidung von „Leistungsfeststellung" und „Leistungsbewertung" von Bedeutung. „Im ersteren Falle verschafft man sich lediglich Information darüber, was die Schüler im Verlaufe des Unterrichtsgeschehens gelernt bzw. nicht gelernt haben. Diese Informationen per se kann der Lehrer in vielfältiger Weise verwenden. Im zweiten Falle geht man einen Schritt weiter, man will die festgestellten Leistungen der Schüler bewerten, wobei verschiedene Bezugspunkte für die Bewertung in Betracht kommen können." Für Testverfahren, die im Dienste der erstgenannten Funktionseinheit stehen, schlägt der Autor die Sammelbezeichnung „Lernsteuerungstests" vor, für Verfahren der zweiten Kategorie die Bezeichnung „Lernkontrolltests". Ohne Zweifel werden mit dieser Einteilung zentrale Aufgabenfunktionen der Leistungsdiagnostik in der Schule getroffen, nämlich die „Lenkung und Steuerung des Lernprozesses" und die „Bewertung der Ergebnisse dieses Vorganges". Während die erste Zielsetzung dem „permanenten Informationsaustausch zwischen Lernenden und Lehrenden" dient und somit eine optimale Instruktion ermöglichen soll, liegt der entscheidende Vorteil der Lernkontrolltests in der „Objektivierung des Bewertungsvorganges", also der im Vergleich zur nicht-testgebundenen subjektiven Leistungsbeurteilung — Benotung nach schriftlicher (z. B. Klassenarbeiten) oder mündlicher Prüfung, Aufsatzbeurteilung usw. — größeren Zuverlässigkeit der Urteilsfindung.

Anhand des ROSEMANNschen Ordnungskonzeptes lassen sich nunmehr die verschiedenen Testformen folgendermaßen zusammenfassen: in *Lernsteuerungstests* (nach bisheriger Benennung die [informellen] kriteriumsbezogenen Schulleistungstests ohne Benotung und — gelegentlich auch — normbezogene Tests für kleinere Lerneinheiten) und *Lernkontrolltests* (die sog. standardisierten Schulleistungstests, informelle normbezogene Tests sowie [informelle] kriteriumsbezogene Tests mit Benotung). ROSEMANN selbst betont

den Charakter seines Ordnungskonzeptes als Diskussionsgrundlage, weshalb wir auch bewußt zunächst darauf verzichteten, die einzelnen Beitragsautoren quasi auf eine einheitliche Terminologie in diesem Band zu verpflichten. Die Vorschläge ROSEMANNS scheinen mir jedoch einer gründlichen Überlegung wert, vor allem im Hinblick auf die praktischen Bedürfnisse der Testanwendung in der Schule.

1.3.2. Objektive Verfahren schulischer Leistungsbeurteilung

1.3.2.1. Lernzieldefinition als Voraussetzung der Leistungsmessung

Anders als beim Programmierten Unterricht (PU), wo man sich von Anfang an vor die Notwendigkeit gestellt sah, *operationalisierte* Lehr-/Lernziele zu formulieren, sind die Probleme einer exakten Unterrichtsplanung im Sinne präziser Zieldefinitionen in der traditionellen Unterrichtslehre verhältnismäßig spät Gegenstand der Diskussion geworden. Virulent wurden solche Problemfragen eigentlich erst in dem Moment, wo das Bemühen um Überprüfung der Unterrichtsziele — analog zum PU — einsetzte, sei es im Rahmen curricularer Innovationen, bei der Erprobung neuer Schulmodelle (Gesamtschule) und damit zusammenhängenden Fragen der Unterrichtsdifferenzierung oder auch, um didaktische und schulpädagogische bzw. therapeutische Fördermaßnahmen auf leistungsdiagnostischer Grundlage zu sichern. Dazu bedarf es in jedem Falle operationalisierter Lernziele, d. h.: Lernleistungsprüfungen sind ohne (vorausgehende) Lernzielbestimmung weder pädagogisch sinnvoll noch meßtechnisch (via Kriteriumstests) durchführbar.

Operationalisierung von Lernzielen meint hier (nach HORN in Kap. 4.1.) die genaue Festlegung der vom Schüler am Ende einer Unterrichtseinheit geforderten Verhaltensweisen (Operationen). „Diese Verhaltensweisen müssen direkt beobachtbar sein. Daher findet man bei Lernzielen häufig Formulierungen wie ‚Der Schüler soll ... nennen (aufschreiben, lösen usw.) können'."

Bei der Festlegung von Lernzielen orientiert man sich heute vielfach an der BLOOMschen Taxonomie, die 6 verschiedene, hierarchisch geordnete, Komplexitätsstufen enthält: Wissen, Verstehen, Anwendung, Analyse, Synthese, Evaluation. Die Kategorien 2 bis 6 betreffen dabei mehr „intellektuelle Fähigkeiten" und repräsentieren Formen zunehmender Komplexität des kognitiven Verhaltens, die auch als „Strategien des Problemlösens" (*Problemlösungsstrategien* = relativ inhaltsunabhängige „Grundmuster") aufgefaßt werden können. Für die praktische Anwendung des BLOOMschen Klassifikationsmodells bei der Unterrichtsplanung gibt HORN wertvolle

Hinweise und erläutert diese an einem Beispiel aus dem Naturkundeunterricht, wobei die Darstellung seines eigenen Lehrplan-Analyseschemas besondere Beachtung verdient. Ausführlich wird dann auf die Konstruktion von Prüfungsaufgaben eingegangen. Auch hierzu bietet HORN eine Reihe praktischer Beispiele. Insgesamt wird somit deutlich, daß die Messung schulischer Lernleistungen vor allen Methodenfragen zunächst von exakten Lernzieldefinitionen abhängt, wobei der Operationalisierung der Lernziele vorrangige Bedeutung zukommt.

1.3.2.2. Informelle Tests (Lernkontrolltests)

Hauptanliegen des folgenden Beitrags von ROSEMANN (s. Kap. 4.2.) ist die Information über Lernkontrolltests, aufgezeigt am Beispiel sog. informeller Leistungsmessung. Dabei soll der Leser mit den Konstruktionstechniken und einigen Problemen informeller Tests soweit vertraut gemacht werden, daß er in der Lage ist, „den einen oder anderen Test selbst zu entwickeln bzw. bestehende Tests kritisch zu beurteilen".

Ein den Ausführungen vorangestelltes Ablaufdiagramm vermittelt einen schnellen Überblick über die einzelnen Arbeitsschritte. Nach der *Operationalisierung der Lernziele* (siehe oben) und der Erstellung einer *Spezifikationstabelle* (in der nicht nur die Anzahl der Testaufgaben oder Items festgelegt, sondern auch eine Entscheidung darüber getroffen wird, welche Lernziele überprüft und somit in den Test aufgenommen werden sollen) erfolgt die eigentliche *Aufgabenkonstruktion*. Hierzu werden vom Autor alle relevanten Item- und Antworttypen besprochen und an Hand einschlägiger Beispiele das Vorgehen veranschaulicht. Ausführlich erörtert ROSEMANN dann die einzelnen Schritte zur Entwicklung der sog. *Testvorform* (Aufgabengruppierung, Erstellung des Testaufgabenheftes einschließlich der Instruktion, Festlegung des Bewertungsschlüssels usw.) sowie die Vorbereitung und Durchführung der *Itemanalyse* (Schwierigkeits- und Trennschärfenanalyse jeder einzelnen Testaufgabe sowie Distraktorenanalyse, Aufgabenselektion bzw. -revision). Schließlich werden die Methoden zur Überprüfung der *Reliabilität* und *Validität* sowie Verfahren zur *Normierung* des Tests dargestellt, wobei eine Reihe von Illustrationsbeispielen wiederum die praktische Arbeit der Testkonstruktion und zugleich das Verständnis testtheoretischer Grundlagen — nunmehr am konkreten Objekt — wesentlich erleichtern dürfte. Der Weg von der Lernzieldefinition bis hin zur *Testendform* ist somit lückenlos beschrieben.

Die Lektüre dieses Beitrags sollte jeden, der sich mit Fragen praktischer Schulleistungsmessung zu beschäftigen hat, verhältnismäßig rasch und umfassend über entsprechende Möglichkeiten (informeller Tests) informieren.

Selbst derjenige Leser, der noch über keine oder nur wenige Kenntnisse und Erfahrungen auf diesem Gebiet verfügt, wird sich hierbei m. E. ohne größere Mühe zurechtfinden.

1.3.2.3. Standardisierte Schulleistungstests

Im nächsten Beitrag (s. Kap. 4.3.) gibt HORN einen Überblick über die wichtigsten zur Zeit verfügbaren sog. standardisierten Schulleistungstests und deren Anwendungsmöglichkeiten im Rahmen der Schülerbeurteilung. Die Unterschiede zwischen *informellen* und *standardisierten* (formellen) Schulleistungstests sind freilich nur gradueller Natur, wie die Ausführungen ROSEMANNS in Kap. 3.3. deutlich gemacht haben. Demnach unterscheiden sich beide Verfahrensansätze vor allem bezüglich der vergleichsweise niedrigeren vs. höheren Reliabilität, des engeren vs. weiteren Anwendungsbereiches, der spezielleren vs. allgemeineren Testinhalte, der Normenfunktion, d. h. verfügbarer Testnormen für bestimmte Klassenstufen (derselben Schulart) vs. für verschiedene Klassen und Schularten sowie hinsichtlich einiger Konstruktionsaspekte, z. B. vorwiegend durch Lehrer vs. Testexperten erstellt (s. S. 158 f.). Die Bezeichnung „standardisierter" Schulleistungstest bezieht sich also keineswegs, wie oft angenommen wird, ausschließlich auf das Normenkriterium.

Unter *inhaltlichem* Gesichtspunkt lassen sich die standardisierten Schulleistungstests folgendermaßen gruppieren: 1. *fächerübergreifende* Tests oder sog. Omnibusverfahren (z. B. Allgemeiner Schulleistungstest für 2. Klassen AST 2 bzw. für 3. Klassen AST 3 usw.); 2. *fachspezifische* Tests (z. B. Erdkundetest Deutschland ETD 5—7, Geschichtstest Neuzeit GTN 8—10, Naturlehretest NLT 9 usw. oder Diagnostischer Englisch-Leistungstest ELT 6—7, Französischer Wortschatztest FWS 9—12 usw.); 3. *lernbereichsorientierte* Tests, etwa Lesetests (z. B. Lesetest LT 2), Rechtschreibtests (z. B. Diagnostischer Rechtschreibtest DRT 3), Rechentests (z. B. Bruchrechentest BRT 6). Im Hinblick auf den *Anwendungsbereich* bzw. die Zielgruppe oder Untersuchungs*population* könnte man schließlich Tests für verschiedene Schulstufen und Schularten unterscheiden: 1. Tests für den *Primar-* vs. *Sekundarstufen*bereich bzw. die einzelnen Schul*klassen;* 2. Tests für die Schul*typen* des Gymnasiums, der Realschule und Hauptschule vs. Grundschule — seltener Gesamtschulen, die fast ausschließlich informelle Tests zur Lernleistungskontrolle verwenden (s. GAUDE u. TESCHNER 1970).

Neben zahlreichen Test- und Aufgabenbeispielen sowie praktischen Hinweisen für den schulischen Einsatz standardisierter Leistungstests bringt HORN am Ende seines Beitrags eine Gesamtübersicht der wichtigsten in der BRD lieferbaren Schultests (Stand 1973). Darüberhinaus kann sich jeder

Leser leicht selbst an Hand der beigefügten Verlagsanschriften durch Anforderung von Prospektmaterial bzw. Gesamtverzeichnissen über die neuesten Testangebote informieren.

1.3.3. Subjektive Verfahren schulischer Leistungsbeurteilung

Zu den „subjektiven" Verfahren rechnen wir alle nicht-messenden Methoden der Schülerbeurteilung, d. h. Verfahrensweisen, die im Sinne der klassischen Meßfehlertheorie (vgl. Meß- bzw. Testgütekriterien) — erfahrungsgemäß — nicht objektiv sind und einen relativ geringen Grad an Zuverlässigkeit und Gültigkeit aufweisen. Im einzelnen fallen hierunter alle Formen konventioneller Notengebung, mündliche Prüfung, Aufsatzbeurteilung u. ä. Aber auch die in der Schulpraxis bislang viel zu wenig beachteten Methoden der (wissenschaftlichen) Verhaltensbeobachtung und einer Reihe von Beurteilungstechniken i. e. S. (vgl. HELLER et al. 1974, Kap. 2.1.3.) sind der Kategorie der subjektiven Verfahren zuzuordnen. Damit befaßt sich nun der folgende Beitrag von LANGHORST (s. Kap. 5.1.).

1.3.3.1. Verhaltensbeobachtung und Schülerbeurteilung

Über die Notwendigkeit der Verhaltensbeobachtung bzw. einzelner Techniken der Beurteilung i. e. S. im Unterricht sollte es eigentlich keine Diskussion mehr geben. Ohne den Einsatz dieser Methode(n) müßten wir oft auf wichtige Informationen verzichten, z. B. über das Schüler- und Lehrerverhalten im Interaktionsspiel des Unterrichtsablaufs, damit zugleich auf nicht via Tests erfaßbare Daten in bezug auf die Unterrichtsplanung, die Curriculumentwicklung, Lernzielbestimmung usw. STAKE formulierte dies in einem anderen Zusammenhang sehr deutlich, wenn er sagt (zit. nach ROSENSHINE 1973, S. 201): „Ohne Informationen über die Lehrmethoden ist es weder möglich, das Wesen eines Curriculums zu verstehen, noch zu wissen, was als nächstes ausprobiert werden muß. In manchen Evaluationsuntersuchungen werden die mit Hilfe der Unterrichtsbeobachtung gewonnenen Daten die wichtigsten und wertvollsten sein."

Nach einem kurzen Überblick über gängige Verfahren der Schülerbeobachtung und -beurteilung im Unterricht diskutiert LANGHORST eine Reihe von Möglichkeiten, diese Verfahrensansätze zu systematisieren und somit auf eine objektivere Basis zu stellen. Dabei gilt es zunächst, die *Hauptphasen der Verhaltensbeobachtung*, nämlich die *Beobachtung i. e. S.* (auf den Beobachtungsgegenstand zentrierte, Unwesentliches selegierende Wahrnehmung), die *Beschreibung* oder *Deskription* (Protokollierung des beobachteten Verhaltens) und die eigentliche *Beurteilung* (Interpretation der Beob-

achtungsdaten), strikt auseinanderzuhalten. Nur so kann die Zahl der subjektiven (unkontrollierbaren) Einflüsse auf ein Minimum reduziert und die Aussagekraft der Ergebnisse erhöht werden.

Im zweiten Teil seines Beitrags geht der Autor ausführlicher auf die Möglichkeiten der Schülerbeurteilung mit Hilfe von *Schätzskalen* (rating scales) ein. Unterrichtsrelevante Formen dieses Vorgehens werden an Hand von praktischen Beispielen demonstriert und kritisch auf ihre Verwendungsmöglichkeiten hin untersucht. In diesem Zusammenhang nimmt die Erörterung der häufigsten *Beurteilungsfehler* einen wichtigen Platz ein, entscheidet doch ihre Kenntnis letzten Endes darüber, „ob der Lehrer die Vielzahl der subjektiven Urteilstendenzen ... beim Umgang mit seinen Schülern in etwa kontrollieren und steuern kann ... Entsprechende Anstrengungen sind schon deshalb lohnenswert, weil — erwiesenermaßen — im Fall einer gerechten Verhaltensbeobachtung sich der Schüler vom Lehrer besser verstanden fühlt."

1.3.3.2. Leistungsbeurteilung durch Notengebung

„Schulnoten sind Maßzahlen für erbrachte Leistungen der Schüler, die der Lehrer nach seinen Erfahrungen und Einschätzungen auf der Notenskala einstuft. Noten kommen also aufgrund eines Urteilsprozesses des Lehrers zustande und sind mit all den Mängeln behaftet, die man bei Urteilsprozessen nachgewiesen hat" (FINGERHUT u. LANGFELDT in Kap. 5.2. dieses Buches). Dabei sind situative Einflüsse genauso beteiligt wie sozial- und persönlichkeitspsychologische Momente (vgl. Beurteilungsfehler). Daraus können wir folgern, daß Schulzensuren prinzipiell schlechte Indikatoren für Schulleistungen sind. Siehe noch INGENKAMP (1971).

Diese Vermutung läßt sich durch zahlreiche empirische Belege stützen. So kommen FINGERHUT und LANGFELDT in dem zitierten Beitrag, der die wichtigsten (neueren) Untersuchungen zu diesem Thema berücksichtigt, zu dem eindeutigen Ergebnis, daß die Lehrerzensuren herkömmlicher Art den meßtheoretischen Forderungen der Objektivität, Reliabilität und Validität nicht oder höchst unvollkommen genügen. „Dies kann den Lehrern jedoch nicht zum Vorwurf gemacht werden. In jeder Unterrichtsstunde wird von ihnen eine Vielzahl schneller und unabhängiger Entscheidungen und Beurteilungen verlangt, die sie nur bewältigen können, wenn sie bestimmte Urteils- und Verhaltensstrategien entwickeln. Diese notwendige Bildung von Stereotypen verhindert aber exakte Urteile." Daraus kann m. E. nur die Folgerung abgeleitet werden, dem Lehrer die nötigen Hilfen in Form objektiver und zuverlässiger Methoden (z. B. formelle und informelle Tests) an die Hand zu geben, um auf diese Weise die subjektiven Urteile abzusichern. Nur unter dieser Voraussetzung wäre es — wenn überhaupt —

sinnvoll, Schulnoten die fast alles im Bildungsgang entscheidende Funktion zuzuerkennen.

Die Absicherung subjektiver Lehrerurteile sollte also *mittelbar* durch Tests u. ä. erfolgen und nicht durch unkritische Angleichung an testtheoretische Modelle versucht werden, da ein solches Unterfangen mancherlei pädagogischen und erzieherischen Absichten widersprechen würde. So führt beispielsweise ZIELINSKI (1973, S. 11) nicht weniger als 10 verschiedene Funktionen der Zensierung auf. Sowohl die Intention der Schule, Lernleistungen zu fördern und damit Schülerleistungen (positiv) zu verändern als auch die Verquickung von pädagogischer Anreizfunktion und reiner Meßfunktion der Schulnoten lassen alle Bemühungen, konsistente (zuverlässige) Leistungsmessungen via Zensuren zu erzielen, praktisch im Ansatz scheitern. Entweder müßte man auf berechtigte schulpädagogische Anliegen hier verzichten, um — nicht weniger berechtigten — meßtheoretischen Anforderungen zu genügen, oder man sollte inkompatible Forderungen fallen lassen und schiedlich zwischen pädagogischen und Meßfunktionen bei der Notengebung trennen (was theoretisch möglich ist) bzw. ganz auf Noten verzichten (was vielen praktisch unmöglich erscheint). Zweifellos wäre die letzte Alternative die konsequenteste Haltung. Da wir aber ihre Realisierungschance momentan als gering erachten, bleibt wohl als *vorläufiger* Weg aus dem Dilemma nur die erste Alternative übrig, deren Gefahrenmomente (Vermischung der Funktionen, Beurteilungsfehler usw.) durch Testhilfen soweit als möglich kompensiert werden müssen. Hierzu werden die formellen und die informellen Schulleistungstests einschließlich sog. Standardarbeiten (vgl. WENDELER 1969) gleichermaßen beitragen können.

Prinzipiell wäre damit auch der Weg gewiesen, *einheitlichere Beurteilungsmaßstäbe* zu erzielen. Nach den Ergebnissen von FINGERHUT und LANGFELDT sowie zahlreicher anderer — in diesem Buch zitierter — Forscher werden Schulnoten fast durchweg fachspezifisch erteilt, wobei sich drei Gruppen herausbilden: „Fächer mit milder (musische Fächer und Religion), mit mittlerer (Nebenfächer) und mit strenger Beurteilung (Hauptfächer)". Die Möglichkeiten der Differenzierung sind außerdem bei milder Beurteilung stark reduziert. Hinzukommen inter- versus intraindividuelle, d. h. lehrerabhängige, Einfluß- bzw. Fehlervariablen (siehe oben). Damit wäre erneut die Notwendigkeit *testunterstützter Notengebung* demonstriert.

Die Möglichkeiten zur objektiven Leistungsbeurteilung sind freilich trotz des wachsenden Testangebots begrenzt, und diese Grenzen können mitunter in der Spezifität des Gegenstandes selbst liegen. Als Paradebeispiel hierfür sei die Aufsatzbeurteilung genannt. Andererseits wird m. E. zu Recht darauf hingewiesen, daß man — aus den verschiedensten Gründen — auf die Übung des Aufsatzes (und seine Bewertung) nicht verzichten sollte. Der Bedeutung des Gegenstandes angemessen beschäftigen sich die letzten zwei

Beiträge dieses Bandes ausführlicher mit der Zensierung von Schüleraufsätzen.

1.3.3.3. Probleme der Aufsatzbeurteilung

In drei aufeinander aufbauenden Untersuchungsserien versuchten NICKEL und WIECZERKOWSKI, „in einer quasi-experimentellen Situation den Einfluß verschiedener vermutlich bedeutsamer Variablen auf die Bewertung von Aufsätzen zu erforschen". Die methodisch sehr sorgfältig angelegte Studie, in der eine Reihe unabhängiger Beurteilergruppen (Schüler, Studenten der Erziehungswissenschaften und Psychologie sowie Referendare und praxiserfahrene Lehrer) Schüleraufsätze von Viertklaßkindern unter weitgehend standardisierten Bedingungen beurteilte, erbrachte — bei systematischer, isolierender Variation einzelner Variablen — folgende Hauptergebnisse (siehe die vollständige Darstellung in Kap. 5.3.).

Keinen Einfluß auf die Aufsatzzensur hat demnach das Geschlecht, und zwar weder des Schülers bzw. Aufsatzschreibers noch des Lehrers bzw. Aufsatzbeurteilers. Ebensowenig scheinen vorausgehende oder nachfolgende Arbeiten gleicher oder ähnlicher Art die Ausbildung individueller Bezugssysteme auf seiten des Beurteilers zu beeinflussen.

Demgegenüber konnten die Autoren folgende Einflußgrößen ermitteln: *Vor-* bzw. *Zusatzinformationen* über den Schüler, insbesondere über dessen sonstige Schulleistungen; *Praxiserfahrung* der Beurteiler, sowohl allgemein in bezug auf das Unterrichten als auch speziell im Hinblick auf Aufsatzbewertungen; verschiedene *Sprachkriterien* des Aufsatzschreibers (Aufsatzlänge, Sprachrichtigkeit vs. Fehlerhaftigkeit, Differenziertheit des Ausdrucksstils, Einfallsreichtum und Originalität, Handlungsganzheit u. ä.). „Eine Analyse des Einflusses verschiedener Sprachkriterien bei annähernder Konstanthaltung des Faktors ‚Länge' und ‚Sprachrichtigkeit' ergab, daß die drei Variablen ‚Originalität der Einfälle', ‚Differenziertheit des sprachlichen Ausdrucks' und ‚Flüssigkeit bzw. Abgeschlossenheit des Handlungsablaufs' allein für die Gesamtzensur bereits einen Schätzungseffekt von 52 % besitzen. Wenn also die Richtigkeit und die Länge der Darstellung als Grundlage für eine Beurteilungsdifferenzierung entfallen, läßt sich die Aufsatzzensur in beträchtlichem Umfang bereits allein aufgrund dieser drei Sprachkriterien voraussagen."

Diese Befunde gewähren einen guten Einblick in den Bedingungskomplex, unter dem subjektive Urteile über Schülerleistungen, aufgezeigt am Beispiel der Aufsatzbeurteilung, zustandekommen. Es ist denkbar, daß diesen Ergebnissen ein größerer Allgemeinheitsgrad zukommt, zumindest was die Beurteilung *sprachlicher* Leistungen betrifft. Darüberhinaus liefert die Arbeit wichtige Kriterien für eine einheitlichere Beurteilungspraxis — ein Thema, dem der nachfolgende Beitrag gewidmet ist.

Das Ausmaß mangelnder Übereinstimmung in der Benotung von Schüleraufsätzen ist tatsächlich erschreckend, wenn man die von WENDELER (s. Kap. 5.4.) zitierten zahlreichen Literaturquellen als repräsentativ für die heutige Praxissituation ansehen darf (genauer: muß). Nach Meinung des Autors ist das Problem der Aufsatzbeurteilung „keineswegs ein Spezialproblem des Sprachunterrichts, sondern das Problem dieser Prüfungsform als solcher". Dabei kann man sich den Gesamtvorgang zweiphasig vorstellen. „Der erste Schritt ist die *Leistungsermittlung*, bei der man die Gesamtleistung oder bestimmte Teilleistungen nach einem Bewertungsschlüssel als ‚richtig', ‚falsch', ‚gut' oder ‚schlecht' beurteilt, meist mit Punktwerten versieht, und bei der man daraufhin einen Gesamtpunktwert bestimmt.

Der zweite Schritt ist die *Zensierung*, d. h. die Zuordnung einer Zensur zu der errechneten Punktzahl." Die Crux der Aufsatzbewertung liegt in der ersten Phase, nämlich in den *Kriterien* der Leistungserfassung und deren *objektiven* Bestimmung. Abweichende Urteile über ein und denselben Aufsatz sind demnach vor allem Folgen unterschiedlicher Kriterienmaßstäbe.

Im Bemühen, einheitliche (verbindliche) Kriterien bzw. Kriteriensysteme für die Aufsatzbeurteilung zu entwickeln, waren besonders einige angelsächsische Versuche erfolgreich. So erstellten DIEDERICH, FRENCH und CARLTON auf induktivem Wege ein Kategoriensystem für die Aufsatzbewertung, das folgende 5 (faktorenanalysierte) *Urteilsdimensionen* enthält: Ideen, innere Form (Gliederung), Lebendigkeit (Originalität), Sprachrichtigkeit und Wortwahl (Flüssigkeit). Darüber sowie über eine Reihe weiterer Ansätze berichtet WENDELER ausführlich in seinem Beitrag. Die Übereinstimmung zahlreicher dieser Befunde mit den in Kap. 5.3. referierten ist unverkennbar und erhärtet ihre praktische Relevanz.

Im zweiten Teil seines Beitrags geht dann der Autor auf Fragen und Probleme der Objektivitätssteigerung im Kontext Aufsatzbeurteilung ein. Hierzu werden wiederum mehrere Ansätze zur Erfassung der *Urteilsobjektivität* referiert sowie zahlreiche praktische Maßnahmen zu ihrer Verbesserung erläutert. Die mit entsprechenden Beispielen versehenen Ausführungen sollten deutlich machen, „daß tatsächlich ein befriedigender Objektivitätsgrad erreichbar ist".

1.4. Ausblick

Die Schulleistungsdiagnostik ist Bestandteil einer umfassenderen *pädagogischen* Diagnostik und steht als solche im Dienste unterrichtlicher und erzieherischer Ziele. Sie kann und will nie Selbstzweck sein. Andererseits stellt sie in mannigfacher Weise die notwendigen Informationen für unumgängliche pädagogische Entscheidungen zur Verfügung, die ohne diese Hilfen

nicht oder nur unzulänglich zu fällen wären. Diese Bewertung diagnostischer Funktionen und Möglichkeiten bedeutet keine Technokratisierung des Unterrichts, wie vielfach behauptet wird, sondern auf rationaler Basis begründete Lehr-/Lernprozesse, wofür wir uns allerdings aussprechen, nicht zuletzt unter dem Gesichtspunkt pädagogischer und bildungspolitischer Forderungen wie der der Verwirklichung von Chancengleichheit im Bildungsgang.

Demgegenüber gilt es, auf die Grenzen diagnostischer Möglichkeiten hinzuweisen. Diagnostische Hilfen in Form von Testergebnissen oder anderen Urteilen (z. B. Aussagen via Schätzskalen) können dem Lehrer die pädagogische Entscheidung selbst nie abnehmen. Diese Einschränkung gilt gleichermaßen im Hinblick auf die praktischen Funktionen und die theoretischen Voraussetzungen pädagogischer Leistungsbeurteilung, die sich ja keineswegs problemlos darstellt. Doch welcher Forschungs- oder Praxisbereich könnte dies heute schon für sich beanspruchen?

Daraus den Schluß zu ziehen, man könne auf die gebotenen — begrenzten — Möglichkeiten diagnostischer Hilfen in der Schule verzichten, wäre m. E. allerdings verhängnisvoll. Ich glaube nicht, daß die Praxis (hier: tagtäglich abgegebener Schülerbeurteilungen) besser sein kann als der jeweilige Erkenntnisstand wissenschaftlicher Forschung hierzu. Beide sind vielmehr in einem gegenseitigen Bedingungsgefüge zu sehen; paradigmatisch verweise ich in diesem Zusammenhang auf die Diskussion um die sog. Selektions- versus Klassifikationsproblematik im Rahmen schulischer Differenzierungsmodelle (vgl. HOPF 1973 sowie bereits HELLER 1970). Hier — wie so oft — sind Theorie und Praxis aufeinander angewiesen, will man sich nicht in praxisferne Modelle oder kurzschlüssige Praktiken verlieren. Dieser Überzeugung entspricht auch die Grundintention dieses Werkes.

Wir möchten unseren kurzen Überblick nicht beschließen, ohne auf zwei Desiderata hingewiesen zu haben. Zum einen wünschte man sich endlich eine umfassende *Theorie der pädagogischen Diagnostik,* die es bislang nur ansatzweise gibt (z. B. GUTHKE 1972). Zum andern dürften die Ausführungen in diesem Band die Notwendigkeit vor Augen führen, *persönlichkeits-* und *sozialpsychologische* Fragestellungen in einer solchen Theorie — stärker als in den vorliegenden Ansätzen — zu berücksichtigen (vgl. ULICH & MERTENS 1973). Das Forschungsfeld und die praktische Bewährung schulischer Leistungsbeurteilung harren weiterer Initiativen.

1.5. Literaturverzeichnis

Fischer, G. H. (Hrsg.): Psychologische Testtheorie. Huber, Bern 1968.

Fricke, R.: Über Meßmodelle in der Schulleistungsdiagnostik. Schwann, Düsseldorf 1972.

Furck, C. L.: Das pädagogische Problem der Leistung in der Schule. Beltz, Weinheim 1961, 1964 (2. Aufl.).

Gaude, P. u. *Teschner, W. P.:* Objektivierte Leistungsmessung in der Schule. Diesterweg, Frankfurt/M. usw. 1970.
Guthke, J.: Zur Diagnostik der intellektuellen Lernfähigkeit. VEB Deutscher Verlag der Wissenschaften, Berlin 1972.
Heller, K.: Aktivierung der Bildungsreserven. Huber und Klett, Bern und Stuttgart 1970.
Heller, K.: Intelligenzmessung. Neckarverlag, Villingen 1973.
Heller, K.; Rosemann, B. und *Gaedike, A.-K.:* Planung und Auswertung empirischer Untersuchungen. Klett, Stuttgart 1974.
Hentig, H. von: Systemzwang und Selbstbestimmung. Klett, Stuttgart 1968, 1970 (3. Aufl.).
Hofer, M. und *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Grundlagentexte 2 zum Funk-Kolleg Päd. Psych. Fischer Taschenbuch, Frankfurt/M. 1973.
Hopf, D.: Möglichkeiten und Grenzen der Anwendung von Tests. In: *Hofer, M.* und *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Frankfurt/M. 1973.
Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Beltz, Weinheim usw. 1971, 1972 (3. Aufl.).
Ingenkamp, K. (Hrsg.): Tests in der Schulpraxis. Beltz, Weinheim usw. 1971, 1973 (3. Aufl.).
Michel, L.: Allgemeine Grundlagen psychometrischer Tests. In: Hb. d. Psychol., Bd. 6: Psychol. Diagnostik (Hrsg. *Heiß, R.*). Hogrefe Göttingen 1964.
Neander, J.: Objektivierte Lernerfolgsmessung in der Gesamtschule — Fortschritt für wen? Die Deutsche Schule, *65* (1973), 35—47.
Rosenshine, B.: Die Beobachtung des Unterrichts in der Klasse. In: *Hofer, M.* und *Weinert, F. E.* (Hrsg.): Pädagogische Psychologie. Frankfurt/M. 1973.
Stene, J.: Einführung in Raschs Theorie der psychologischen Messung. In: *Fischer, G.* (Hrsg.): Psychologische Testtheorie. Bern 1968.
Symposium des Europarats in Berlin vom 11.—19. 11. 1971. Unveröffentl. Bericht der Arbeitsgruppe 2 (Tutor: Dr. *Weiß, R.*, Protokollant: Dipl.-Psych. *Büscher, P.*).
Ulich, D. und *Mertens, W.:* Urteile über Schüler. Beltz, Weinheim usw. 1973.
Wendeler, J.: Standardarbeiten — Verfahren zur Objektivierung der Notengebung. Beltz, Weinheim usw. 1969, 1973 (5. Aufl.).
Zielinski, W.: Die Beurteilung von Schülerleistungen. In: Funkkolleg „Pädagogische Psychologie" (Studienbegleitbrief 12). Beltz, Weinheim usw. 1973.

2. Schulleistung als pädagogisch-psychologisches Problem

Einleitender Kommentar

Wenn von ‚Schulleistung' bzw. ‚Schulleistungsbeurteilung' die Rede ist, so werden die meisten Leser zunächst an Zensuren, Schulleistungstests u. ä. denken. Zweifellos sind damit Indikatoren angesprochen, die in engem Zusammenhang zur Schülerbeurteilung stehen. Die Frage, was hinter solchen Indikatoren steht, d. h. welche Faktoren die Schulleistung(en) bedingen, ist damit freilich nicht beantwortet. Ebensowenig konnten die bisher vorgeschlagenen — mehr oder weniger begrifflich-formalen — Definitionen zur inhaltlichen Klärung der Phänomene ‚Schulleistung' beitragen.

Der hier von LANGFELDT und FINGERHUT vorgelegte Diskussionsbeitrag, in dem „anhand verschiedener faktorenanalytischer Untersuchungen eine strukturelle Beschreibung der Schulleistung versucht" wird, bereichert nicht nur die theoretische Diskussion über Formen und Inhalte schulischer Lernleistung, er vermittelt zugleich wichtige Erkenntnisse im Hinblick auf eine adäquate Schulleistungsdiagnostik. Demnach repräsentieren die traditionellen Leistungsurteile im wesentlichen „die Fähigkeit zum Erbringen geforderter Leistungen im Schulsystem".

Wenn die Autoren auch die Vorläufigkeit ihrer Ergebnisse betonen, so ist damit (wohl zum ersten Mal) der systematische Versuch geglückt, die faktorielle Struktur des Konstruktes ‚Schulleistung' modellhaft darzustellen (s. S. 42 ff.). Einer empirisch-operationalen Bestimmung dessen, was üblicherweise als Schulleistung bezeichnet wird, ist hiermit zweifellos der Weg geebnet.

Der folgende Beitrag beschäftigt sich nun mit den kognitiven und nichtkognitiven Faktoren, die schulisch geforderte Lernleistungen bedingen oder beeinflussen können. In ihrem Sammelreferat behandelt A. K. GAEDIKE die wichtigsten Untersuchungen zu diesem Themenkomplex. Neben dem Aufweis intellektueller Voraussetzungen wird dabei die Rolle zahlreicher nichtkognitiver Faktoren, z. B. Variablen der Schülerpersönlichkeit, des Lehrerverhaltens und des familiären Milieus, im Hinblick auf das Schulleistungsverhalten herausgearbeitet und ihre Bedeutung für eine Optimierung des Unterrichts unterstrichen.

Die hieraus abgeleiteten pädagogischen Schlußfolgerungen sowie zahlreiche konkrete Hinweise für eine angemessene Schulleistungsförderung dürften Lehrer und Eltern gleichermaßen interessieren. Darüber hinaus ergeben sich Konsequenzen für die Leistungs*beurteilung*, wobei exakte Lernzieldefinitionen unter dem Gesichtspunkt der Ermöglichung und einer gerechteren Bewertung von Schülerleistungen der Autorin besonders wichtig erscheinen. Die Höhe schulischer Lernleistungen ist keine ‚unipolare' Angelegenheit.

2.1. Empirische Ansätze zur Aufklärung des Konstruktes „Schulleistung"

Hans-Peter Langfeldt und Walter Fingerhut

Obwohl die Literatur zum Problemkreis „Schulleistung" bereits sehr zahlreich ist und weiter zunimmt, kann die Aussage, wonach „wir immer wieder feststellen können, daß der pädagogische Leistungsbegriff weder formal noch inhaltlich genügend definiert ist" (INGENKAMP 1967, S. 1) auch zum gegenwärtigen Zeitpunkt noch als gültig angesehen werden. Wenn auch eine verbindliche Definition des pädagogischen Leistungsbegriffes — eingeschränkt auf den Begriff „Schulleistung" — noch aussteht, so ist doch festzustellen, daß mit ihm pragmatisch umgegangen wird. In dem Maße, in dem es gelingt, das Phänomen „Schulleistung" empirisch angemessen zu beschreiben, werden die impliziten Annahmen und Voraussetzungen des Schulleistungsbegriffes deutlich. Im Gegensatz etwa zu FIPPINGER (1969) sind wir der Meinung, daß die Definition eines Konstruktes, das „Schulleistung" erklären könnte, eher über den Weg empirischer Beschreibungen zu erreichen ist als über die Interpretation einer geisteswissenschaftlichen Persönlichkeitstheorie, wie es der genannte Autor mit Hilfe der Theorie WELLEKS versucht.

Eine empirische Beschreibung dessen, was allgemein als Schulleistung verstanden wird, muß zunächst diejenigen Indikatoren berücksichtigen, von denen man üblicherweise erwartet, daß sie schulische Leistungen erfassen: Schulnoten und Schultestergebnisse. Die Zusammenhänge dieser beiden Indikatoren mit Variablen, die nachweislich etwas mit Schulleistung zu tun haben (siehe besonders die Kapitel „Determinanten der Schulleistung" und „Leistungsbeurteilung durch Notengebung" in diesem Buch) ergeben ein Strukturgefüge, dessen Beschreibung und Interpretation unter Umständen diejenigen Faktoren deutlich werden läßt, die Schulleistung im gegenwärtigen System bedingen. Aus der Diskussion dieser Faktoren könnte die Definition eines Konstruktes entstehen, das „hinter" der Schulleistung steht und sie verdeutlicht bzw. erklärt.

Zur Aufklärung einer solchen Struktur halten wir trotz methodischer Einwände (z. B. KALVERAM 1970) die Faktorenanalyse für geeignet. Sie wird hier verstanden als Methode zur Ordnung und Systematisierung einer Vielzahl von Einzelzusammenhängen. Sie macht „eine differenzierte Hypothese über die Struktur des Zueinanders der Variablen und Faktoren möglich . . ., ohne daß man vorher eine bestimmte Struktur annehmen oder bereits kennen muß" (ÜBERLA 1971, S. 3).

Im folgenden soll anhand verschiedener faktorenanalytischer Untersuchungen eine strukturelle Beschreibung der Schulleistung versucht werden, um die Diskussion im obengenannten Sinne anzuregen. Dazu stehen allerdings nur relativ wenige geeignete Untersuchungen zur Verfügung, de-

nen zudem andere Fragestellungen zugrunde lagen. Die folgende Zusammenstellung kann daher nur als vorläufig verstanden werden. Ein Anspruch auf vollständige und repräsentative Erfassung der einschlägigen Literatur wird nicht erhoben.

Der gebräuchlichste Indikator für Schulleistung ist immer noch das Schulzeugnis. DENIG & WEIS (1970) faktoranalysierten acht publizierte Korrelationsmatrizen von Schulzeugnissen der Sekundarstufe (Realschule und Gymnasium).[1] Es ergaben sich Lösungen mit mindestens drei und höchstens fünf Faktoren. Als durchgängiges Ergebnis kann festgehalten werden: Die Zeugnisse der Sekundarstufe werden durch drei stabile Faktoren charakterisiert:

1. Ein *Fremdsprachenfaktor* mit hohen Ladungen in der (den) jeweiligen Fremdsprache(n) und niedrigen Ladungen in Deutsch und Mathematik.
2. Ein *mathematisch-naturwissenschaftlicher* Faktor mit hohen Ladungen in mindestens zwei der Fächer Chemie, Mathematik, Physik und niedriger Ladung in Biologie.
3. Ein *Faktor der Sachfächer*, der durch hohe Ladungen der Fächer Erdkunde, Geschichte und Biologie gekennzeichnet ist.

In dieser durchgängigen Gruppierung sind Fächer wie Kunsterziehung, Musik und Leibeserziehung nicht vertreten. Ihre jeweilige faktorielle Zugehörigkeit kommt eher zufällig zustande.

Es liegt nahe, diese Struktur im Sinne unterschiedlicher Anforderungen der einzelnen Gymnasialfächer zu verstehen. In einer sehr differenzierten Interpretation können DENIG & WEIS jedoch deutlich machen, daß diese Betrachtungsweise nicht ausreicht.

Als weitere Einflußgrößen, welche diese Struktur bedingen, können angenommen werden: Das Fachlehrerprinzip mit typischen Fächerkombinationen (z. B. eine Fremdsprache und Deutsch oder Mathematik und Physik) sowie der Kontext, in dem Fachnoten vergeben werden. Wenn ein Lehrer in einer Klasse zwei Fächer unterrichtet, so ist anzunehmen, daß die Noten beider Fächer höher korrelieren als wenn sie von zwei Lehrern erteilt würden. Wenn eine Fachnote für die Versetzung bedeutsam ist, provoziert dies möglicherweise ein anderes Urteilsverhalten als wenn dieses Fach für die Versetzung ohne Folgen bleibt. *Die Anforderungsstruktur der Fächer oder*

[1] Literatur nach *Denig/Weis* (1970, 230–232): *Bobertag, O.:* Korrelationsstatistische Untersuchungen über die Unterrichtsleistungen der Schüler einer höheren Lehranstalt. Z. angew. Psychol. 1915, *10*, 169—187. *Höger, D.:* Analyse der Intelligenzstruktur bei männlichen Gymnasiasten der Klassen 6 bis 9 (Untersekunda bis Oberprima). Psychol. Forsch. 1964, *27*, 419—474. *Lienert, G. A.* & *Hopp, A. D.:* Über die Interkorrelationen von Gymnasialzensuren. Weinheim, Beltz, o. J. *Tent, L.:* Die Auslese von Schülern für weiterführende Schulen. Göttingen, Hogrefe, 1969. *Thyen, H.:* Über Geschlechtsunterschiede in den Schulfähigkeiten. Z. päd. Psychol. 1935, *36*, 325—355.

auch Begabungssonderheiten der Schülerpopulation reichen somit als Erklärungsfaktoren der Zeugnisstruktur nicht aus.

FUNKE (1972) analysierte die Zeugnisse von 103 Grundschülern und die letzten Volksschulversetzungszeugnisse einer parallelen Stichprobe von 103 lernbehinderten Sonderschülern. Bei den Sonderschülern zeigte sich eine dreifaktorielle Zeugnisstruktur, die durch „Schulisches Verhalten", „Schulische Leistung in den Hauptfächern" und „Schulische Leistung in den Nebenfächern" gekennzeichnet ist. Bei den Volksschülern ergaben sich dagegen nur zwei Faktoren: „Allgemeine schulische Leistung" und „Schulische Disziplin und Ordnung" (FUNKE 1972, S. 108 ff.). In einer älteren Untersuchung konnte bei den Sonderschulzeugnissen von 100 Lernbehinderten eine ähnlich einfache Struktur ermittelt werden. Es wurden „eine Dimension, die die schulische Leistung im engeren Sinne betrifft (Deutsch und Rechnen) sowie eine zweite Dimension, die sich in der Notengebung für die Kopfnoten manifestiert" festgestellt (ZIMMERMANN 1968, S. 496). An Grundschulen und an Sonderschulen für Lernbehinderte werden im allgemeinen nahezu alle Noten vom Klassenlehrer erteilt. Dies kann ein Grund sein, der die einfachere Zeugnisstruktur erklärlich macht.

Trotz der geringen Anzahl von Untersuchungen und möglicher Artefakte kann mit einiger Vorsicht festgestellt werden: *Unabhängig vom Schultyp, in dem eine Leistung zu erbringen ist, deutet sich an, daß die Schulleistung — so wie sie durch Schulnoten erfaßt wird — bei weitem nicht so differenziert ist, wie es Zeugnisse mit zehn und mehr Einzelnoten nahelegen.*

Zur Darstellung eines Konstruktes, das Schulleistung erklären könnte, reicht die Analyse von Schulnoten allein natürlich nicht aus. Wie die umfangreiche einschlägige Literatur belegt, nahm der Zusammenhang zwischen Schulleistung und Intelligenz besondere Aufmerksamkeit in Anspruch. HÖGER (1964) stellte eine der ersten faktorenanalytischen Untersuchungen vor, in denen Schulnoten und Intelligenztestergebnisse gemeinsam analysiert wurden. Bei 519 Gymnasiasten der Untersekunda bis Oberprima wurden die Noten erfaßt und ein Intelligenztest (Intelligenz-Struktur-Test [IST] von AMTHAUER) durchgeführt. Die Analyse der Korrelationsmatrix ergab folgende Faktoren (HÖGER 1964, S. 445):

„Faktor A vertritt die rationale und formale Erfassung und Verarbeitung realer Gegebenheiten der Umwelt, wie sie in guten Leistungen bei den mathematisch-naturwissenschaftlichen Schulfächern zum Ausdruck kommt.

Faktor B erweist sich als IST-Faktor, der wahrscheinlich komplexer Natur ist. Er repräsentiert die Fähigkeit, den IST zu bewältigen und entzieht sich einer präziseren Interpretation.

Faktor C repräsentiert die Fähigkeit, Vorstellungen aller Art zu bilden, zu speichern und zu reproduzieren.

Faktor D ist bipolar. Er vertritt auf dem einen Pol die sprachfreie Intelligenz, auf dem anderen die verbale Ausdrucksfähigkeit.

Faktor E erscheint als fremdsprachlicher Faktor mit einem zusätzlichen Akzent in Richtung des allgemeinen Sprachverständnisses.

Faktor F steht in enger Verwandtschaft zu MEILIS Faktor der Komplexität. Er repräsentiert die Fähigkeit, ein gegebenes Material psychisch präsent zu halten und es dabei zu strukturieren bzw. umzustrukturieren."

Bei diesen Faktoren ist zu beachten, daß die vier Faktoren A, C, D und E hoch mit den Schulnoten und B bzw. F mit den IST-Untertests geladen sind. Nach dieser Untersuchung ist also festzustellen, daß Schulleistung (wie sie durch Gymnasialzensuren erfaßt wird) relativ wenig mit Intelligenz (gemessen durch den IST) zu tun hat.

Zu einem ähnlichen Ergebnis bei Grundschülern gelangt JANTZEN (1971), der die bekannten Untersuchungen von KEMMLER (1967) replizierte, wobei er in die Faktorenanalyse neben Intelligenztestvariablen auch Schulleistungsvariablen mit aufnahm. Sowohl bei schlechten als auch bei guten Grundschülern des dritten Schuljahres fand sich eine recht deutliche Struktur von elf Faktoren (diese hohe Faktorenzahl wird durch die relativ vielen unterschiedlichen Einzelvariablen verständlich). Die Schulleistungsvariablen laden jedoch nur auf drei der elf Faktoren. Die Interpretation JANTZENS, die Notenunterschiede guter und schlechter Schüler seien daher „überhaupt nicht" auf Leistungsunterschiede zurückzuführen, ist aber wohl zu weitgehend.

Eine unveröffentlichte Faktorenanalyse (FINGERHUT & LANGFELDT 1971), die auf Daten (Schulnoten, Schultest- und Intelligenzleistungen sowie biographischen Informationen) von 1756 Grundschülern der vierten Klasse beruht, erbrachte vier Faktoren, die wie folgt charakterisiert werden können:

1. *„Schulische Leistungsfähigkeit":* Über 0,30 laden auf diesem Faktor die Schulnoten (Deutsch, Rechnen, Heimatkunde), sämtliche Untertests des verwendeten Schultests (Allgemeiner Schulleistungstest für 4. Klassen [AST 4] von FIPPINGER) sowie die Untertests „logisches Denken" und „Erkennen von Regeln" des verwendeten Intelligenztests (Leistungsprüfungsystem [LPS] von HORN).
2. *„Genereller Notenfaktor":* Auf diesem Faktor laden (über 0,30) außer den genannten Fachnoten die sogenannten Kopfnoten (Betragen, Fleiß, Aufmerksamkeit, Ordnung) und die biographischen Daten der Schüler (Alter und Geschlecht).
3. *„Test-Intelligenz":* Auf diesem Faktor laden (über 0,30) ausnahmslos die verwendeten Untertests des LPS.
4. *„Rechenleistungen":* Auf diesem Faktor laden (über 0,30) die Rechennote, die entsprechenden Rechen-Untertests des AST 4 und die LPS-Untertests zum „logischen Denkvermögen", zur „Wahrnehmungsgenauigkeit" und zum „Kopfrechnen".

Schulleistungsindikatoren (Noten und Schultests) setzen sich demnach doch nicht so eindeutig von Begabungsindikatoren (Intelligenztests) ab (siehe

Faktoren 1 und 4), wie es die vorhergehenden Untersuchungen vermuten ließen. In den erwähnten Arbeiten wurden in der Regel nur Schulnoten als Indikatoren verwendet. Trotzdem war bisher gesagt worden, daß der zusammenhang zwischen Schul*leistung* und Intelligenz gering sei. Es kann jedoch nur gesagt werden, daß Schul*noten* und Intelligenz relativ unabhängig voneinander sind. Die Ergebnisse dieser Analyse zeigen: *Schulnoten können nicht selbstverständlich als adäquates Abbild der Schulleistung angesehen werden.* Untersuchungen, die als Schulleistungsindikatoren ausschließlich Schulnoten verwenden, können daher nur in mäßigem Umfange etwas über die *tatsächlichen* Bedingungen der Schulleistung aussagen.

Bestimmte statistische Kriterien (Höhe der Kommunalitäten und Eigenwertabfall) legen den Schluß nahe, daß die dargestellte Analyse von Fingerhut & Langfeldt überstrukturiert ist. Die Daten wurden daher einer neuerlichen Faktorenanalyse unterzogen. Die graphische Darstellung zeigt eine zweifaktorielle Lösung.

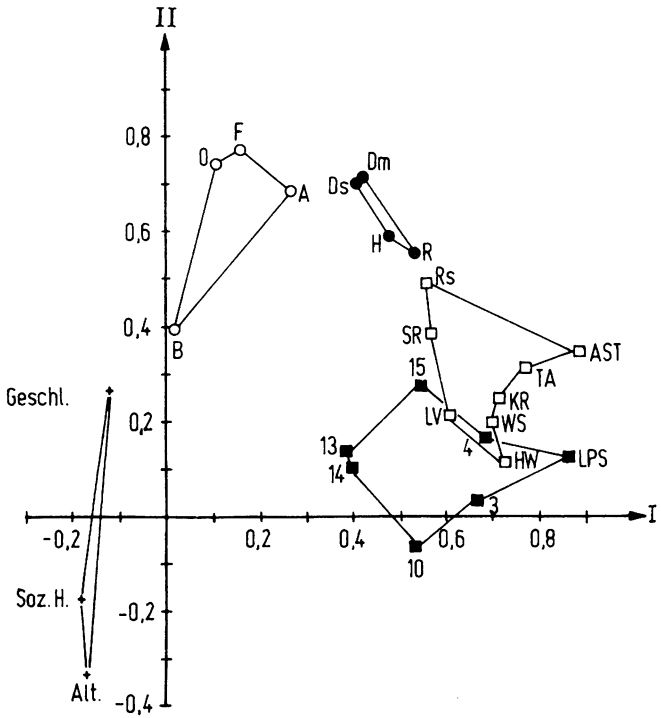


Abb. 1: Faktorielle Struktur des Konstruktes "Schulleistung" (nach Daten von FINGERHUT & LANGFELDT 1971)

Durch entsprechende Verbindungslinien werden fünf Komponenten deutlich:

1. Intelligenztestleistungen im LPS (HORN 1962)
   —■— mit der Nummer der Untertests:

   3: Regelerkennen, logisches Denken (reasoning)
   4: Regelerkennen, logisches Denken (reasoning)
   10: Erkennen des Wesentlichen (closure)
   13: Wahrnehmungstempo (perceptual speed)
   14: Bemerken von Fehlern (accuracy)
   15: Schnelles Addieren (number)
   LPS: Gesamtpunktwert aus den verwendeten Untertests

2. Schultestleistungen im AST4 (FIPPINGER 1966)
   —□— mit den Abkürzungen der Untertests:

   LV: Leseverständnis
   WS: Wortschatz
   KR: Kopfrechnen
   SR: Schriftliches Rechnen
   TA: Textaufgaben
   RS: Rechtschreibung
   HW: Heimatkundliches Wissen/Kartenverständnis
   AST: Gesamtpunktwert aus allen Untertests

3. Fachnoten
   —●— mit den entsprechenden Abkürzungen:

   Dm: Deutsch mündlich
   Ds: Deutsch schriftlich
   R: Rechnen
   H: Heimatkunde

4. Kopfnoten
   —○— mit den entsprechenden Abkürzungen

   B: Betragen
   F: Fleiß
   A: Aufmerksamkeit
   O: Ordnung

5. Biographische Daten
   -+- mit den entsprechenden Abkürzungen
   Alt.: Alter
   Geschl.: Geschlecht
   soz. H.: soziale Herkunft

In der Abbildung können deutlich fünf abgrenzbare Cluster unterschieden werden:

1. Testleistungen im Intelligenztest (LPS von HORN, 1962),
2. Testleistungen im Schultest (AST 4 von FIPPINGER, 1966),
3. Fachnoten,
4. Kopfnoten,
5. biographische Daten.

Die Fachnoten und die Schultestleistungen liegen zwischen dem Cluster der „Begabung" und dem der Kopfnoten. Kopfnoten können als „Wohlverhalten und Anpassung an das System Schule" interpretiert werden.

Aus der Distanz des Notenclusters und des Schultestclusters zu dem Intelligenztestcluster einerseits und dem Cluster der Kopfnoten andererseits kann man — etwas akzentuiert — schließen: *Schulleistung im gegenwärtigen Schulsystem ist eine Funktion von tatsächlicher Leistung aufgrund von Begabung* (Faktor I) *und Anpassung an dieses System* (Faktor II).

Diese Interpretation wird gestützt durch die faktorenanalytischen Untersuchungen von SEITZ & LÖSER (1969) an 17jährigen Gymnasiasten, von SEITZ (1970) an Schülern des dritten Grundschuljahres und SEITZ (1971) an Volksschülern der sechsten Klasse. In diese Analysen wurden neben Intelligenztestergebnissen und Schulnoten auch Persönlichkeitstestergebnisse mit aufgenommen. Durchgehend ergab sich innerhalb der multifaktoriellen Lösungen jeweils ein „Notenfaktor", auf dem jeweils auch solche Persönlichkeitsmerkmale hohe Ladungen zeigten, die mit sozialer Anpassung in Zusammenhang stehen. Bei den Gymnasiasten laden auf dem Notenfaktor die Merkmale „Zurückhaltung und Isolierung", „Unsicherheit über eigene Leistungsfähigkeit" und „egozentrische Autonomie" (gemessen durch das High School Personality Questionnaire [HSPQ] von CATTELL). Bei den Schülern des dritten Schuljahres waren es „Erwartung von Leistungsmißerfolg", „Selbstüberzeugung" und „pessimistisch-ängstliche Depression" (gemessen durch das Early School Personality Questionnaire [ESPQ] von COAN & CATTELL). Bei den Schülern des sechsten Schuljahres lagen auf dem Notenfaktor die Variablen „Selbstunterschätzung—Selbstschätzung", „optimistische Unbekümmertheit — pessimistische Vorsicht" und „Unsicherheit im sozialen Umgang — Selbstsicherheit im sozialen Umgang" (gemessen durch das Children's Personality Questionnaire [CPQ] von PORTER & CATTELL).

Es steht außer Frage, daß die referierten Untersuchungen noch keine genaue oder endgültige Beschreibung des Konstruktes „Schulleistung" gestatten. Wir wollen unseren Beitrag daher in erster Linie als Diskussionsbeitrag und Anregung verstanden wissen. Als vorläufige Zusammenfassung kann festgehalten werden: Schulleistung sollte auf dem Hintergrund eines Konstruktes erklärt werden, das vorläufig mit „Fähigkeit zum Erbringen geforderter Leistungen im Schulsystem" umschrieben werden muß. Es be-

steht Grund zu der Annahme, daß im gegenwärtigen System Schulleistung durch Begabungsaspekte und die Fähigkeit zur sozialen Anpassung bedingt ist.

Literaturverzeichnis

Denig, F. u. *Weis, V.:* Zur Faktorenstruktur der Fachnoten an deutschen Sekundarschulen. ZeF, 1970, *4*, 210—232.

Fingerhut, W. u. *Langfeldt, H. P.:* Schülermerkmale, Lehrermerkmale und ihre Beziehungen zu Schulnoten. Unveröffentlichte Diplomarbeit, Marburg, 1971.

Fippinger, F.: Allgemeiner Schulleistungstest für vierte Klassen (AST 4). Weinheim, Beltz, 1966.

Fippinger, F.: Zum Problem der Schulleistungsdiagnostik: Lehrerurteil und Schulleistungstest. Päd. Rdschau, 1969, *23*, 869—880.

Funke, E. H.: Grundschulzeugnisse und Sonderschulbedürftigkeit. Berlin, Marhold, 1972.

Höger, D.: Analyse der Intelligenzstruktur bei männlichen Gymnasiasten der Klassen 6 bis 9 (Untersekunda — Oberprima). Psychol. Forsch., 1964, *27*, 419 bis 474.

Horn, W.: Leistungsprüfsystem (LPS). Göttingen, Hogrefe, 1962.

Jantzen, W.: Untersuchungen zur Faktorenstruktur von Intelligenz und Schulleistungen bei guten und schlechten Schülern dritter Grundschulklassen (und zur Einstellungsstruktur ihrer Lehrer). ZeF, 1971, *5*, 44—62 und 92—106.

Kalveram, K. T.: Über Faktorenanalyse. Kritik eines theoretischen Konzepts und seine mathematische Neuformulierung. Arch. ges. Psychol., 1970, *118*, 92—118.

Kemmler, Lilly: Erfolg und Versagen in der Grundschule. Göttingen, Hogrefe, 1967.

Seitz, W.: Über den Zusammenhang von Persönlichkeitseigenarten, Schulnoten und HAWIK-Leistungen bei Volksschülern. Psychol. Beitr. 1970, *13*, 579—602.

Seitz, W.: Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztestleistungen bei Volksschülern. Z. exp. angew. Psychol., 1971, *18*, 307—336.

Seitz, W. u. *Löser, G.:* Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztestleistungen bei Volksschülern. Z. exp. angew. Psychol., 1969, *16*, 651—679.

Überla, K.: Faktorenanalyse. Berlin, Springer, 1971 [2].

Zimmermann, K. W.: Über die Beziehungen zwischen Schulnoten von lernbehinderten Sonderschülern. Z. f. Heilpäd., 1968, *19*, 465—471.

2.2. Determinanten der Schulleistung

Anne-Katrin Gaedike

Mit diesem Beitrag soll eine Übersicht über Forschungsarbeiten zur Frage der Schulleistungsdeterminanten gegeben werden. Zu diesem Zweck ist im vorhinein die Klärung der beiden Begriffe „Schulleistung" und „Determinante" erforderlich. Für „Schulleistung" findet sich in der einschlägigen Literatur weder eine verbindliche Definition noch eine einheitliche Verwendung des Begriffs. Zwischen Schulleistung, Schulerfolg und Schulzensur wird z. B. häufig nicht exakt unterschieden, so daß die Vergleichbarkeit der dargestellten Untersuchungsergebnisse nicht ohne weiteres gewährleistet ist. Wenn mit „Leistung" jedes Resultat einer gezielten Aktivität gemeint ist, so kann man unter „Schulleistung" umfassend die Summe derjenigen Leistungen verstehen, die unter schulischen Bedingungen erbracht werden. In der Praxis werden jedoch lediglich solche Schüleraktivitäten als Leistungen anerkannt, die auf ein bestimmtes, durch den Lehrplan fixiertes Ziel hin ausgerichtet sind. In diesem Zusammenhang sei auf den vorstehenden Artikel von LANGFELDT & FINGERHUT verwiesen, der sich ausführlich mit dem Konstrukt „Schulleistung" befaßt. Wir wollen versuchen, im folgenden der Klarheit halber zu allen zitierten Untersuchungen das Schulleistungsverständnis des betreffenden Autors oder zumindest seine Methode der Schulleistungserfassung im Sinne einer quasi operationalen Definition mit anzugeben.

„Determinante" bedeutet nach FRÖHLICH (1968, S. 55) soviel wie „Inbegriff der Ursache für das Auftreten eines Phänomens, so wie es hier und jetzt erscheint". Da von Ursachenforschung in der pädagogischen Psychologie weitgehend noch nicht gesprochen werden kann und man sich bisher zur Annahme von Ursache-Wirkungs-Zusammenhängen lediglich auf das Auffinden korrelativer Beziehungen beschränkt hat, sollen im Rahmen dieses Artikels alle Variablen als mögliche Determinanten der Schulleistung aufgeführt werden, die gesetzmäßig mit der Variation der Schulleistung in Verbindung stehen. Aus dieser Vorgehensweise ergibt sich, daß dabei nicht exakt getrennt werden kann zwischen „verursachenden" Determinanten und „enthaltenen" Faktoren. Intelligenz z. B. könnte sowohl ein Faktor der Schulleistung sein im Sinne von „Teil" als auch eine Determinante im Sinne von „Bedingung".

Wir erörtern zunächst die kognitiven Determinanten und anschließend verschiedene nicht-kognitive. Eine Hierarchie der Bedeutsamkeit ist durch die gewählte Reihenfolge nicht impliziert.

2.2.1. Kognitive Faktoren der Schulleistung

2.2.1.1. Korrelative Beziehungen zwischen verschiedenen Intelligenztests und der Schulleistung

Zur Frage der Beziehung zwischen Schulleistung und Intelligenz, d. h. ob „gute" Schüler auch zugleich „intelligente" Schüler sind und umgekehrt, wird bereits seit Beginn dieses Jahrhunderts geforscht. Und noch immer ist man zu keinem eindeutigen Ergebnis gekommen, vermutlich deshalb nicht, weil es an einer allgemeinverbindlichen Definition sowohl für Schulleistung als auch für Intelligenz mangelt (siehe hierzu auch LÖSCHENKOHL 1973). Intelligenztests, die eher schulnahe Aufgaben enthalten, korrelieren beispielsweise höher mit der Schulleistung als andere; bei solchen, die einen Verbal- und einen Handlungsteil haben, steht der erstere in wesentlich engerem Zusammenhang mit der Schulleistung als der zweite. Wird die Schulleistung testmäßig erfaßt, ergeben sich deutlich höhere Zusammenhänge. Außerdem nimmt das Ausmaß des Zusammenhangs mit zunehmendem Lebensalter der Schüler ab (WEISS 1964). In Schultypen mit höheren schulischen Anforderungen sind die Zusammenhänge zwischen Schulleistung und Intelligenz niedriger als in Schultypen mit geringeren schulischen Anforderungen (WEISS 1969, TENT 1969).

Eine Zusammenstellung der Arbeiten vor 1930, die alle recht hohe Korrelationskoeffizienten für den Zusammenhang zwischen Intelligenz und Schulleistung ermittelten ($r = .49$ bis $r = .98$), ist bei FIPPINGER (1966) zu finden. Diese frühen Untersuchungen von BURT, LOBSTEIN, STERN, KESSELRING, MANIG, ROLOFF, SANDER und LÄMMERMANN stützen sich jedoch durchweg auf wenig gesicherte Methoden sowohl diagnostischer als auch experimentell-statistischer Art. Spätere, exaktere Arbeiten bringen im Durchschnitt niedrigere, jedoch sehr unterschiedliche, von den jeweiligen Testverfahren abhängige Ergebnisse hervor. GUILFORD (1965) berichtet von „typischen" Korrelationen zwischen Testwerten der verbalen Intelligenz und den Schulzensuren von $r = .50$ bis $r = .70$. Im einzelnen stellen sich die Ergebnisse der neueren Untersuchungen wie folgt dar:

STELZL (1949) ermittelte in seiner Untersuchung, die allerdings auf so wenig leistungsfähigen Tests wie dem Düsseldorfer Lückentest, dem Fabeltest, dem Besorgungstest und der MASSELONschen Probe beruht, Korrelationen zwischen Intelligenz- und Leistungsrang von durchschnittlich $r = .52$. Im Rahmen der Validitätsprüfung zu seinem Intelligenz-Strukturtest (IST) berechnete AMTHAUER (1953) zwischen dem IST und der Schulleistung einen Zusammenhang von $r = .46$; als Maß für die Schulleistung galt dabei das arithmetische Mittel aus allen Zeugniszensuren. Bei 287 Gymnasiasten der Unterstufe stellte STEINHAUSER (1957) nur einen Zusammenhang von

r=.22 fest, wobei die Intelligenz mit dem HAWIE-Verbalteil gemessen und das Schulleistungsmaß aus dem arithmetischen Mittel von Deutsch, Latein, Englisch und Mathematik bestimmt wurde. CLOSTERMANN (1959) und ZILER (1959) berichten über Korrelationen zwischen dem Mann-Zeichen-Test nach GOODENOUGH und der Schulleistung um r=.45, wobei allerdings zu bedenken ist, daß der Mann-Zeichen-Test eher ein Entwicklungstest als ein Intelligenztest ist. BEER (1960) ermittelte einen Zusammenhang zwischen Intelligenz und Schulleistung von r=.40 für den ersten Klassenzug der österreichischen Hauptschule (Mittelschule) und von r=.45 für den zweiten Klassenzug (Hauptschule). Als Intelligenzmaß verwandte er den Verbalteil des HAWIE und als Maß für die Schulleistung stand das Notenmittel aus Deutsch und Mathematik. EWERT (1960) berechnete eine multiple Korrelation zwischen allen Schulnoten und Ergebnissen aus dem Progressiven Matrizen-Test von RAVEN und erhielt dabei einen Gesamt-Koeffizienten von r=.59. Die größte Beziehung besteht nach dieser Untersuchung zwischen den Intelligenztestergebnissen und der Zensur im Rechnen, nämlich r=.49. MEILI (1961) erhielt Korrelationen um r=.40 (von r=.03 bis r=.79) zwischen seinem Analytischen Intelligenztest (AIT) und den nicht näher definierten Schulleistungen. BURGER (1963) ermittelte zwar einen relativ hohen Zusammenhang zwischen dem IST und der Schulleistung (von r=.50 bis r=.81), korrigiert sich jedoch 1965, indem er diese Beziehung keinesfalls für die Extremgruppe der Hochbegabten zutreffend feststellt. 23,8 % der Hochbegabten seiner Stichprobe hatten bis zum Erreichen der vierten bzw. fünften Klasse der höheren Schule eine Klasse bereits einmal wiederholt, also keine ausreichenden Schulleistungen erbracht. Korrelationskoeffizienten zwischen Intelligenz- und Schulleistungswerten wurden für diese Gruppe jedoch nicht berechnet. Der von HÖGER (1964) eruierte Zusammenhang zwischen IST-Werten und Schulnoten schwankt zwischen r=.06 und r=.35, liegt also besonders niedrig. HÖGER interpretiert dieses Ergebnis dahingehend, daß er vermutet, die beiden Merkmalssysteme (IST und Schulnoten) ließen „weitgehend unterschiedliche psychische Gegebenheiten zur Auswirkung kommen". Für 148 Volksschüler der 4. Klasse bestimmte KOHL (1964) das Ausmaß der Intelligenz mit dem Hamburg-West-Yorkshire-Test (HWY) und fand Korrelationen zur Deutschnote von r=.60 und zur Rechennote von r=.53. Trotz dieser relativ hohen Korrelationen gelangte er zu der Feststellung, daß 32 % der Kinder mit guten oder befriedigenden Ergebnissen im Intelligenztest schlechte Schulleistungen auswiesen. Die Interpretation dieser diskrepanten Befunde überläßt KOHL dem Leser. SCHMITZ (1964) konnte an 124 Klassen des 4. Grundschuljahres in Nordrhein-Westfalen zwischen Intelligenz und Schulleistungsniveau eine Korrelation von r=.54 feststellen. Als Intelligenzmaß diente ihm ebenfalls der HWY; die Schulleistung wurde mit einem Diktat und zwei Rechenarbeiten der „Soltauer

Arbeitsgemeinschaft für Leistungsmessung in der Volksschule" erfaßt. Die Korrelationen zwischen dem von WEISS (1964) willkürlich gewogenen Notenmittel und verschiedenen Untertests sowie dem Gesamtergebnis des HAWIK variieren von r=.007 bis r=.54. Es zeigt sich dabei ein deutlicher Unterschied zwischen Verbal- und Handlungsteil zugunsten eines engeren Zusammenhangs zwischen dem HAWIK-Verbalteil und der Schulleistung. Ebenso deutlich wurde der Unterschied hinsichtlich der Korrelation zwischen erstem und zweitem Klassenzug der österreichischen Hauptschule. Demnach wird die Schulleistung in dem unserer Hauptschule entsprechenden zweiten Klassenzug eher durch die testmäßig erfaßte Intelligenz bedingt als in dem unserer Mittel- oder Realschule vergleichbare ersten Klassenzug. Zwischen dem Leistungs-Prüf-System (LPS) von HORN und dem arithmetischen Mittel aller Zeugniszensuren ergab sich in der Untersuchung von HAHN (1965) an 173 10- bis 15jährigen Hauptschülern ein Zusammenhang von r=.53. TENT (1965), der 942 Volksschüler der 4. Klasse mit dem LPS von HORN untersuchte und ihre Testwerte mit den Noten des 2. bis 4. Schuljahres korrelierte, fand entsprechende Zusammenhänge von r=.47 bis r=.55. An einer Gruppe verhaltensgestörter oder milieugeschädigter 8 bis 14 Jahre alter Kinder interessierte BLÖSCHL (1966) der Zusammenhang zwischen der letzten Zeugnisnote und den Werten im Begabungs-Test-System (BTS) von HORN und im HAWIK. Im ersten Fall ergab sich ein Korrelationskoeffizient von r=.38, im zweiten Fall von r=.34. FIPPINGER (1966) kam in seiner ausführlichen Untersuchung an 86 Volksschülern im Alter von 9 bis 12 Jahren (4. Schuljahr) auf Korrelationskoeffizienten von r=.54 für die Beziehung zwischen HAWIK-Werten und dem Schulleistungstest HI 19, von r=.32 für den Zusammenhang zwischen HAWIK-Werten und dem Lehrerurteil sowie auf einen Koeffizienten von r=.64 für die Beziehung zwischen BTS-Werten und dem Lehrerurteil. Dabei wies jeweils der Verbalteil des HAWIK eine deutlich größere Relevanz für die Diagnose der Schulleistung auf als der Handlungsteil. In der Untersuchung von SCHNELL (1966) — zit. nach BEER, KUTALEK, SCHNELL (1968) — ergab sich für die 6. Stufe aller Schultypen ein durchschnittlicher Korrelationskoeffizient zwischen Intelligenz und Schulleistung von r=.44. Die Intelligenz wurde auch in dieser Arbeit mit dem BTS gemessen, die Schulleistung durch das arithmetische Mittel aus Deutsch, Mathematik und Fremdsprachen bestimmt. WEBER (1966) erfaßte bei einer Gruppe von 71 Schülern des 5. und 6. Schuljahres Rangkorrelationskoeffizienten von r=.74 und r=.89 für den Zusammenhang zwischen dem Intelligenzquotienten im HAWIK und den Noten im Rechnen und Rechtschreiben. Diese extrem hohen Werte können aufgrund der methodischen Mängel der Arbeit zustande gekommen sein: Weder die Intelligenzquotienten noch die Noten für die Testarbeiten waren in der Stichprobe normal verteilt.

Für Schüler des naturwissenschaftlichen Zweiges der Oberschule errechnete TODT (1967) in seiner „Untersuchung zur Vorhersage von Schulnoten" einen signifikanten Zusammenhang zwischen der durchschnittlichen Schulnote und dem Intelligenz-Standardwert des WILDE-Intelligenz-Tests (WIT) von r=.37. Für Schüler des sprachlichen Zweiges ergab sich bemerkenswerterweise kein signifikanter Zusammenhang. Bei einem Extremgruppenvergleich von „guten" und „schlechten" Schülern konnte KEMMLER (1967) aufzeigen, daß die Intelligenz — gemessen als normalisierter Mittelwert von 7 Einzelvariablen des Primary-Mental-Ability-Tests (PMA) von THURSTONE — die beiden Gruppen ebensogut trennt wie ein Diktat. Betrachtet man jedoch die Streuung der Gesamtstandardwerte in dieser Untersuchung, so ergibt sich für die 5 % besten Schüler ein Bereich von 91 bis 129 IQ-Werten und für die 12 % schlechtesten Schüler ein Bereich von 72 bis 113 IQ-Werten. M. a. W.: Die Lehrerurteile in bezug auf „gute" und „schlechte" Schüler ermöglichen zwar eine globale Extremgruppenklassifizierung hinsichtlich der Intelligenz, für den einzelnen Schüler jedoch ist die Klassifizierung nach dem Lehrerurteil sehr unsicher (siehe auch HELLER 1970, S. 168 f.). Schlechte Schüler sind zwar darin gleich, daß sie Mißerfolg in der Schule haben, dieser kann jedoch auf verschiedene Ursachen zurückzuführen sein und nicht allein auf eine mangelnde Intelligenz. Die umfangreiche Untersuchung von WEISS (1967) an 3 967 Vierzehnjährigen verschiedener Schultypen läßt einen Zusammenhang zwischen LPS-Werten und dem von WEISS willkürlich gewogenen Notenmittel zwischen r=.21 und r=.33 erkennen. Auch in dieser Untersuchung — wie schon in der von 1964 — ergaben sich höhere Zusammenhänge bei Volksschülern gegenüber Gymnasiasten. Wesentlich höhere Zusammenhänge konnten ermittelt werden, wenn die Schulleistung durch standardisierte Leistungstests bestimmt wurde. Die Korrelationskoeffizienten lagen dann durchweg zwischen r=.28 und r=.53 und waren alle signifikant. BEER, KUTALEK & SCHNELL (1968) bestimmten einen relativ niedrigen mittleren Zusammenhang zwischen Intelligenz (nach dem BTS von HORN) und Schulleistung von r=.44 und wiesen damit in ihrer Untersuchung auf die Bedeutsamkeit der Aufklärung weiterer Faktoren hin, die die Varianz der Schulleistungen mitbestimmen. Der von HELLER & SCHIRMER (1973) entwickelte Wortschatztest für Sehbehinderte (WST [Sb]) korreliert in den Klassen 4 bis 9 bei verschiedenen Schulzensuren (Rechnen, Aufsatz, Deutsch) zwischen r=.16 und r=.61; wenn eine Korrektur für die mangelnde Reliabilität der Schulzensuren in die Berechnung mit einbezogen wird, variieren die Koeffizienten von r=.3 bis r=.9. Die besonders hohen Korrelationen ergeben sich aus der Konzeption des Tests als reinem Verbal-Intelligenz-Test, der naturgemäß einen engen Bezug zu verbalen Schulleistungen hat. Die Korrelationskoeffizienten für die Beziehung

zwischen WST (Sb) und der Zensur im Rechnen liegen im unteren des angegebenen Bereichs, r=.16 bis r=.43 (korr. r=.3 bis r=.6).

Übersicht der verschiedenen Untersuchungsergebnisse zur Korrelation zwischen Schulleistung und Intelligenz

Autor	Jahr	r	N
STELZL	1949	.52	82
AMTHAUER	1953	.45	350
STEINHAUSER	1957	.22	287
CLOSTERMANN; ZILLER	1959	.45	126
BEER	1960	.40	292
EWERT	1960	.59	163
MEILI	1961	.03 bis .79	200
BURGER	1963	.50 bis .81	189
HÖGER	1964	.06 bis .35	519
KOHL	1964	.53 bis .60	148
SCHMITZ	1964	.54	3600
WEISS	1964	.01 bis .40	581
HAHN	1965	.53	173
TENT	1965	.47 bis .55	942
BLÖSCHL	1966	.34 bis .38	125
FIPPINGER	1966	.32 bis .64	86
SCHNELL	1966	.44	1397
WEBER	1966	.74 bis .89	71
TODT	1967	.37	208
WEISS	1967	.28 bis .53	3967
BEER, KUTALEK & SCHNELL	1968	.44	1264
HELLER & SCHIRMER	1973	.16 bis .61	606
	(korrigiert:	.30 bis .90	

Die durchweg positiven Korrelationen zwischen Maßen der Schulleistung und der Intelligenz der aufgeführten Untersuchungen belegen, daß ein wesentlicher Faktor der Schulleistung (was immer das auch sei) die testmäßig erfaßte Intelligenz ist. Eine Schwierigkeit bei Interpretation und Vergleich dieser Untersuchungsergebnisse liegt darin, daß die einzelnen Autoren verschiedene Schulleistungsmaße verwenden. Auch das häufig verwendete Notenmittel läßt keine eindeutige Interpretation zu, da die einzelnen Noten unterschiedlich gewichtig sind für den Schulerfolg an verschiedenen Schulen (eine allgemeinverbindliche Gewichtung wird sich kaum finden lassen, da die Schulen verschiedene Schwerpunkte für die Ausbildung festsetzen) und es außerdem auch etwas anderes bedeutet, „ob ein Schüler durchweg mittel-

mäßige Leistungen zeigt, oder ob er sehr gute und sehr schlechte Noten nebeneinander hat" (SUERMANN 1971, S. 45). Für die Schullaufbahnberatung bzw. für Auswahlverfahren zum Übertritt auf weiterführende Schulen ist es bedeutsam, daß Intelligenztests für einen Zeitraum von einem bis zu sechs Jahren den Erfolg auf weiterführenden Schulen mit einer Genauigkeit voraussagen, die nach GEBAUER (1965) und HITPASS (1963) zwischen $r=.40$ und $r=.65$ sowie nach BURGER (1963) zwischen $r=.50$ und $r=.81$ schwankt. Dagegen liegen die Erfolgskoeffizienten für Noten aus schriftlichen Aufnahmeprüfungen erheblich niedriger.

Trotzdem sollten die Schulleistungsprädikatoren „Intelligenztests" nicht überbewertet werden. JANSSEN (1970) entwickelt dazu folgenden interessanten Gedankengang: Er geht in seinen Überlegungen von der theoretischen Beziehung zwischen Reliabilität und Validität eines Prädiktors (P) und eines Kriteriums (K) nach MAGNUSSON (1969) aus:

$$r_{PK(emp)} = r_{PK(wahr)} \cdot \sqrt{r_{PP} \cdot r_{KK}}$$

Das heißt, die geschätzte empirische Korrelation zwischen Prädiktor und Kriterium $[r_{PK(emp)}]$ ist das Produkt aus der „wahren" Beziehung zwischen Prädiktor und Kriterium $[r_{PK(wahr)}]$ und dem geometrischen Mittel der Reliabilitäten der beiden Meßwertreihen. Die Schwierigkeit, daß der „wahre" Zusammenhang praktisch eine unbekannte Größe ist, kann dadurch behoben werden, daß man verschiedene „wahre" Zusammenhänge definiert und unter diesen Bedingungen die Erwartungswerte unterschiedlich reliabler Prädiktoren miteinander vergleicht.

Auf diese Weise stellte JANSSEN fest, daß die prognostische Korrelation zwischen Intelligenztest und Kriterium (Schulerfolg) mit Notwendigkeit genauer ist als die Korrelation zwischen dem Prädiktor „Schulnote" und dem Kriterium, wenn man sich auf die in der Literatur angegebenen Reliabilitätsschätzungen für Schulzensuren und Intelligenztests bezieht. Der Unterschied der prognostischen Genauigkeit der Prädiktoren Intelligenztest und Schulnote ist jedoch nach UNDEUTSCH (1969) wesentlich deutlicher, als nach JANSSEN theoretisch erwartet werden darf. JANSSEN erklärt diese Diskrepanz zwischen Theorie und Praxis damit, daß die Intelligenzwerte (z. B. von GEBAUER & HITPASS) erst erhoben wurden, nachdem die Auslese der Ungeeigneten bereits abgeschlossen war. Durch diese unterschiedliche Prozedur haben die beiden Prädiktor-Datenmengen „Noten-der-Aufnahmeprüfung" und „Intelligenztestwerte" ihre statistische Gleichwertigkeit eingebüßt, da die Streuungsreduktion der beiden Prädiktoren unterschiedlich ist. Nach GUILFORD (1965) bedeutet das, daß allein schon wegen seiner geringeren Streuungsreduktion der indirekte Prädiktor (hier: Intelligenztestwerte) enger mit dem Kriterium korrelieren muß als der direkte Prädiktor (hier:

Noten der Aufnahmeprüfung). Um also die prognostische Genauigkeit des schulpädagogischen Prädiktors mit der des psychometrischen zu vergleichen, muß man die unterschiedliche Streuungsreduktion korrigieren (LIENERT 1967), was jedoch in bisherigen Vergleichsuntersuchungen noch nicht geschehen ist. Aus diesem Grunde wurde seither die tatsächliche prognostische Effizienz der klassischen schulpädagogischen Selektionsverfahren unterschätzt.

Ein weiterer von JANSSEN dargelegter Gesichtspunkt bei Auswahlverfahren für weiterführende Schulen ist der einer optimalen Entscheidungsstrategie nach TAYLOR & RUSSEL (1939) oder CRONBACH & GLESER (1965) als optimale Verknüpfungen oder Kombination quantitativer Entscheidungsinformationen (Validität eines Tests; Quote der a priori Erfolgreichen; Quote der Erfolgreichen nach Anwendung des Tests; Quote derjenigen, die ausgelesen werden sollen; Kosten, die bei der Beschaffung neuer Information — z. B. zusätzlicher Tests — entstehen).

Um nun wenigstens 80 % erfolgreiche Real- und Oberschüler zu prognostizieren, könnte man auf zweierlei Weise vorgehen: Entweder man entwikkelt einen Test, der mit dem Kriterium „Schulerfolg" r=.90 korreliert, oder man greift auf einen weniger validen Test zurück und hält die Selektionsquote sehr klein. In der Praxis sähe das z. Z. so aus, daß man bei einer Quote a priori Erfolgreicher von 30 % und einer Testvalidität von ca. r=.60 nur 5 % aller Bewerber nach TAYLOR & RUSSEL (1939) in weiterführende Schulen aufnehmen dürfte, sofern man die Quote der Erfolgreichen von 30 % auf 80 % steigern möchte (da es einen Test, der mit dem Schulerfolg mit r=.90 korreliert, nicht gibt). Siehe hierzu auch die Diskussion um selektive Fehlentscheidungen in der pädagogischen Diagnostik zwischen KORNMANN (1972) und MANDEL & KRAPP (1972). Man wird daher eine Lösung suchen müssen, die weniger ökonomisch, dafür aber realisierbar ist, z. B. wie in England (s. YATES 1957, PIDGEON 1962, PEEL 1962, National Foundation for Educational Research 1963) schulpädagogische und psychometrische Prädiktoren gemeinsam berücksichtigen, um die Wahrscheinlichkeit einer richtigen Entscheidung zu erhöhen. Es müßte dabei allerdings sicher sein, daß alle benutzten Prädiktoren verschiedene, gleichermaßen relevante Information für die Prognose erfassen. Das ist der Fall zumindest für Schulzensuren und Intelligenztestwerte, die nur in mittlerer Höhe von .50 bis .70 miteinander korrelieren (nach GUILFORD). Die Frage ist, ob man bei der Kombination der verschiedenen Prädiktoren nach einem kompensatorischen Modell (d. h. niedrige Werte in einem Prädiktor können durch hohe Werte in einem anderen kompensiert werden) vorgeht oder nach einem disjunktiven (d. h. für jeden Prädiktor müssen festgelegte Mindestwerte erreicht werden). Wir wollen im Rahmen dieses Beitrages nicht näher auf die Diskussion um diese Frage eingehen.

Zusammenfassend läßt sich sagen, daß der Schulerfolg unter den augenblicklichen Bedingungen keinesfalls allein aufgrund eines einzigen Kriteriums diagnostiziert oder prognostiziert werden darf, da keines der z. Z. gebräuchlichen Schulleistungskriterien allein ausreichend genau ist. Bevor wir im folgenden (Punkt 3) aufzeigen, welche Variablen den Schulerfolg bzw. die Schulleistung mitbedingen — welche Kriterien also miterfaßt werden müssen, um eine gesicherte Schulerfolgsdiagnose oder -prognose stellen zu können —, wollen wir zunächst die *einzelnen* Intelligenzfaktoren im Zusammenhang mit der Schulleistung betrachten.

2.2.1.2. *Korrelative Zusammenhänge zwischen verschiedenen Intelligenzfaktoren und der Schulleistung*

Notenmäßig oder durch Tests erfaßte Schulleistungen stehen, wie wir oben ausführten, nachgewiesenermaßen in Abhängigkeit von der *allgemeinen* Intelligenz. Es ist zu vermuten, daß sich größere Abhängigkeiten ergeben, wenn man die kognitiven Einzelfunktionen getrennt im Zusammenhang mit der Schulleistung betrachtet.

Bottenberg & Wehner (1970) gingen bei der Überprüfung ihrer ähnlich formulierten Hypothese so vor, daß sie für eine Gruppe von 80 Abiturienten alle Abitureinzelnoten, die Werte in den Untertests des IST sowie die Werte einer großen Anzahl von nach French faktoriell geordneten Tests für kognitive Einzelfunktionen erhoben und gemeinsam einer Faktorenanalyse (Zentroid-Verfahren, orthogonale Rotation der extrahierten Faktoren nach dem Varimax-Kriterium) unterzogen. Sie konnten dabei vier statistisch relevante Faktoren extrahieren:

Faktor A spiegelt den Einfluß der Allgemeinen Intelligenz im Bereich der Schulleistungen wider. Es laden auf dem Faktor entsprechend interpretierte IST-Subtests sowie Schulnoten derjenigen Fächer, „die allgemein eher als begabungsmäßig unspezifisch erachtet werden, in denen es darauf ankommt, sich gehäuft Informationen aus verschiedenen Kultur- und Naturbereichen anzueignen und durch Beziehungsstiftung und kombinatorischen Umgang auszubauen und durchzuordnen", d. h. Geschichte, Erdkunde, Biologie, Deutsch, Chemie und zuletzt Mathematik (Bottenberg & Wehner 1970, S. 21). Die Interpretation der Schulfächer in dieser Weise darf unserer Meinung nach jedoch höchstens in der Oberstufe der höheren Schule vorgenommen werden, weil gemeinhin gerade die Fächer Geschichte, Erdkunde und Biologie in den niedrigeren Klassen wenig zu tun haben mit „Beziehungsstiftung und kombinatorischem Umgang", eher mit Informationsspeicherung.

„*Faktor B*" beschränkt sich auf den Bereich kognitiver Funktionen, an denen er einen Zug vorherrschend visueller Flexibilität sichtbar macht, die in

positivem Bezug zur Leistungsgeschwindigkeit steht". Auf diesem Faktor lädt keine der Schulnoten signifikant, d. h. Schulleistung steht mit visueller Flexibilität in keinem bedeutsamen Zusammenhang. Uns erstaunt dieses negative Ergebnis, da wir annehmen, daß zumindest für Leistungen in den naturwissenschaftlichen Fächern sinnliche Wahrnehmungsfähigkeit und Umstrukturierungsfähigkeit im visuellen Bereich eine der Voraussetzungen ist (vgl. dazu auch WENZL 1934).

Den *Faktor C* nennen BOTTENBERG & WEHNER interpretierend „kurzfristige Materialspeicherung, Finden oder/und Verwenden einer sacheigenen Lösungsregel". Auf diesem Faktor laden am höchsten die IST-Subtests „Merken", „Rechenaufgaben", „Zahlenreihen" sowie die Zensuren für Mathematik und für Latein. Demzufolge dürfte mit Faktor C ein intellektuelles Moment angesprochen sein, das in den oftmals als verwandt gekennzeichneten Schulfächern Mathematik und Latein zum Tragen gelangt" (s. noch ARNOLD 1969, LIENERT & HOPP 1964).

Faktor D weist auf ein Moment sprachlichen Verständnisses hin. Hier laden Wortschatztests und die aktuell-sprachlichen Schulfächer wie Englisch und Französisch hoch.

Die Faktorenanalyse läßt mit Vorbehalt deutlich erkennen, daß sich Schulleistungen über den engen konventionellen Intelligenzbereich der allgemeinen Intelligenz hinaus auch auf besondere kognitive Funktionen stützen.

KEMMLER (1967) ging bei der näheren Differenzierung des Zusammenhangs zwischen Intelligenz und Schulleistung ebenfalls faktorenanalytisch vor. Sie verglich die beiden für 9- bis 10jährige „gute" und „schlechte" Volksschüler (Lehrerurteil) getrennt berechneten Ergebnisse der Faktorenanalyse über die Untertests des PMA von THURSTONE. Für die „guten" Schüler ließen sich 10 interpretierbare Faktoren extrahieren, für die „schlechten" nur 8; zur Kritik dieser Klassifizierung s. S. 50 oben. KEMMLER interpretiert dieses Ergebnis als Bestätigung der „Differenzierungshypothese" (BURT 1954, WEWETZER 1958, LIENERT 1960), wonach mit höherem Intelligenzniveau eine differenziertere Intelligenzstruktur zu erwarten ist. Wir sehen das Ergebnis eher als ein Artefakt an, da die letzten beiden Faktoren bei den „guten" Schülern nur noch bei 2 bzw. 5 von 21 Variablen Ladungen aufweisen, und von diesen ist jeweils nur eine signifikant. Sowohl „gute" als auch „schlechte" Schüler erzielen hohe Ladungen auf dem Verbalfaktor, d. h. der Schulerfolg ist in starkem Maße von verbalen Leistungen abhängig. Für beide Gruppen kristallisieren sich ebenfalls die weiteren 4 THURSTONE-Faktoren „Schnelligkeit", „räumliches Erfassen", „Zahlenverständnis" und „schlußfolgerndes Denken" heraus. Für die Gruppe der „schlechten" Schüler hat dabei der Faktor „Zahlenverständnis" die höchste Ladung, d. h. die größte Bedeutung für die Leistungen der „schlechten" Schüler (aber wenig Relevanz für den Schulerfolg), während die „gu-

ten" Schüler ihre Leistungen in erster Linie (nach dem „Sprachverständnis") durch „schlußfolgerndes Denken" (reasoning) erreichen. Ob hierbei die Unterschiede zwischen den beiden Schülergruppen signifikant sind, wurde nicht überprüft.

Auf die weiteren von KEMMLER extrahierten (wenig bedeutsamen) Faktoren wollen wir nicht näher eingehen. Sie haben unserer Meinung nach keine nennenswerte Aussagekraft, da in die Faktorenanalyse nur Untertests des PMA eingegangen sind, die Faktoren also nicht durch „PMA-fremde" Untertests gestützt werden. Lediglich bei den „schlechten" Schülern ist der Faktor 6 auffällig durch seine vergleichsweise hohe prozentuale Gesamtladung. Bei Faktor 6 handelt es sich vermutlich um so etwas wie Gestalterfassung im Bereich von Buchstaben und Wortgebilden, d. h. um Lesefähigkeit und Rechtschreibleistung. Er dokumentiert die Wichtigkeit von Lesen und Rechtschreibung für den Schulerfolg. Wir möchten die hohe Ladung der „schlechten" Schüler gerade auf diesem Faktor vorsichtig dahingehend interpretieren, daß in dieser Gruppe eine nicht unerhebliche Anzahl von Legasthenikern enthalten sein mag, deren schlechte Schulleistungen durch Schwierigkeiten im Lesen und Rechtschreiben bedingt sind.

Wenn man wie WEISS (1964) bei der Untersuchung der differentiellen Beziehung zwischen Schulleistung und Intelligenz nur von *einem* Intelligenztest mit verschiedenen Untertests ausgeht, so ergeben sich auch hier — wenn auch weniger deutliche — Hinweise darauf, daß man nicht von einem Zusammenhang zwischen Intelligenz und Schulleistung schlechthin sprechen sollte, sondern besser angibt, welcher Intelligenzaspekt und auch welche Art von Schulleistung gemeint ist. Bei WEISS korrelieren, abgesehen vom Gesamt-IQ-Wert, mit dem gewogenen Mittel der Schulzensuren (also dem Gesamtmaß für Schulleistung) am höchsten der Wortschatztest, der Untertest „Allgemeines Wissen" und der Verbal-IQ des HAWIK. Die Korrelationen sind in dieser Untersuchung bei Gymnasiasten zwar geringer, zeigen aber eine ähnliche Tendenz, nämlich die größere Bedeutsamkeit des Verbalfaktors. Als mindestens gleichwertig erwiesen sich die Faktoren „space" und „perceptual speed and accuracy" — im Gegensatz zu den Ergebnissen bei BOTTENBERG & WEHNER, die keine Beziehung zwischen Schulleistungen und Wahrnehmungsgeschwindigkeit bzw. -genauigkeit feststellen konnten. Die Deutschnote wird, wie zu erwarten, am stärksten vom Verbal-Faktor beeinflußt, während die Mathematiknote am engsten mit „reasoning" zusammenhängt. Hierfür liegt Übereinstimmung zu den Resultaten von BOTTENBERG & WEHNER vor. Ausgesprochen gering sind dagegen die Zusammenhänge zwischen Intelligenzfaktoren und Fremdsprachen. Bei den Jungen weisen der Verbal-Faktor und Wortflüssigkeit den relativ engsten Zusammenhang mit der Schulnote auf, bei den Mädchen nur Genauigkeit und Ausdauer. Besonders erstaunlich sind die sehr geringen bis negativen Kor-

relationen zwischen Fremdsprache und „reasoning". WEISS ist bedauerlicherweise nicht auf eventuelle Beziehungen zwischen naturwissenschaftlichen Schulfächern und Intelligenzfaktoren eingegangen.

Einen Hinweis für die Unterschiede im Ausmaß des Zusammenhangs von Intelligenz und Schulleistung zwischen Grund- und Oberschule gibt die Untersuchung von ROLOFF (1957). Er verglich die IST-Test-Profile von Grundschülern mit denen von Oberschülern. Besonders deutliche Unterschiede stellen sich dabei beim Untertest „Analogiebildung" und bei der Abstraktionsfähigkeit (Untertest „Gemeinsamkeitenfinden") heraus. Wenn auch ROLOFF die Intelligenz-Struktur-Unterschiede nicht statistisch absicherte, könnte man aufgrund seiner Ergebnisse doch vorsichtig vermuten, daß die Unterschiede im Zusammenhang zwischen Intelligenzfaktoren und Schulleistungen bei Grund- und Oberschülern durch eine unterschiedlich ausgeprägte Intelligenzstruktur dieser beiden Schülergruppen beeinflußt ist.

Eine weitere Untersuchung zur Intelligenzstruktur von Schülern (451 Schüler des 4. Schuljahres) hat SIMONS (1969) durchgeführt. Ziel der Arbeit war die Prüfung der Frage, ob Under- und Overachiever typische Leistungsschwerpunkte in einer Intelligenztestbatterie erkennen lassen, und ob zwischen den Vergleichsgruppen Differenzen in der Intelligenzstruktur deutlich werden. Als Under- bzw. Overachiever werden Schüler bezeichnet, deren (von Lehrern beurteilte) Schulleistung erheblich nach oben oder unten von den Leistungen abweicht, die man aufgrund ihrer Intelligenz erwartet. Da der gemeinsame Varianzanteil von Schulleistungen und nichtkognitiven Variablen bei diesen Schülern entsprechend größer ist als bei den sogenannten Achievern, d. h. bei kongruenten Begabungs- und Schulleistungsverhältnissen, kann ihre Schulleistung weit weniger präzise durch Intelligenztests vorhergesagt werden. Aus diesem Grunde sollte man bei der Betrachtung der signifikanten Resultate aus der Arbeit von SIMONS bedenken, daß charakteristische Profile dieser Gruppen weniger Bedeutung haben als Profile von Gruppen, deren Schulleistung durch Intelligenztestergebnisse zuverlässiger vorhergesagt werden kann. Bedauerlicherweise unterzieht auch SIMONS seine Daten keiner eigentlichen Profilanalyse (s. z. B. LIENERT 1961), sondern vergleicht lediglich die einzelnen Untertests miteinander. Dabei erreichen die Underachiever im Wortschatztest der von KEMMLER & LANGHEINRICH (1967) überarbeiteten Form des „Primary Mental Ability" Tests (PMAT) von THURSTONE sehr signifikant schlechtere Leistungen als die Overachiever. Für die Overachiever deutet sich außerdem ein Leistungsvorsprung im Zahlentest des PMAT an. Während die Underachiever in den Untertests „Bildwortschatz", „Raumerfassung" und „Wahrnehmungsgeschwindigkeit" hochsignifikant bessere Leistungen erzielten als im Wortschatztest, erreichten die Overachiever genau umgekehrt in diesen Tests signifikant geringere Leistungen als im Wortschatztest. Mit anderen

Worten: Overachiever leisten in den schulgebundenen Untertests mehr, während die Underachiever in jenen Tests höhere Punktewerte erreichen, deren Bewältigung weniger von erworbenem Schulwissen abhängt. Diese Ergebnisse konnten in etwa bestätigt werden von FRANKEL (1960), CARMICAL (1964), COLEMAN & RASOF (1963) und KEMMLER (1967). Da der Schulerfolg in erster Linie bedingt ist durch verbale Leistungen (siehe WEISS 1964, BOTTENBERG & WEHNER 1970, KEMMLER 1967), wird deutlich, warum Over- und Underachiever bei gleichen Gesamt-Intelligenztest-Ergebnissen in der Schule erfolgreicher bzw. weniger erfolgreicher sind als erwartet.

Zusammenfassend läßt sich sagen, daß durch die zitierten Untersuchungen die starke Abhängigkeit des Schulerfolgs von verbalen Fähigkeiten, besonders in den unteren Klassen, als gesichert angesehen werden kann. Für einige Schulfächer (Mathematik, Latein, nach KEMMLER auch für die Durchschnittsnote) scheint zusätzlich der Faktor „reasoning" (schlußfolgerndes Denken und Abstraktionsfähigkeit) bestimmend zu wirken. Alle anderen intellektuellen Fähigkeiten werden demnach in der Schule weniger gefordert bzw. gefördert.

Ebenso mangelt es in unseren Schulen an der Anerkennung bzw. Unterstützung von „kreativen" Leistungen. Eine der umfassendsten Definitionen für kreatives Denken stammt von NEWELL et al. (1962):

Denken ist kreativ, wenn es folgende Bedingungen erfüllt:

a) das Ergebnis ist neuartig und für den Denkenden oder die Kultur von Wert,
b) es ist unkonventionell,
c) es ist hoch motiviert und beharrlich oder von hoher Intensität,
d) das Problem war zunächst vage und nicht spezifiziert. Ein Teil der Aufgabe war die Formulierung des Problems (nach GETZELS & MADAUS 1969).

TAYLOR äußerte sich auf einer Konferenz über Kreativität in Salt Lake City zur Bedeutung der durch Tests erfaßten Intelligenz folgendermaßen: „Intelligenz ist eine Erfindung unserer westlichen Kultur, die darauf achtet, wie schnell jemand völlig unwichtige Probleme lösen kann, ohne irgend einen Fehler zu machen" (nach KEMMLER 1969). Diese Aussage über die Intelligenz läßt sich cum grano salis auch auf die Schulleistung übertragen, wenn man sich etwa die in heutigen Lehrplänen definierten Lernziele ansieht, die vielfach auf Erwerb von Faktenwissen (Geschichtszahlen) und Auswendiglernen von Gedichten oder das Finden vorprogrammierter Lösungen für praxisferne Mathematikaufgaben abzielen. Von hierher wird auch die Problematik der Underachiever verständlich. Sie zeigen nicht nur andere Intelligenzschwerpunkte als für die Schulleistung erforderlich, sondern oft auch mehr Leistungen, die den oben aufgeführten Kreativitätsbedingungen folgen. Sie langweilen sich in dem konvergent ausgerichteten

Unterricht und stören diesen mit ihrem divergenten Denken, was den Lehrer wiederum dazu veranlaßt, sie als „schlechte" Schüler einzustufen. GETZEL & JACKSON (1962), die „creativity" eher zu den kognitiven Stilen zählen als zu Verhaltenseigenschaften wie z. B. WEWETZER (s. u.), betonen in ihrer Sechs-Jahres-Studie, daß „kreative" Schüler in den Schulen immer wieder zugunsten der Intelligenten übersehen werden, daß sich schöpferische Schüler weniger angepaßt verhalten und daher weniger Förderung erfahren. „Offenbar sind diese Kreativen weniger bereit zu vertrauen und hinzunehmen" (HASELOFF 1966). Im deutschsprachigen Raum konnten AMELANG & VAGT (1970) bestätigen, daß die bessere Vorhersagbarkeit der Schulleistungen weiblicher gegenüber männlichen Schülern auf ihr größeres Ausmaß der Integration in das Schulsystem zurückzuführen sei (besonders gute Betragensnoten).

FELDHUSEN, TREFFINGER & ELIAS (1970) mußten in ihrer Untersuchung zur Voraussageeffizienz verschiedener Prädiktoren (Ängstlichkeit, Selbsteinschätzung der kreativen Fähigkeiten, konvergentes und divergentes Denken) ebenfalls feststellen, daß sich Schulerfolg am ehesten durch Variablen des konvergenten Denkens vorhersagen läßt. Und auch KEMMLER (1967) mißt für Underachiever (Schüler mit hohen Intelligenzwerten, aber geringen Schulleistungen) einen höheren Originalitätswert im Einfallsreichtum. Diese hohe Originalität macht sich ihrer Vermutung nach in der Schulroutine eher als mangelnde Anpassung an die gewohnten Denkschemata denn als Bereicherung der Schulleistung bemerkbar. Die Overachiever (Schüler mit relativ geringen Intelligenzwerten, aber guten Schulleistungen) dagegen zeichnen sich durch besonders gute Arbeitshaltung, durch Anpassung, Stetigkeit und Anstrengungsbereitschaft aus. Dabei ist der Prozentsatz der Mädchen bei den Overachievern größer als der der Jungen, bei Underachievern ist das Verhältnis umgekehrt.

LÜCKERT (1969, S. 274) konstatiert: „Die Zusammenfassung kreativer Begabungen (in schulische Leistungsgruppen, d. Verf.) ist mit größeren Schwierigkeiten verbunden, da im gegenwärtigen Schulsystem vom Lehrplan her die hier erforderliche Freizügigkeit und vom Unterrichtsverfahren her die hier erforderlichen Anregungen und Anleitungen nicht gewährleistet sind." WEWETZER (1970, S. 57) berichtet von einer Untersuchung zur Rangordnung der Schülermerkmale durch Lehrer: „Die Rangordnung (der Merkmale) bei den deutschen Lehrern: (Der Schüler ist) vertrauenswürdig, gewissenhaft, höflich, rücksichtsvoll, beliebt bei anderen Kindern, ruhig und entspannt. — Das wäre also die „Maßeinheit" bei der Beurteilung kindlichen Verhaltens in der Schule. Weit ist der Weg zur Kreativität." WEWETZER meint hier mit Kreativität jedoch eher eine Verhaltenseigenschaft, weniger eine kognitive Fähigkeit, wie wir sie im Rahmen dieses Aufsatzes selbst verstanden wissen wollen.

Als Resümee unseres Referats über Untersuchungen, die sich die Aufdeckung der kognitiven Faktoren der Schulleistung zur Aufgabe gestellt hatten, ergibt sich die Feststellung, daß längst nicht alle kognitiven Fähigkeiten für Erfolg in der Schule garantieren. Es kommt also nicht darauf an, intelligent oder gar kreativ zu sein, sondern bevorzugt werden Schüler, die in ganz bestimmter Weise intelligent sind (vorwiegend verbal).

Untersuchungen, die sich mit der Beeinflussung der Schulleistung durch nicht-kognitive Faktoren befassen, werden in den folgenden Kapiteln erörtert.

2.2.2. Nicht-kognitive Faktoren der Schulleistung

2.2.2.1. Motivation und Arbeitshaltung

Um die Vermutung zu untersuchen, daß die Güte der Schulleistung evtl. durch das Ausmaß der Motivation, genauer der Leistungsmotivation, mitbestimmt wird, ist eine Klärung des Begriffs „Leistungsmotivation" erforderlich. Bei Durchsicht der Literatur zu diesem Bereich finden sich jedoch ebensoviele unterschiedliche, ungenaue und sich mehr oder weniger überschneidende Definitionen für „Motivation" wie für andere psychologische Konstrukte, z. B. „Intelligenz" oder „Begabung". Als das wesentlichste Kennzeichen der Leistungsmotivation im Sinne der Antriebsquelle für Leistungen wird bei allen Autoren die zukunftsperspektivische Zielbezogenheit erachtet, die sich in den Zielsetzungen, durch das jeweilige Verhalten Erfolg zu erreichen oder Mißerfolg zu vermeiden, niederschlägt. Was von dem einzelnen dabei als Erfolg oder Mißerfolg betrachtet wird, ist individuell sehr verschieden. McCLELLAND (1965) definiert allgemeinverbindlich die Leistungsmotivation als Auseinandersetzung mit einem Gütemaßstab, der persönlichen Anspruch an die sachliche Leistung manifestiert. Den Gütemaßstab als zentrales Element der Leistungsmotivation gibt ebenfalls HECKHAUSEN (1965, S. 604) an, wenn er Leistungsmotivation definiert als „das Bestreben, die eigene Tüchtigkeit in all jenen Tätigkeiten zu steigern oder möglichst hoch zu halten, in denen man einen Gütemaßstab für verbindlich hält und deren Ausführung deshalb gelingen oder mißlingen kann". Er unterscheidet dabei drei verschiedene Gütemaßstabskategorien: sachbezogene Gütemaßstäbe (Vollkommenheitsgrad des Tätigkeitsprodukts), personenbezogene Gütemaßstäbe (Vergleich mit früheren Leistungen) und sozialbezogene Gütemaßstäbe (Vergleich mit Leistungen anderer). Sie können nebeneinander bestehen; die Situation bestimmt, nach welchem Gütemaßstab die Leistung vom Individuum bewertet wird. Dabei ist das Erleben von Erfolg und Mißerfolg weitaus mehr von dieser persönlich gesetzten Leistungsnorm abhängig als von der objektiven Güte der Leistung. Ein bestimmter Hand-

lungseffekt „wird erst dadurch zum Erfolg oder Mißerfolg, daß eine Beziehung zu einem angestrebten Ziel, einem Ideal oder einer sonstigen Norm besteht, die als momentanes Maß für den Handlungseffekt in seiner Bedeutung als Leistung gilt" (HOPPE 1930, S. 10). Es ist nun leicht vorstellbar, daß nicht alle Schüler den Gütemaßstab der Schule als eigene Erfolgsnorm für verbindlich ansehen, was in erster Linie auf die Genese der Leistungsmotivation zurückzuführen ist. „Die Entwicklung der Leistungsmotivation vollzieht sich mit erheblichen individuellen Unterschieden, die sowohl im Zeitpunkt des Auftretens als auch im Grad der individuellen Ausprägung deutlich werden. Solche Abweichungen sind auf divergierende Erziehungsstile und Erziehungspraktiken der familiären Umwelt zurückzuführen" (WASNA 1972, S. 29). Die Erfahrungen, die das Kind macht, während es nach Selbständigkeit und Unabhängigkeit, also eigenständigen Leistungen strebt, und besonders die Art und Weise, wie sein Lernen durch enge Bezugspersonen gesteuert wird, haben — wie auch die Untersuchung von WINTERBOTTOM (1958) gezeigt hat — entscheidende Bedeutung für die Entwicklung des Leistungsstrebens sowohl für das spätere Ausmaß als auch für die Qualität. Wenn z. B. Leistungsstreben zu Mißerfolg führt, wird das Verhalten gemindert, frühe Bekräftigung der kindlichen Aktivitäten dagegen fördert die Leistungsmotivation, ebenso affektive Zuwendung und hohe (nicht zu hohe) Leistungsansprüche der Eltern.

Die Untersuchung von WASNA (1972) beschäftigt sich mit der Frage, wie Leistungsmotivation und Schulerfolg miteinander in Beziehung stehen. Als Meßinstrument diente der Thematic Apperception Test (TAT) nach MCCLELLAND, ein projektiver Geschichtenerzähltest. Es werden mehrdeutige Bilder vorgelegt, zu denen der Proband spontan Geschichten erzählen soll. Die einzelnen Aussagen werden unter anderem nach den Kategorien „Hoffnung auf Erfolg" (HE) und „Furcht vor Mißerfolg" (FM) ausgewertet. WASNA ging in ihrer Arbeit von der Hypothese aus, daß schwache Schüler eine geringere Gesamtmotivation und niedrigere HE-Werte im TAT haben als gute Schüler. Da gute Schüler häufiger erfolgreich sind, wäre zu erwarten, daß sie eine höhere Erfolgsmotivation haben als leistungsschwache Schüler, und diese wiederum mehr Furcht vor Mißerfolg zeigen und insgesamt eine niedrigere Leistungsmotivation haben (s. auch die Ergebnisse von MEYER et. al. 1965). Wider Erwarten unterscheiden sich gute und schlechte Schüler in bezug auf die Gesamtmotivation überhaupt nicht, auch nicht bei gleichem Intelligenzniveau, außerdem sind die HE-Werte bei den Schwachen höher, die FM-Werte (Furcht vor Mißerfolg) niedriger als bei den Guten, also genau umgekehrt als erwartet. WASNA interpretiert dieses paradox scheinende Ergebnis durch die Annahme, die leistungsbezogenen Aussagen in den TAT-Geschichten der schwachen Schüler seien eher der Irrealitätsebene des Wunsches als der Realitätsebene zuzurechnen. Die geringen

FM-Werte interpretiert sie als Abwehrreaktionen. „Wenn sich die erwarteten Zusammenhänge zwischen Erfolgsmotivation und allgemeinem Schulerfolg nicht bestätigen lassen, statt dessen jedoch deutlich wird, daß unbewußte Abwehrreaktionen und realitätsferne Wünsche in den leistungsbezogenen Inhalten der Geschichten zum Ausdruck kommen, wird der TAT als Meßinstrument der Leistungsmotivation immer fragwürdiger" (WASNA 1972, S. 87). Die Untersuchungen von WASNA ermöglichen also keinerlei Aussagen über den Zusammenhang zwischen Leistungsmotivation und Schulerfolg, da sie ein ungeeignetes Untersuchungsinstrument zur Grundlage haben. Bedauerlicherweise ist der TAT als Meßinstrument der Leistungsmotivation noch immer sehr verbreitet. VONTOBL (1970) z. B. behauptet, das inhaltsanalytische Verfahren nach der TAT-Methode habe sich unter den verschiedenen Meßmethoden in seiner sehr komplexen Untersuchung zum Leistungsbedürfnis im Zusammenhang mit der sozialen Umwelt als das befriedigendste Verfahren erwiesen. Er ermittelte eine höhere Leistungsmotivation jeweils für Personen der „nicht-traditionellen" Umwelt („nichtbäuerlich"), für Personen höherer Sozialschichten (besonders des Mittelstandes), für Katholiken und für „Gebildete" jeweils im Vergleich zu ihrer Komplementärgruppe. Dabei konnte er jedoch nicht nachweisen, daß diese ermittelten Motivationswerte tatsächlich der „intrinsischen" Leistungsmotivation entstammen und nicht, wie den Ergebnissen von WASNA zufolge, der irrationalen Wunschebene.

Wenn auch nicht direkt für die Schulleistung, so konnte doch eine positive Beziehung zwischen Intelligenz und Lernmotivation aufgedeckt werden. Nach SONNTAG, BAKER & NELSON (1958), KAGAN et al. (1958), KAGAN & MOSS (1959), ZIGLER & BUTTERFIELD (1968) sowie SKOLNIK (1966) und MCCLELLAND (1966) steigt der Grad der Intelligenz bei Kindern mit hoher Leistungsmotivation im Alter zwischen $4^1/_2$ und 15 Jahren deutlich an. Bei Niedrigmotivierten bleibt der IQ gleich oder fällt ab. Durch die positiven Zusammenhänge zwischen Intelligenz und Schulleistung kann vorsichtig auf eine gewisse Wirkung der Leistungsmotivation auf die Schulleistung geschlossen werden. Zumindest *mit*entscheidend für die Schulleistung ist demnach die Weckung oder Steigerung der Lern- und Leistungsmotivation. Wie kann sie geweckt bzw. gesteigert werden?

Diese Frage legte ROSENFELD (1966) seinen Untersuchungen zugrunde. Er ging nicht von der Überlegung aus, ob höher motivierte Kinder bessere Schulleistungen erbringen, sondern setzte dies bereits als Tatsache voraus und untersuchte nun, unter welchen schulischen Bedingungen die Bereitschaft der Schüler zu Aktivitäten im Unterricht am größten und stabilsten ist. Er versuchte also nicht, mit fragwürdigen Methoden zu ermitteln, wie die Kinder sind, sondern wie sie unter welchen Bedingungen werden können. Er kommt dabei zu dem Schluß, daß „durch welche Zwecke auch immer Lern-

handlungen dynamisch involviert werden, entscheidend ist vermutlich — für den Prozeß wie für seinen Effekt — die subjektive Bedeutung des mit der Lernhandlung verbundenen Zweckziels" (S. 198). Die Verbindung zwischen Zweckmotivation und Lernhandlung erweist sich demnach als vorteilhafter als eine Selbstzweckmotivierung des gleichen Vorganges, „und zwar vorteilhafter, was die systematische Steuerung des Prozesses, den emotionalen Erlebniswert und schließlich auch den Lerneffekt betrifft". In allen jenen Situationen, in denen den Kindern das Lernen mit einem bestimmten und für sie verbindlichen Zweck verbunden scheint, werden sie erheblich stärker motiviert, als wenn sie um des Lernens willen („für die Schule") lernen sollen. ROSENFELD versäumte es allerdings zu untersuchen, ob dieser motivierende Lernzweck tatsächlich auch für alle seine Versuchspersonen verbindlich ist und ob sich nicht vielleicht noch deutlichere Unterschiede zwischen zweckorientiertem und selbstzweckorientiertem Lernen herausstellen würden, wenn das Ausmaß der Zweckverbindlichkeit für die Versuchspersonen konstant gehalten würde. Dieses festzustellen ist allerdings nicht einfach.

In einem zweiten Versuch untersuchte ROSENFELD unterschiedliche Unterrichtsinhalte für jeweils gleiche Unterrichtsfächer hinsichtlich ihrer aktivitätsstimulierenden Wirkung. Auch hier ergab sich dasselbe Resultat, daß nämlich Aufgaben, die mit einer gewissen Zweckhaftigkeit verbunden waren, gegenüber abstrakteren den größeren Zuwendungswert hatten, selbst wenn die zweckbehafteten Aufgaben in ihrer Durchführung nicht besonders interessant oder abwechslungsreich waren.

Unsere eigene Hypothese zur Wirkung der sog. Leistungs- oder Lernmotivation auf die Schulleistung sehen wir durch die Ergebnisse ROSENFELDS eher gestützt: daß nämlich nicht das *Ausmaß* der Leistungsmotivation für den Schulerfolg bestimmend ist, sondern eher der *Inhalt*, auf den sich die Leistungsmotivation bezieht. Was Leistung ist, bestimmt in der Schule der Lehrplan, außerhalb der Schule bestimmen dies jedoch die jeweiligen soziokulturellen Leistungsnormen und Werteinstellungen der Sozialgruppe, der der einzelne Schüler angehört. Und hier bestehen vielfältige Differenzen, wie z. B. Untersuchungen zur Begabtenreserve (HELLER 1970, ARNOLD 1968, LÜCKERT 1969) belegen. Wessen Leistungsmotivation auf andere Gebiete als die durch die Schulfächer repräsentierten gerichtet ist, für wen also Schulunterricht nur geringe „Zweckhaftigkeit" besitzt, der wird sich diesen Schulfächern weniger aktiv zuwenden und weniger erfolgreich sein. Demnach besteht möglicherweise lediglich ein Zusammenhang zwischen „auf-schulische-Inhalte-gerichteter" Leistungsmotivation und Schulerfolg.

In enger Beziehung zu der „Schulleistungsmotivation" stehen die sog. Arbeitshaltungen wie Konzentration, Fleiß, Ausdauer. Sie sind die Qualitäten des aus der Leistungsmotivation resultierenden Verhaltens. „Verhalten kommt durch ein bestimmtes Motiv in Gang, oder bereits vorliegendes

Verhalten wird intensiviert. Das Tätigsein an sich ist Kriterium für motiviertes Verhalten" (KRUSE & ROGGE 1971, S. 106). Wir meinen dagegen mit WINTERBOTTOM (1958), nicht allein das „Tätigsein an sich", sondern auch die Art und Weise des Tätigseins (konzentriert, mit Ausdauer, fleißig ...) kennzeichnen, ob ein Verhalten motiviert oder unter äußerem Druck geschieht. Motiviertes Verhalten wird, da es mit mehr Engagement geschieht, auch eher zu Erfolg führen, so daß sich dadurch wiederum das Verhalten bzw. die Motivationslage stabilisiert (s. Bekräftigungsprinzip der Lerntheorien), womit der Kreis zwischen Motivation, Verhalten und Schulerfolg geschlossen ist.

Das Hauptaugenmerk bei der Frage nach der Wirkung der Leistungsmotivation auf die Schulleistung ist also auf die äußeren Anregungsvariablen zu legen, die die Lern- und Leistungsmotivation sowie entsprechendes Verhalten aktivieren und dadurch Schulerfolg bedingen. Nach HECKHAUSEN (1971) sind solche situationsgebundenen Anregungsvariablen:

a) der Erreichbarkeitsgrad des in der Lernsituation gestellten Leistungsziels,
b) der Anreiz von Aufgaben,
c) der Neuigkeitswert des dargebotenen Lehrstoffes.

Dabei meint er mit Erreichbarkeit die *individuelle Schwierigkeit*, den Leistungsaufwand, den der einzelne für die Bewältigung der Aufgabe erbringen muß. „Die Anregungswirkung ist am größten, wenn eine dosierte *mittlere* Diskrepanz zwischen der Umweltsituation und dem erreichten sachstrukturellen Entwicklungsstand, zwischen Angebot und bereits gespeichertem Vorrat besteht" (S. 200). Hierin stimmt ROSENFELD (1966) mit HECKHAUSEN überein, wenn er in seiner Theorie zur Motivierung die Setzung innerer Widersprüche („die fortwährende Setzung und gleichzeitige Lösung dieses Widerspruchs ist eben die Bewegung"; zit. nach ENGELS 1952, S. 198), d. h. die Gegensatzbildung zwischen vermehrter Forderung und minderer Leistungsmöglichkeit betont, wobei diese Setzung innerer Widersprüche nur dann motivierend wirkt, wenn sie weder eine Überforderung noch einen Pseudo-Widerspruch als Unterforderung darstellt. Auch die Untersuchungen von WINTERBOTTOM (1958), von ROSEN & D'ANDRADE (1959) weisen auf die positive Wirkung dieser mittleren Forderungen hin: Hohe Leistungsansprüche der Eltern und hohe Leistungserwartungen begünstigen die Ausbildung von Gütemaßstäben, d. h. einen Leistungsanspruch des Kindes an sich selbst; extrem hohe Leistungserwartungen der Eltern scheinen sie jedoch eher zu hemmen, ebenso wie extrem niedrige.

Das gleiche gilt für den Anreiz von Aufgaben: mittelschwere Aufgaben besitzen den höchsten Anreizwert. ATKINSON & FEATHER (1966) und HECKHAUSEN (1967) konnten zeigen, daß mittlere Schwierigkeitsgrade, bei denen ein Erfolgs- oder Mißerfolgsausgang ungefähr gleich wahrscheinlich ist, den höchsten Motivierungswert haben (s. a. VONTOBEL 1970). Auch der Neuig-

keitsgehalt eines Lernstoffes hat die größte Wirkung für die Leistungsmotivierung, wenn „die Erwartungsschemata des Schülers in einem mäßigen Grad durchbrochen werden. Dann erscheint der Stoff interessant, überraschend oder komplex" (HECKHAUSEN 1967, S. 196; s. a. BERLYNE 1960 und FESTINGER 1957).

Das heißt also: Das wichtigste Mittel, den Schüler im Unterricht zu motivieren, ist eine fortlaufend optimale Dosierung des Schwierigkeitsgrades je nach Entwicklungsstand der Schüler. Im gegenwärtig üblichen Klassenunterricht ist diese Dosierung allerdings nur auf den Entwicklungsstand der ganzen Klasse abzustimmen und nicht, wie es eigentlich sinnvoll wäre und in den schwedischen „Schulkliniken" bereits gehandhabt wird, auf den einzelnen Schüler. Denn ein „mittlerer" Erreichbarkeits-, Schwierigkeits- und Neuigkeitsgrad, der für die ganze Klasse (also den Durchschnitt) verbindlich sein soll, wird mindestens für alle über- und unterdurchschnittlichen Schüler kein „mittlerer", sondern ein zu hoher oder zu niedriger sein. Diese Schüler werden dann über- oder unterfordert.

2.2.2.2. Persönlichkeitsvariablen

Zu den Auswirkungen verschiedener Persönlichkeitsvariablen der Schüler auf ihre Schulleistungen liegen einige Arbeiten vor, die hier kurz wiedergegeben werden sollen. Wir möchten jedoch von vornherein auf unsere starken Bedenken derartigen Untersuchungen gegenüber hinweisen: Persönlichkeitsdimensionen werden mit Tests gemessen, die meist sehr wenig reliabel (zuverlässig) und valide (gültig) sind und für die in der Regel keine ausreichenden deutschen Normen vorliegen. Die Schwierigkeit, einen meßtheoretisch einwandfreien Persönlichkeitstest zu konstruieren, liegt unseres Erachtens in dem Persönlichkeitskonzept selbst begründet.

ALLPORT (1960) hat aus 50 verschiedenen Definitionen zusammenfassend „Persönlichkeit" so formuliert: „Persönlichkeit ist die dynamische Ordnung derjenigen psychologischen Systeme im Individuum, die seine einzigartige Anpassung an die Umwelt bestimmen" (S. 27). Es gibt also etwas, nämlich die Persönlichkeit, das überdauernd den Menschen von seinen Mitmenschen unterscheidet. Ob es nun so etwas Überdauerndes (vor allen Dingen bei Kindern) überhaupt gibt, ist die eine Frage, ob nicht viel eher Umwelteinflüsse permanent modifizierend wirken, die andere, die mit der Frage nach der Meßbarkeit dieser überdauernden Persönlichkeitszüge direkt zusammenhängt. Zumindest die zweite Frage kann durch die z. Z. gebräuchlichen projektiven Verfahren und Persönlichkeitsfragebögen nicht positiv beantwortet werden, da ihre mangelnde Reliabilität doch eher darauf schließen läßt, daß ein wenig stabiles Merkmal gemessen wird. Wir wollen bei der Betrachtung der im folgenden dargestellten Untersuchungsergebnisse statt von Persön-

lichkeitseigenschaften davon ausgehen, daß hier bestimmte Verhaltenseigenschaften in ihrem Einfluß auf den Schulerfolg untersucht wurden, die durchaus nicht „überdauernd" sein müssen, sondern veränderbar sind.

2.2.2.2.1. Ängstlichkeit

Zahlreiche, vorwiegend amerikanische, Untersuchungen lassen es wahrscheinlich erscheinen, daß ängstliche Schüler in vielfacher Weise ihren weniger ängstlichen Mitschülern unterlegen und dadurch in ihrem schulischen Fortkommen behindert sind (FELDHUSEN & KLAUSMEIER 1962, COWEN et al. 1965, HILL & SARASON 1966, SARASON 1966). Das Phänomen ließ sich u. a. auch in Schweden (LJUNG 1960) und in Australien (COX 1962) beobachten. Im deutschen Sprachraum trat in jüngsten Untersuchungen von ZIELINSKI (1967), NICKEL & SCHLÜTER (1970), SCHNELL (1972) sowie von NICKEL, SCHLÜTER & FENNER (1973) der ungünstige Einfluß erhöhter Angst auf die Schulleistungen hervor. Eine Übereinkunft darüber, wie die Begriffe „Angst", „Ängstlichkeit", „ängstliches Verhalten" zu definieren sind, haben alle diejenigen Forscher, die sich mit diesem Konstrukt beschäftigt haben, noch nicht finden können. GÄRTNER-HARNACH (1972) charakterisiert dieses Konstrukt zusammenfassend nach seinen vier verbindlichen Merkmalen: Angst entsteht als eine Reaktion auf eine Gefährdung, die von außen kommt oder in einer Person zu suchen ist (z. B. Leistungsforderung) bzw. als Antizipation einer derartigen Bedrohung; sie wird als unangenehm wahrgenommen; sie bewirkt einen Anstieg des physiologischen Aktivierungsniveaus; und sie beeinflußt in der Regel das Verhalten. Unter den Theorien über die Auswirkungen von Angst auf Leistung scheinen uns die von MANDLER & SARASON (1952) und von ATKINSON & MCCLELLAND (1953) am relevantesten. Sie sehen die Leistungsangst als etwas Erlerntes an, das einerseits aufgabengerechte Handlungen hervorruft, andererseits aber auch störende Verhaltensweisen und Abwehrreaktionen in Gang setzt, wobei auch „Hoffnung auf Erfolg" und „Furcht vor Mißerfolg", also Formen der Leistungsmotivation, eine intervenierende Rolle spielen. Inwieweit sich nun die Leistungsangst fördernd oder störend auf den Handlungserfolg auswirkt, ist in zahlreichen Untersuchungen zu erforschen versucht worden. Die Ergebnisse von BRANDSTÄTTER, FRANKE & ROSENSTIL (1966) weisen darauf hin, daß Ängstlichkeit (gemessen mit dem IPAT Anxiety Scale Questionnaire von CATTELL und SCHEIER 1963) Quantität und Qualität von Leistungen beeinflußt. Richtung (Minderung oder Steigerung) und Ausmaß des Einflusses scheinen vom Ängstlichkeitsniveau und vom Belastungsgrad der Aufgabe bzw. Situation abhängig zu sein. Dabei ergab sich eine kurvilineare Regression von Kenntnissen (gemessen mit dem Differentiellen

Kenntnistest von TODT) auf Ängstlichkeit. Dies bedeutet: Die Leistungen steigen bis zu einem Ängstlichkeitsoptimum an (bis hierhin ist Ängstlichkeit also leistungsfördernd) und fallen dann wieder ab (ab diesem Ausmaß ist Ängstlichkeit zunehmend leistungshemmend). Das Ängstlichkeitsoptimum liegt in der Untersuchung von BRANDSTÄTTER, FRANKE und ROSENSTIL bei stark belastenden Aufgaben niedriger als bei weniger belastenden Tests. Das scheint zwar plausibel, da jedoch die Belastetheitsgrade der von den Autoren verwendeten Fähigkeitstests von ihnen selbst ohne experimentelle Überprüfung subjektiv eingestuft worden waren, wollen wir auf dieses Ergebnis nicht näher eingehen. Für den Schulunterricht wesentlicher erscheint uns das Resultat, daß ängstliches Verhalten in einem bedeutsam höheren Zusammenhang steht zu Beobachtungstempo als zu Denkgenauigkeit, dem Intelligenzniveau und auch dem Durchschnitt der Abiturnoten, also der Oberschulleistung. Demnach beeinträchtigt Ängstlichkeit vor allen Dingen Leistungen, die unter starkem Zeitdruck stehen, z. B. Klassenarbeiten bzw. Prüfungsleistungen.

Das geringe Ausmaß der Beeinflussung von Intelligenztestwerten durch Testangst (ebenfalls gemessen mit dem IPAT) konnten auch BÄUMLER & BREITENBACH (1970) durch eine Faktorenanalyse von Intelligenztestwerten, Angstwerten und Leistungsmotivationswerten feststellen. Das hieße, daß sich das Ausmaß der Ängstlichkeit von Schülern besonders in Fächern, deren Intelligenzkomponente groß ist (vgl. Abschnitt 2.2.2.), weniger störend bemerkbar macht. Die Untersuchungsergebnisse von SARASON, HILL & ZIMBARDO (1964) bestätigen diese Vermutung, da auch sie eine größere negative Beziehung zwischen Ängstlichkeit und Schulleistung als zwischen Ängstlichkeit und Intelligenz aufzeigen und bei den einzelnen Schulleistungen wiederum einen größeren Zusammenhang zwischen Ängstlichkeit und Lesenlernen als zwischen Ängstlichkeit und Rechnen (Rechnen hatte sich als intelligenzabhängiger erwiesen als das Lesen). Die Faktorenanalyse von SEITZ (1971) über Persönlichkeitsdimensionen des Children's Personality Questionnaire (CPQ) nach PORTER & CATTELL (1963), verschiedene Frankfurter Schultests, den AIT (Analytischen Intelligenztest von MEILI), den CFIT (Culture Fair Intelligence Test von CATTELL) sowie alle Schulnoten erbrachte so gut wie keine Zusammenhänge zwischen den Angstdimensionen und den Schulleistungen. Lediglich die Dimensionen „Unsicherheit im sozialen Umgang" versus „Selbstsicherheit im sozialen Auftreten" weisen positive Korrelation zu den Zensuren in Deutsch, Rechnen, Erdkunde und Naturkunde auf.

Auch in der Validitätsstudie von TEWES (1973) zur HANES (KJ) von BUGGLE, GERLICHER & BAUMGÄRTEL (1968) ergab sich lediglich ein schwach negativer Zusammenhang zwischen Neurotizismus (Angsterleben u. a.) und Schulleistung, und hier auch nur für Mädchen.

Bei einer Untersuchung von Over- und Underachievern konnte WEINERT (1965) sogar überhaupt keine signifikanten Unterschiede bezüglich der Ängstlichkeit (gemessen mit der deutschen Version der Children Manifest Anxiety Scale) feststellen. Die unterschiedlich hohen Schulleistungen von Schülern mit gleichem Intelligenzniveau sind demnach nicht auf ein verschieden großes Ausmaß an Ängstlichkeit zurückzuführen. Insgesamt bestätigen jedoch die Ergebnisse von WEINERT die Befunde von SARASON (1966), nach denen ängstliche Kinder etwas schlechtere Schulleistungen erzielen als weniger ängstliche. Diese Tendenz ist bei Mädchen deutlicher ausgeprägt als bei Jungen, das gleiche Ergebnis erzielte auch ZIELINSKI (1967), der einen stärkeren negativen Zusammenhang zwischen der Schulleistung und Werten der ins Deutsche übersetzten Children Manifest Anxiety Scale feststellte als WEINERT (beide Kindergruppen unterschieden sich nicht hinsichtlich des Alters). Ebenso ergab sich bei NICKEL, SCHLÜTER & FENNER (1973) ein signifikanter korrelativer Zusammenhang zwischen Schulleistung und Schulangst (gemessen mit der deutschen Fassung der Test Anxiety Scale for Children [TASC] von SARASON et al. 1958) in dem Sinne, daß fast ohne Ausnahme schlechte Schüler mehr Schulangst aufwiesen. Mit der Schulangst sank gleichzeitig die Mitarbeit im Unterricht. Auch in dieser Untersuchung erwiesen sich Mädchen wiederum als ängstlicher im Vergleich zu den Jungen. Außerdem nahm die Schul- und Prüfungsangst mit ansteigender Klassenstufe ab.

Die Unterschiede zwischen Jungen und Mädchen führen wir, wie auch andere Autoren (NICKEL, SCHLÜTER & FENNER 1973, SARASON 1958, 1960), auf ihre verschiedenen Rollen im sozialen Kontext zurück. Mädchen fällt es leichter als Jungen, Ängstlichkeit zuzugeben, da ängstliches Verhalten innerhalb der Mädchenrolle sozial gestattet ist. Jungen hingegen müssen eher „stark", „mutig" und „tapfer" sein, ängstliches Verhalten wäre für sie nicht rollenkonform. Der per Fragebogen festgestellte Ängstlichkeitsunterschied zwischen Jungen und Mädchen ist möglicherweise also nur ein Unterschied im Verbalverhalten! Die altersmäßigen Unterschiede lassen sich nach demselben Prinzip erklären: jüngere Kinder „dürfen" ängstlicher sein als ältere, oder sie können auf Gewöhnungsprozesse (Gewöhnung an die Angst hervorrufende Situation) zurückzuführen sein.

Den zitierten Untersuchungsergebnissen zufolge stellt sich der Zusammenhang zwischen Schulleistung und Angst zum einen keinesfalls als gesichert dar, zum anderen abhängig von den verwendeten Meßverfahren, vom Schulfach, vom Geschlecht und vom Alter der untersuchten Schüler. Selbst in den Untersuchungen, in denen ein Zusammenhang aufgedeckt zu sein scheint, lassen sich keine Aussagen über die *kausale* Relation machen. Haben Kinder keinen Schulerfolg, weil sie ängstlich sind? Werden Kinder ängstlich, weil sie keinen Schulerfolg erzielen können? Oder besteht gar der von NICKEL, SCHLÜTER & FENNER vermutete circulus vitiosus: Angst→schlechte

Leistungen→mehr Angst→noch schlechtere Leistung... Hier liegt noch ein weites unerforschtes Gebiet vor uns.

Das Dilemma der bisherigen Untersuchungen liegt in der Bestimmung der Ängstlichkeit ausschließlich durch Fragebögen. Die Wahrscheinlichkeit ist gering, daß das verbale Fragebogenverhalten übereinstimmt mit dem Verhalten in realen Situationen, besonders dann, wenn es um sozial so wenig anerkannte Eigenschaften wie Ängstlichkeit geht. Wahrscheinlicher ist die Beantwortung des Fragebogens weitgehend nach „Social Desirability", besonders für diejenigen Kinder, die die Fragebogenbeantwortung selbst bereits als angsterzeugende Belastung empfinden. Eine umfassende Erörterung zu dem Angst-Konstrukt sowie auch zu diesem Teilproblem findet sich bei GÄRTNER-HARNACH (1972). Unsere Annahme wird gestützt durch die hohen Lügenwerte, die die Validitätsergebnisse von TEWES (1973) stark beeinträchtigen. Solange sich jedoch auch nur der kleinste Hinweis für einen negativen Zusammenhang zwischen ängstlichem Verhalten und Schulleistung ergibt, sollte der Lehrer Erzeugung von Angst (z. B. durch Androhung von Strafe oder durch Überforderung) vermeiden.

2.2.2.2.2. Selbstachtung

In engem, möglicherweise verursachendem Zusammenhang zu Angst oder ängstlichem Verhalten steht diejenige Haltung eines Menschen sich selbst gegenüber, die zusammenfassend „Selbstachtung" genannt wird. Da der Zusammenhang zwischen Schulleistung und Werten in verschiedenen Angstfragebögen bisher nicht eindeutig geklärt ist, bietet das Vorgehen von SEITZ einen erfolgversprechenden Ansatz. Er geht bei seinen bereits zitierten Untersuchungen über den Zusammenhang zwischen Persönlichkeitsvariablen und Intelligenzleistungen einerseits sowie Schulleistungen andererseits von 12 zu „Typen" zusammengefaßten Dimensionen einer deutschen Fassung des Childrens Personality Questionnaire (CPQ) nach PORTER & CATTELL (1963) aus und berechnet verschiedene Faktorenanalysen, in die diese Persönlichkeitsdaten, verschiedene Intelligenzwerte und Schulzensuren eingehen (SEITZ & LÖSER 1969, SEITZ & METZELDER 1970, SEITZ 1970a, SEITZ 1970b, SEITZ 1971). Die für unsere Fragestellung interessanteste Analyse ist die simultane Faktorenanalyse von Persönlichkeitsdaten und Schulleistungen.

Es zeigen sich dabei deutlich positive Zusammenhänge zwischen „mißtrauischer Angst" und „Feinfühligkeit" auf der einen Seite und verbalen Leistungen auf der anderen. SEITZ deutet diesen Leistungsvorteil Ängstlicher vorwiegend als Kompensation von Ichschwäche; die besseren sprachlichen Differenzierungsmöglichkeiten sind dabei auch auf die höhere Sensibilität dieser Personengruppen zurückzuführen. Für die schulischen Leistungen des

sprachlichen Ausdrucks erweisen sich jedoch die beiden Persönlichkeitskomponenten „Erwartung von Leistungserfolg" und „Selbstüberzeugung" als wesentlicher. Über das Ursache-Wirkungs-Verhältnis dieser positiven Korrelation läßt sich zunächst nichts aussagen. Ist der Schulerfolg von Kindern größer, weil sie Leistungserfolg erwarten (vgl. Leistungsmotivation) und von sich selbst überzeugter sind als andere, oder ist die Erwartung von Leistungserfolg und die Selbstüberzeugung von Kindern größer, weil sie Erfolg in der Schule haben? Dieselbe Frage mußten wir uns schon bei den Zusammenhängen zwischen Leistungsmotivation und Schulleistung sowie zwischen Schulangst und Schulleistung stellen. Eine Kreisprozeßwirkung erscheint uns die wahrscheinlichste Antwort auf diese Fragen zu sein, so daß auch bei der Selbstachtung nicht von einer überdauernden, sondern eher von einer durch Erfolg modifizierbaren Persönlichkeitseigenschaft gesprochen werden kann. Die Ergebnisse von SEITZ belegen diese Annahme dadurch, daß die Kovarianz der Schulnoten mit der Persönlichkeitsdimension „Selbstschätzung" allgemeiner und deutlicher ist, wenn sie nicht gleichzeitig mit auf die HAWIK-Subtests bezogen ist. Diese noch engere Beziehung der „selbstüberzeugten Erfolgserwartung" zu den Schulnoten im Vergleich zu den Intelligenztestwerten spricht für unsere Kausalinterpretation, wonach das Ergebnis schulischen Erfolges (das seinerseits entsprechende Intelligenz voraussetzt) für die durch SEITZ aufgewiesenen Kovariationen stark mitverantwortlich ist. Bestreben des Lehrers sollte es demnach sein, den beschriebenen Kreisprozeß in Gang zu setzen, d. h. dem Schüler zunächst unabhängig von den geforderten Lernzielen Erfolge zu vermitteln, die die „Selbstschätzung" fördern und dadurch wiederum zu besseren Leistungen im Sinne der Lernziele führen.

2.2.2.2.3. Extraversion

Mit Extraversion wird gemeinhin die Summe einer Reihe von Eigenschaften des „extravertierten Persönlichkeitstypus" bezeichnet, dessen Interessen sich nach außen (auf andere Menschen und auf die übrige Umwelt) richten. In EYSENCKS Persönlichkeitssystem (EYSENCK & RACHMAN 1970) stellt die Dimension „Extraversion/Introversion" einen auf dem Typenniveau gesicherten Persönlichkeitsfaktor dar; hier scheint eine der eingangs in Frage gestellten „überdauernden Persönlichkeitsvariablen" nachgewiesen zu sein. Eine intensive Stellungnahme zu dem EYSENCKschen Konzept würde in diesem Beitrag zu weit führen. Es soll hier nur soviel gesagt werden, daß EYSENCK mit viel zu kleinen Versuchspersonen-Stichproben gearbeitet hat, um Repräsentativaussagen machen zu können. Außerdem hat sich in Kontrolluntersuchungen ergaben, daß die beiden EYSENCK-Faktoren „Neurotizismus" und „Extra- versus Introversion" nicht unabhängig voneinander

sind, so daß nicht eindeutig von eigenständigen Persönlichkeitsvariablen gesprochen werden darf.

Interessant an dem Ansatz von EYSENCK ist die durch zahlreiche Untersuchungen gestützte Theorie der unterschiedlichen Erregungs- und Hemmungsprozesse bei Intro- und Extravertierten. Unter „Erregung" versteht er dabei im großen und ganzen „eine Aktivierung des Kortex' und eine allgemeine Erleichterung des Lernens, Erinnerns und Handelns" (EYSENCK 1970, S. 37). „Hemmung" definiert er global als einen Prozeß des ZNS (Zentralnervensystem), der mit den ablaufenden kognitiven, motorischen und perzeptiven Tätigkeiten des Organismus interferiert, also diese stört. Schnell und stark ablaufende Hemmungsprozesse werden nach EYSENCK vielfach bei Menschen gefunden, die hohe Extraversionswerte in dem von EYSENCK entwickelten Fragebogen zur Extra-/Introversion erreichen. Stark Introvertierte neigen dagegen eher zu entsprechenden Erregungsprozessen. Die Bedingungen für Lernprozesse sind demnach bei mehr introvertierten Menschen günstiger als bei mehr extravertierten. Allerdings beziehen sich nur wenige von EYSENCK zur Stützung seiner Hypothese angeführten Untersuchungen (meist eigene oder aus seinem Institut hervorgegangene) auf die Hemmung bzw. Aktivierung rein kognitiver Leistungen. Ob also aufgrund dieser Laborergebnisse auch Aussagen über den Zusammenhang von Lernleistungen in der Schule und Extra- bzw. Introversion gemacht werden können, ist fraglich. TEWES (1973) hat in seiner Validierungsstudie zur HANES (K), einer Übersetzung und Zusammenfassung der Skalen vom Junior EYSENCK Personality Inventory (JEPI) von Sybil EYSENCK (1965) auf die Frage des Zusammenhangs zwischen schulischen Lernleistungen und der nach EYSENCK erfaßten Extraversion eine Antwort zu finden versucht. Die Hypothese EYSENCKS, daß hoch Extravertierte schwache Schulleistungen, hoch Introvertierte dagegen eher gute Schulleistungen erbringen, konnte nur tendenziell und auch nur bei Mädchen bestätigt werden. Außerdem zeigte sich ein moderierender Einfluß des Neurotizismus: der Zusammenhang zwischen Extraversion und Leistung ist bei Kindern mit hohen Neurotizismuswerten anders als bei wenig neurotischen Kindern. Das Ergebnis läßt sich also nicht eindeutig interpretieren. TEWES vermutet, daß die Emotionalität (Extra/Introversion) weniger mit der absoluten Leistungshöhe als vielmehr mit dem Ausmaß der Leistungsschwankung, also mit der Arbeitshaltung korreliert (vgl. die Zusammenstellung von Einzelbefunden zur Symptomatologie bei Over- und Underachievern von WEINERT 1965 und KOSCHAT 1966). Auch RANK (1962) konnte in ihrem Vergleich von „guten" mit „schlechten" Schülern feststellen, daß dem Lehrerurteil gemäß jeweils mehr „gute" als „schlechte" Schüler konzentriert, beherrscht, stetig, ordentlich, aktiv, engagiert und selbständig arbeiteten (sie verglich allerdings ausschließlich Prozentwerte; über das Signifikanzniveau der Unterschiede wird nichts ausge-

sagt). Wenn dem aber so ist, brauchen wir nicht von der „schicksalhaften Bestimmung" der Schulleistung durch eine überdauernde Persönlichkeitseigenschaft (hier Extraversion) zu sprechen, sondern können uns auf das Training und die Ausbildung eines günstigen Leistungsverhaltens wie Konzentration, Ausdauer, Sorgfalt, Planung, Selbständigkeit etc. konzentrieren und somit die Leistungen in der Schule positiv beeinflussen.

2.2.2.3. Lehrerverhalten

Wie wir im Abschnitt über die Lern- und Leistungsmotivation gesehen haben, kann der Lehrer wesentlich dazu beitragen, die Leistungen seiner Schüler dadurch positiv zu beeinflussen, daß er ihre Bereitschaft zur Mitarbeit durch die richtige Auswahl des Lernstoffes aktiviert. Welche weiteren Abhängigkeiten zwischen Schulleistung und Lehrerverhalten bestehen, wollen wir im folgenden darlegen.

2.2.2.3.1. Setzung „sachfremder" Leistungsmotivation

HECKHAUSEN (1967) unterscheidet die Motivation aufgrund (seiner Meinung nach) zwei verschiedener Genesen: die intrinsische und die extrinsische Motivation. *Intrinsisch* ist die Motivation, wenn die Bewältigung einer Aufgabe oder auch schon die Arbeit an der Aufgabe als angestrebter Erfolg oder als Bekräftigung erlebt wird. Wenn jedoch andere Bekräftigungen durch Leistungsstreben zu erlangen erhofft werden (HECKHAUSEN nennt als Beispiel die Erfüllung des Bedürfnisses nach Identifikation, nach Zustimmung, nach Abhängigkeit und nach Geltung, wobei weder operationalisiert wird, was mit „Bedürfnis", noch was mit diesen speziellen Bedürfnissen gemeint ist), und wenn es dem Schüler darum geht, in Aussicht gestellte Strafen zu vermeiden, so nennt HECKHAUSEN diese Art „sachfremden" Leistungsstrebens *extrinsische* Motivation. Er behauptet, die Lernmotivation würde so in extremem Maße von nicht unmittelbar sachbezogenen Bekräftigungstechniken und damit nicht zuletzt von höchst individuellen Persönlichkeitseigenarten des Lehrers abhängig. In Lernsituationen außerhalb des formalisierten Unterrichts oder in Abwesenheit der motivierenden Lehrperson bliebe die Lernmotivierung aus. Aus lerntheoretischer Sicht bietet sich dagegen eher die Vermutung an, daß sich die sachfremde (extrinsische) Leistungsmotivation über den Weg der nicht allein aufgabenbezogenen Bekräftigung sehr wohl zu einer sachbezogenen (intrinsischen) Motivation stabilisieren läßt, und zwar nach demselben Mechanismus, nach dem sich auch z. B. das „extrinsische Reinlichkeitsstreben" im Laufe der Entwicklung zum „intrinsischen" verselbständigt. Entscheidend ist dabei die Dauer der Koppelung von Leistung und sachfremder Bekräftigung, die für die Stabilisierung der sach-

bezogenen Leistungsmotivation erforderlich ist (zu beachten sind die in diesem Bereich vermutlich großen interindividuellen Unterschiede), sowie ein geeigneter Verstärkungsplan. Am geeignetsten für die Ausbildung „wünschenswerter" Eigenschaften hat sich ein intermittierender Bekräftigungsplan erwiesen (z. B. FOPPA 1966), d. h. zunächst wird jede erwünschte Verhaltensweise (hier Leistungsverhalten) bestärkt, die meiste Zeit der Behandlung hindurch wird die Bekräftigung ab und zu fortgelassen und zum Ende hin werden die Belohnungen langsam immer weiter eingeschränkt. In der Verhaltenstherapie (EYSENCK & RACHMAN 1970, BEECH 1969, WOLPE 1970) wird dieses Prinzip bereits erfolgreich durchgeführt. Außerdem sollte man die Veränderung der sozialen Situation beachten, die für das Kind möglicherweise dadurch eintritt, daß es sich — wenn auch zunächst aufgrund sachfremder Motivationen — leistungsangepaßt verhält. Es wird vermutlich weniger schlechte Noten bekommen und dadurch weniger Sanktionen oder Druck, vielleicht sogar statt dessen mehr Zuwendung oder andere Belohnungen von seiten der Eltern erhalten, was eine zusätzliche Verstärkerwirkung bedeutet. Eine Sonderstellung unter den sachfremden Motivationen gesteht HECKHAUSEN (1967, S. 197) dem Bedürfnis nach Identifikation zu, sofern der Schüler dem Lehrer im Hinblick auf dessen Leistungstüchtigkeit ähnlich werden möchte. „Dies kann recht unmittelbar zu einer Wertschätzungssteigerung von leistungsthematischen Sachbereichen führen und auf die Dauer auch die Leistungsmotivation beeinflussen" (vgl. BANDURA & MISCHEL 1965). Warum gerade dieses „Bedürfnis" einen Sonderfall darstellt, führt er nicht weiter aus. Eine plausible Erklärungsmöglichkeit aufgrund experimenteller Befunde bieten BANDURA & MISCHEL (1965), BANDURA, ROSS & ROSS (1963) und BANDURA (1969). Sie verwenden statt des psychoanalytisch belasteten Begriffs der Identifikation den lediglich beschreibenden Begriff der Imitation bzw. des „Modellearning" (Lernen nach einem Modell). Kinder imitieren Verhaltensweisen eines Modells oder Vorbildes, wenn

a) das Modell für sein Verhalten bekräftigt wird,
b) das Modell eine Machtposition inne hat,
c) das Modell sozial begehrenswerte Eigenschaften aufweist.

Besonders b) trifft für die Rolle des Lehrers in unserem Schulsystem zu, so daß die Wahrscheinlichkeit groß ist, daß die Schüler Verhaltensweisen des Lehrers nachahmen.

Nach TAUSCH (1972) ist eine zusätzliche Bedingung für das Imitationslernen die „persönliche Glaubwürdigkeit" des Lehrers für den Schüler und nach KLAUSMEIER & GOLDWIN (1966) der „Enthusiasmus" sowie „beherrschtes Engagement" des Lehrers. Bleibt er impulsiv und unstet, dann profitieren nur Hochbegabte, während die anderen eher verschüchtert werden.

Die Leistungsmotivation stellt sich somit als eine erlernbare Eigenschaft dar. Für die frühe Kindheit sieht auch HECKHAUSEN (1971, S. 201) ihre Genese in dieser Weise, wenn er schreibt: „Die Übermittlung von Motivausprägungen von der älteren auf die heranwachsende Generation geht nicht nur über die Erziehungsmittel von Lohn und Strafe oder von selbständigkeits- oder abhängigkeitsfördernden Maßnahmen. Die Kinder lernen auch unmittelbar am Vorbild des Erziehers ...", oder wenn er frühkindliche Trainingsaufgaben zur Leistungsmotivation in Form von Comics anbietet, in denen es um für die Kinder besonders begehrenswerte Personen geht. HECKHAUSEN glaubt jedoch, daß sich die Leistungsmotivation bereits bis zum Schuleintrittsalter unveränderbar zu einer extrinsischen Leistungsmotivation verfestigt habe, und danach nur noch nichtüberdauerndes extrinsisches Leistungsstreben angeregt werden könne. Wir meinen dagegen aufgrund der Befunde aus der Verhaltenstherapie, daß es in der oben angeführten Weise mit einem geeigneten Langzeittraining auch noch nach dem Vorschulalter möglich sein müßte, die Leistungsmotivation von Schülern positiv zu beeinflussen, d. h. daß ein auf ein solches Motivationstraining ausgerichtetes Lehrerverhalten die Schulleistung auch auf Dauer positiv beeinflussen kann.

2.2.2.3.2. Kognitive Stile

Sogenannte kognitive Stile im Sinne von Denkmustern, Abfrage- und Assoziationsmustern sowie Problemlösungsstrategien oder die Art und Weise, wie ein Kind Informationen aufnimmt und verarbeitet und welche Probleme es überhaupt als solche wahrnimmt, all dies wird bereits in früher Kindheit relativ fest ausgebildet.

HESS & SHIPMAN (1965) konnten schon bei Vierjährigen diesbezügliche Unterschiede feststellen, die mit der sozialen Schichtung einhergingen. Über die Genese wissen wir noch recht wenig; sie wird jedoch ebenso nach Lernprinzipien (durch Erfolg und Mißerfolg, Bekräftigung und Bestrafung, Imitation und Identifikation) erfolgen, wie die Entwicklung anderer nichtangeborener Verhaltensweisen. Es ist z. B. einleuchtend, daß ein Kind, das frühzeitig viel von seiner Umwelt visuell wahrnehmen kann, Lösungen später eher im Anschaulichen sucht, während ein anderes Kind, mit dem viel gesprochen wurde, noch ehe es selber verbalisieren konnte, lieber einen verbalen Lösungsweg vorziehen wird. In der Schule besteht dann leicht die Schwierigkeit, daß der Lehrer andere Denkmedien bevorzugt als der Schüler. Eine Resonanz zwischen Lehrer und Schüler ist dadurch erschwert, und der Schüler erzielt schlechtere Leistungen als bei einem Lehrer, dessen Abfragemuster mit dem Assoziationsmuster des Schülers übereinstimmt.

Neben den verschiedenen Denkmedien spielt vermutlich ebenfalls die Denkdynamik eine entscheidende Rolle für den Erfolg im kognitiven Be-

reich. KAGAN u. a. (1964) unterscheiden zwischen Impulsivität und Reflexivität bei der Lösung von Aufgaben und konstatieren einen störenden Einfluß des impulsiven Stils auf das schulische Lernen durch vorschnelles und darum fehlerhaftes „Drauflosiegen". Das Verhalten des Lehrers sollte unseres Erachtens nicht allein darauf angelegt sein, bei den Schülern „günstige kognitive Stile" (HECKHAUSEN 1971) — gemeint sind damit nur die Stile, die für die konvergenten Leistungen in der Schule förderlich sind — herauszubilden, sondern ihnen möglichst viele unterschiedliche Stile und deren Anwendungsbedingungen zu vermitteln (da wo z. B. Kreativität, d. h. divergentes Denken gefordert wird, ist Impulsivität erheblich „günstiger" als Reflexivität allein). Zudem sollten Pädagogen versuchen, sich auf die jeweiligen kognitiven Stile ihrer Schüler einzustellen und ihren Unterrichtsstoff auf verschiedene Stile abzustimmen. Untersuchungen über derartige Resonanzoptimierungen zwischen den kognitiven Stilen von Lehrern und Schülern liegen nicht vor.

2.2.2.3.3. Unterrichtsatmosphäre

Besonders in jüngster Zeit beschäftigen sich etliche Untersuchungen mit der Interaktion zwischen verschiedenen emotionalen Verhaltensweisen von Lehrern und den Lernfortschritten ihrer Schüler. Ein interessantes Experiment zu diesem Fragenbereich nahm BERNSTEIN bereits 1957 an Tieren vor. Er ordnete per Zufall die gleiche Anzahl von Ratten je einem „Versuchsleiter" zu, der den Tieren ein starkes Ausmaß von „handling" (Zuwendung) zukommen ließ, einem mit mittlerem „handling" und einem, der gar keine Zuwendung entgegenbrachte. Nach 100 Lernversuchen im Labyrinth erreichten die Tiere, die ein großes Ausmaß an liebevollem Kontakt erhalten hatten, zu 70 % das Lernkriterium. Ratten mit mittlerem erhaltenen „handling" erreichten das Kriterium nur zu 5 % und die Tiere ohne liebevolle Zuwendung zeigten gar keinen Lernerfolg.

Dieses Experiment ist unter anderem eine Erklärung für den sog. ROSENTHAL-Effekt. ROSENTHAL (1968) hatte zwei Gruppen von „Versuchsleitern" je eine Anzahl gleich lernfähiger Ratten zum Einüben eines bestimmten Verhaltens ausgehändigt. Der einen „Versuchsleiter"-Gruppe wurde gesagt, sie hätten es mit besonders intelligenten Ratten zu tun, bei der anderen Gruppe wurden die Tiere als besonders dumm bezeichnet. Tatsächlich erwiesen sich nach dem Lernversuch die Ratten der ersten Gruppe denen der zweiten in ihren Lernleistungen überlegen. Das Resultat von BERNSTEIN läßt nun vermuten, daß die „Voreingenommenheit", es mit besonders intelligenten bzw. dummen Ratten zu tun zu haben, die „Versuchsleiter" in ihrem Ausmaß an emotioneller, bekräftigender Zuwendung da-

hingehend beeinflußt hat, daß „intelligente" Ratten mehr und „dumme" Ratten weniger in ihrem Lernverhalten bestärkt wurden.

Die hier interessierende Frage ist, inwieweit diese Ergebnisse auch auf Lehrer- und Schülerverhalten übertragbar sind. Eine ausführliche Zusammenstellung von Untersuchungen, die sich mit der Auswirkung der emotionalen Dimension im Erzieher- oder Vorgesetztenverhalten auf die Verhaltensmodifikation bei Kindern oder Untergebenen befassen, findet sich bei TAUSCH & TAUSCH (1970).

Danach korreliert in der Untersuchung von RYANS (1961a) freundliches, verstehendes, warmes Lehrerverhalten besonders bei Grundschülern positiv mit sog. produktivem Schülerverhalten wie aktiver Teilnahme am Unterricht, Selbstkontrolle (wieso letzteres als produktiv bezeichnet wird, sei dahingestellt!). Der eigentliche Lernerfolg wird jedoch nur geringfügig in dieser Untersuchung durch das freundliche Unterrichtsklima begünstigt. In der Untersuchung von CHRISTENSEN (1960) wurde ein positiver Zusammenhang zwischen der emotionalen Wärme der Lehrerin und dem Leistungsgewinn der Schüler in Sprachen und Arithmetik deutlich. Das Ergebnis ist unseres Erachtens jedoch nicht eindeutig interpretierbar, da das Lehrerverhalten (die emotionale Wärme des Lehrers) durch die Schüler selbst anhand von 40 Fragebogenitems eingeschätzt wurde. Das subjektive Erleben eines Lehrers durch den Schüler ist jedoch abhängig von vielen zusätzlichen Schülerfaktoren, z. B. von früheren Lehrererfahrungen in diesem Schulfach, vom Interesse an diesem Schulfach, von den Leistungen in diesem Schulfach u. ä. Ein im Fach Englisch guter und dafür interessierter Schüler wird etwa seine Englischlehrerin in diesem Fragebogen von CHRISTENSEN positiver einstufen als ein in diesem Fach schlechter und uninteressierter Schüler. Hier erhebt sich der Streit um die Henne und das Ei: Was war zuerst da, die Freundlichkeit des Lehrers oder die Fähigkeit des Schülers? Wir meinen, auch in diesem Fall müßte wohl von einer Wechselwirkung gesprochen werden.

Etwas besser geplant im Sinne der Eindeutigkeit der Resultate sind die zahlreichen Untersuchungen von TAUSCH und verschiedenen Mitarbeitern (siehe in TAUSCH & TAUSCH 1970). Hier wurde das Verhalten der Lehrer von neutralen „ratern" (Einschätzern) beurteilt, so daß Lehrereinschätzung und Schülervariablen unabhängig voneinander waren. In diesen Untersuchungen zeigte sich in erster Linie ein positiver Zusammenhang zwischen Verhaltensdimensionen des Lehrers („emotionale Wärme, Wertschätzung, Zuneigung") einerseits und positiven gefühlsmäßigen Erfahrungen des Schülers, positiver Lehrer-Schüler-Beziehungen, der Förderung seelischer Reife (es wird keine Erläuterung gegeben, was damit gemeint ist), der Beliebtheit des Lehrers sowie der Selbständigkeit des Kindergartenkindes im Sprachverhalten andererseits. Außerdem zeigte sich vielfach eine Minderung der

Schulangst durch das oben beschriebene Lehrerverhalten. Direkte Beziehungen zum Leistungsverhalten des Schülers wurden nicht aufgewiesen.

Bei McKeachie (1961) dagegen zeigte sich eine derartige Beziehung zu Schulnoten. Er fand seine Hypothese gestützt, daß Studenten, die sich in ihrem Kontaktstreben (dieses wurde allerdings mit dem TAT gemessen, Kritik dazu siehe S. 65) voneinander unterscheiden, unterschiedlich durch die „Wärme" oder die Freundlichkeit des Dozenten beeinflußt würden. Die unterschiedliche Beeinflussung machte sich in den Noten bemerkbar. Dabei war der Einfluß bei den beiden Geschlechtern genau konträr: bei kontaktbedürftigen Studentinnen bewirkte größere Freundlichkeit bessere Zensuren, kontaktbedürftige Studenten andererseits schnitten in den Kursen mit größerer Lehrerfreundlichkeit schlechter ab. Anders, als daß das Maß für Kontaktbedürfnis bei beiden Geschlechtern etwas Unterschiedliches bedeutet (d. h. bei Studentinnen wird es durch Dozentenfreundlichkeit erfüllt, bei Studenten nicht), können wir uns diesen Unterschied nicht erklären.

Aufgrund all dieser Untersuchungsergebnisse läßt sich kein einfaches „Kochrezept" wie „Eine freundliche Unterrichtsatmosphäre durch emotional zugewandtes Lehrerverhalten fördert das Leistungsverhalten der Schüler" oder „Eine freundliche Unterrichtsatmosphäre beeinflußt die Schulleistungen nicht positiv" angeben. Als eindeutig feststehend kann jedoch die Tatsache angesehen werden, daß emotional zugewandtes, „warmes" Lehrerverhalten die Schulleistungen nicht negativ verändert, auch nicht bei Kindern, die ein mehr gegensätzliches Erzieherverhalten gewohnt sind. Inwieweit die Leistungsmodifizierung in positive Richtung möglich ist, scheint den Untersuchungsergebnissen zufolge von zusätzlichen Variablen abzuhängen (Geschlecht, Kontaktbedürfnis, Arrangement des Unterrichtsstoffes, Alter usw.).

2.2.2.3.4. Direktives Verhalten

Als eine zweite Hauptdimension des Lehrerverhaltens neben der emotionalen Zuwendung hat sich in zahlreichen, z. B. bei Tausch & Tausch (1970) zusammengestellten, Untersuchungen die Direktivität des Verhaltens im Sinne von direkt beobachtbarer Lenkung bzw. Führung herausgestellt.

Die Auswirkung dieser Dimension des Erzieherverhaltens zeigt sich — wie auch in sozialpsychologischen Experimenten zum Führungsverhalten (z. B. Sheriff [1955], French, Israel & As [1960], Lippit u. White [1952] u. a.) — abhängig vom jeweiligen Aufgabentyp. Für die Übermittlung bzw. Aneignung von Kenntnissen und Fertigkeiten erwies sich der mehr lenkende Unterrichtsstil als effektiver, bei Aufgaben jedoch, die eigenständige Denkarbeit erfordern (z. B. für Problemlösungen und kreatives Lernen), ist ein nichtdirektes Verhalten des Lehrers förderlicher als ein direktives.

NICKEL, SCHLÜTER & FENNER (1973) deckten zudem einen hochsignifikanten Zusammenhang zwischen Dirigismuswerten von Lehrern (gemessen mit dem Fragebogen zur direktiven Einstellung [FDE] von BASTINE 1971) und der Qualität der Unterrichtsmitarbeit sowie der Konzentration der Schüler (gemessen mit dem d2-Test von BRICKENKAMP 1962) auf. Konzentrationsleistungen nehmen demnach mit steigender Direktivität ab. Hinsichtlich der Mitarbeit reagieren Jungen und Mädchen unterschiedlich auf direktives Verhalten des Lehrers: Mädchen beteiligen sich zunehmend mehr am Unterricht, Jungen dagegen zunehmend weniger. Ohne daß die Schüler darauf Einfluß haben, hängen ihre Leistungen (besonders Konzentrationsleistungen) demnach mit davon ab, wie direktiv der Lehrer den Unterricht gestaltet, d. h. wieweit er die Schüler in ihrer Eigeninitiative einengt.

Wenn man jedoch nicht nur allein den Leistungsaspekt beachtet, sondern auch das Erleben der Schüler, das z. B. für die Motivierung zu Leistungsverhalten in der Schule bedeutsam ist, so ließ sich in allen Untersuchungen feststellen, daß eine Verminderung von Lenkung und mehr Gewährung von Freiheiten den Schüler zufriedener sein ließen. Unter mehr nichtdirektiven Bedingungen erleben Kinder den Unterricht weniger als lästige Pflicht, sondern vielmehr als Bekräftigung, die ihre Interessen auch noch über die Schulstunde hinaus für den Unterrichtsgegenstand aktiviert (s. THIERSCH 1971). Es ist also zunächst zu fragen, welche Lernziele angestrebt werden sollen, um dann das dafür effektivste Lehrerverhalten hinsichtlich der Dimension Lenkung/Dirigismus bestimmen zu können. Zudem ist zu bedenken, daß diese Verhaltensdimension „Lenkung/Dirigismus" im statistischen Sinne nicht unabhängig ist von der nach TAUSCH & TAUSCH (1970) benannten Hauptdimension „emotionale Zuwendung", beide korrelieren signifikant miteinander (TAUSCH 1970). Man kann demnach nicht die Aussage machen, daß bei Eigenständigkeit und Kreativität erfordernden Arbeiten ein nichtdirektives Erziehungsverhalten leistungsfördernder sei als ein direktives, ohne den leistungshemmenden oder zufriedenheitsmindernden Anteil des emotional abgewandten Verhaltens an dem nicht-direktiven zu beachten. Es ließe sich z. B. sehr gut vorstellen, daß ein direktives Verhalten bei gleichzeitig hohem Ausmaß an emotionaler Zuwendung ebenso effektiv für Leistung und Zufriedenheit ist wie ein nicht-direktives Führungsverhalten; oder umgekehrt, daß ein nicht-direktives Verhalten mit gleichzeitig verwirklichter emotionaler Abwendung ebenso wie das in den bisherigen Untersuchungen als direktiv bezeichnete Lehrerverhalten weder für Kreativitätsleistungen noch für die Zufriedenheit der Schüler vorteilhaft ist.

Als weitere Variable für die Effektivität der Leistungsförderung des Schülers durch den Lehrer ist die vorherige Erziehungserfahrung des Schülers zu nennen. Kinder, die bisher von ihren Eltern stark dirigistisch eingeengt worden waren, können nicht von einem nicht-direktiven Lehrerverhal-

ten profitieren (z. B. THIERSCH 1971), es sei denn, der Wechsel würde langsam schrittweise vollzogen und der weitere dirigistische Elterneinfluß wäre auszuschließen. Kinder mit starken Lenkungs- und Führungserfahrungen werden durch die plötzliche Nicht-Direktivität des Lehrers verunsichert statt zu Eigeninitiative initiiert. Es erweist sich demnach auch an diesem Punkt unserer Erörterungen, daß es nicht *ein* effektives Lehrerverhalten gibt, sondern ein komplexes Repertoir verschiedener je individuell einzusetzender Verhaltensformen, um unterschiedlichen Schülern in unterschiedlichen Situationen gerecht werden zu können. Es besteht ohne Zweifel ein Bedingungszusammenhang zwischen den Schulleistungen eines Schülers und dem Lehrerverhalten, nur läßt er sich nicht in eine allumfassende Formel fassen, sondern stellt sich unter variablen Bedingungen immer wieder anders dar.

2.2.2.3.5. Werthaltungen

In unserer einleitenden Diskussion über die Definition der Schulleistung und ihre Messung wiesen wir bereits darauf hin, daß bei der Bestimmung der Leistung eines Schülers durch den Lehrer breiter Spielraum besteht für die subjektive, den jeweiligen Werthaltungen und Vorurteilen des Lehrers entsprechende Interpretation und Bewertung der Schülerleistungen. So wird möglicherweise zwar nicht die „wahre" erbrachte Leistung durch Lehrerverhalten beeinflußt, wohl aber die Erscheinungsweise, also die Wirkung im sozialen Umfeld. Die Zensurengebung und vor allen Dingen die Auslese für weiterführende Schulen wird weitgehend, wie z. B. LATSCHA (1963) in seiner Schweizer Untersuchung und STEINKAMP (1972) in der BRD nachweisen konnten, von charakterlichen Qualitäten wie Disziplin, Sauberkeit, ordentlichem Betragen usw. und weniger von der „wahren" Leistung bestimmt (weitere Erläuterungen zur Notengebung und zum Problem der subjektiven Leistungsbewertung finden sich in den nachstehenden Beiträgen von LANGHORST, FINGERHUT & LANGFELDT, NICKEL & WIECZERKOWSKI und WENDELER). Beachtenswert in diesem Zusammenhang erscheint uns die von KEMMLER (1967) und von AMELANG & KÜHN (1972) festgestellte Wechselwirkung zwischen dem Geschlecht des Lehrers und dem des Schülers in bezug auf die Leistungsbeurteilung. Schüler gleichen Geschlechts wie der Lehrer werden demnach schlechter beurteilt als Schüler des anderen Geschlechts. Einen Hinweis für die Abhängigkeit der Schulleistungsbeurteilung von Werthaltungen und Einstellungen der Lehrer gibt auch die Untersuchung von STEINKAMP (1972) zum Einfluß der Herkunftsschicht der Lehrer auf die Schülerbeurteilung. Volksschullehrer aus unteren sozialen Schichten messen der Herkunftsschicht im Hinblick auf den Erfolg an höheren Bildungsinstitutionen überdurchschnittliche Bedeutung bei und empfehlen über-

durchschnittlich viele Schüler aus den oberen sozialen Schichten für die Oberschule — jeweils im Vergleich zu ihren aus allen übrigen Sozialschichten stammenden Kollegen. Methodisch und statistisch ist zwar an dieser Arbeit mancherlei Kritik zu üben, die inhaltliche Tendenz der Untersuchungsergebnisse sollte jedoch nicht außer acht gelassen werden.

Die Forschung zum Einfluß von Persönlichkeitseigenschaften, Verhaltensweisen, Einstellungen und Werthaltungen der Lehrer auf die Leistungen ihrer Schüler sollte unseres Erachtens erheblich intensiviert werden, da hier ein vielversprechender Ansatz für die optimale Förderung von Schulleistungen liegt. In Deutschland waren solche Erörterungen, von einigen Ausnahmen abgesehen, meist unergiebig. Einige blieben moralisierend appelativ, andere sind politisch ideologisch belastet. Beide Arten der Vorgehensweise führen zu praxisfernen Pauschalierungen. Wir meinen, Leistungen der Schüler müssen nicht allein durch Veränderung der Schüler, sondern auch sehr wohl durch Anpassung der Lehrer bzw. durch die richtige Auswahl der Lehrer für die entsprechenden Schüler gesteigert werden. Das Ausmaß der Effektivität der hier aufgeführten Dimensionen des Lehrerverhaltens kann aber nicht jeweils generell beantwortet werden. Jeder Aspekt muß in sich wieder spezifiziert werden im Hinblick auf verschiedene Schüler, Situationen und Aufgaben. „Die Wirkung des Lehrerverhaltens ist verschieden in unterschiedlicher historischer, sozio-kultureller Umgebung und in unterschiedlichem Anspruchsniveau der Schulen, in verschiedenen Schulorganisationsformen und unter Streß oder in Spontaneität. Die Wirkung des Lehrerverhaltens ist verschieden je nach den unterschiedlichen sozialen und psychologischen Voraussetzungen der Schüler, nach den verschiedenen schichtspezifischen Sprach- und Verhaltensmustern und nach der Diskrepanz zwischen den Verhaltensmustern der Schüler und denen der Lehrer; sie ist verschieden nach Geschlecht und Alter der Schüler und vor allem nach ihren aus dem Elternhaus und den bisherigen Lernerfahrungen vorgeprägten Verhaltensstrukturen, der Autoritätsanfälligkeit, dem Selbstvertrauen, der Ängstlichkeit, der Leistungsmotivation. Die Wirkung des Lehrerverhaltens ist verschieden in bezug auf die unterschiedlichen angestrebten Leistungen, auf die verschiedenen Lerninhalte und Fächer und vor allen Dingen in bezug auf die Beherrschung eines vorgegebenen Stoffes oder ein offenes problemlösendes Verhalten“ (Thiersch 1971, S. 483 f.).

2.2.2.4. Soziokulturelles Milieu (Familie, Peergroups)

Im vorausgegangenen haben wir mehrfach darauf hingewiesen, daß verschiedene Determinanten der Schulleistung wiederum ihrerseits von soziokulturellen Bedingungen wie Erzieherverhalten der Eltern, Werte und Normen der sozialen Gruppe, der das Kind angehört, äußeren Lebensbedingun-

gen, Arbeitsmöglichkeiten u. dgl. determiniert sind. Wir wollen hier die festgestellten Zusammenhänge zwischen Schulleistung und Milieuvariablen schildern.

2.2.2.4.1. Äußere, objektive Bedingungen

Es steht außer Frage, daß räumliche und akustische Bedingungen Einfluß ausüben auf Leistungsverhalten, in unserem Fall z. B. auf die Erledigung der Hausaufgaben und somit auf die Schulleistung. Es steht ebenfalls außer Frage, daß die räumlichen und akustischen Verhältnisse bei Familien der unteren Einkommensklassen, die meist mit mehr Personen zusammen auf engerem Raum wohnen als andere Gruppen, weniger günstig sind als bei Familien höherer Einkommensklassen. Dadurch haben die Unterschichtkinder schon von den Arbeitsbedingungen her geringere Chancen für gute Leistungen als Oberschichtkinder (Unter- und Ober- bzw. Mittelschicht wird in den meisten Untersuchungen zu diesem Themenbereich definiert durch die Höhe des Einkommens bzw. den Berufsstatus des Familienoberhauptes, eine nicht ganz befriedigende Klassifikation). KOSCHAT (1966) konnte z. B. in seiner Untersuchung an „positiv" und „negativ diskrepanten" Schülern (Over- und Underachiever) feststellen, daß die negativ diskrepanten Schüler im Vergleich zu einer Kontrollgruppe in beengteren und ungemütlicheren Wohnverhältnissen leben und nicht ungestört arbeiten können. Ähnliche Feststellungen machten auch BEER, KUTALEK & SCHNELL (1968), RÖSLER (1967), KEMMLER (1967) und RANK (1962). Auch zwischen der Gestaltung des Nachmittags der Kinder und ihren Schulleistungen erwies sich ein bedeutsamer Zusammenhang in den Untersuchungen von BEER, KUTALEK & SCHNELL (1968), RANK (1962) und KEMMLER (1967). Schüler mit guten Schulleistungen halten sich demnach (abgesehen von der durchschnittlichen Spielzeit) häufiger zu Hause auf als solche mit schlechten Leistungen. Mit diesem Ergebnis ist jedoch noch keineswegs bewiesen, daß ein längerer bzw. häufigerer Aufenthalt im Elternhaus für die Schulleistungen förderlicher ist als ausgedehnte Aufenthalte außerhalb des Elternhauses. Es könnte durchaus sein, daß die guten Schüler der zitierten Untersuchungen noch bessere Leistungen erbringen könnten, wenn sie mehr Anregungen durch häufigeren Aufenthalt außerhalb des Elternhauses hätten, oder daß die schlechten Schüler durch Beschränkung ihrer Außenaufenthalte noch geringere Leistungen erbringen würden. In den zitierten Untersuchungen wurden lediglich Gewohnheiten von guten und schlechten Schülern dargestellt; ob diese Gewohnheiten mitbestimmend für die erbrachten Leistungen sind, muß weiter analysiert werden. Ebenso kritisch sollten in diesem Zusammenhang Untersuchungsergebnisse (z. B. von RANK 1962) betrachtet werden, die den Zusammenhang von Schulleistung und Lebensgewohnheiten, wie Zeitpunkt

der Schularbeiten, Dauer der Schularbeiten, Hilfeleistungen im Haushalt, Teilnahme an außerschulischen Veranstaltungen, Kinobesuch, Freizeitlektüre usw., darstellen.

Eine weitere, mehrfach untersuchte Umweltvariable in bezug auf die Schulleistung ist die Familienkonstellation. ZIELINSKI (1966) machte folgende Feststellungen: Einzelkinder und ältere Geschwister aus Familien mit zwei Kindern haben z. T. signifikant bessere Leistungen aufzuweisen als ältere Geschwister aus Familien mit drei Kindern (siehe auch RANK 1962 u. KEMMLER 1967). Älteste und jüngste Mädchen erzielen signifikant bessere Schulleistungen als Mädchen in mittleren Positionen. Ältere Kinder aus Zweikinderfamilien mit einer jüngeren Schwester sind den älteren Geschwistern mit einem jüngeren Bruder in bezug auf Intelligenz, Schulleistung und schulische Leistungsbeurteilung überlegen. Kinder, die bis zum 11. Lebensjahr Vater oder Mutter verloren hatten, erzielten signifikant geringere Mittelwerte im Intelligenztest und im Schulleistungstest als Kinder ohne solche Verluste. Kinder dagegen, deren Eltern im 1.—6. Lebensjahr den Verlust eines Elternteils zu beklagen hatten, hatten im Schulleistungstest einen signifikanten Leistungsvorsprung gegenüber einer Kontrollgruppe. ZIELINSKI bietet verschiedene auf Theorien zum Verhalten in Sozialrollen basierende Interpretationsversuche für seine Ergebnisse an, die wir hier jedem Leser selbst überlassen möchten.

Für bedeutsamer jedoch als die dargestellten, äußeren Umweltbedingungen erachten wir die bedauerlicherweise nur schwer objektiv erfaßbaren Milieuvariablen wie Erziehungsverhalten, Wertvorstellungen und Normen innerhalb der sozialen Gruppe, der das Kind angehört.

2.2.2.4.2. Werthaltungen, Normen

In dem Abschnitt zur Leistungsmotivation und Arbeitshaltung schnitten wir bereits das Problem der unterschiedlichen Leistungsziele bei verschiedenen Schülern an. Welche Ziele durch eigene Anstrengungen und Leistungen erreicht werden sollen, ist eine Frage der Werteinstellung. Es wäre banal zu betonen, daß Ziele, die nicht bedeutsam, prestigeträchtig, interessant oder andersartig wertvoll für den einzelnen sind, auch nicht erstrebenswert erscheinen.

In zahlreichen Untersuchungen von Schülern und ihren Eltern zeigten sich bedeutsame Unterschiede zwischen Angehörigen höherer und niederer Sozialschichten (hoch und niedrig wiederum meist lediglich definiert durch die Höhe des Familieneinkommens bzw. den Berufsstatus) hinsichtlich solcher Zielvorstellungen. Ober- und Mittelschicht vertrauen eher ihrem individuellen Einsatz, ihrer Energie, ihren Leistungen, um voranzukommen; Aufstieg ist das Ziel. Die Unterschicht dagegen sieht Aufstieg nicht als ein

durch eigene Leistung zu erreichendes Ziel an, sondern als vom Schicksal bestimmt. HYMAN (1966) charakterisiert die Vorstellungen der Unterschicht wie folgt: „Die Komponenten dieses Wertsystems (der Unterschicht) beinhalten ... eine geringere Betonung der traditionellen hohen Erfolgsziele, ein erhöhtes Bewußtsein für den Mangel an Gelegenheit, Erfolg zu erreichen, und eine geringere Betonung der Ziele, die ihrerseits Mittel wären, Erfolg zu erreichen" (S. 430). Auch ROSEN (1967) kam zu dem Ergebnis, daß man in der Mittelschicht mehr „aktivistisch", „zukunft- und individualistisch" orientiert ist, d. h. durch eigene Anstrengung das im Leben erreichen zu können glaubt, was man sich erträumt, und daß diese Einstellung einhergeht mit höherem Leistungsstreben, die Unterschicht dagegen weist bei geringerem Leistungsstreben eine mehr „passivistische", „gegenwartsbetonte" und „familistische" Orientierung auf. Ausführlichere Darstellungen hierzu finden sich bei MOLLENHAUER (1971), SPITZMÜLLER (1969), OEVERMANN (1966) und BAUR (1972). Wie leicht einsichtig wird, fließen derartige Wertvorstellungen zum Aufstiegs- und Leistungsverhalten auch in Erziehungsverhalten ein und bestimmen das Bildungsinteresse mit.

GLIDEWELL (1961) faßt in dem Begriff der Erziehungshaltung (parental attitude) den gesamten Komplex von Einstellungen, Vorstellungen und tatsächlichem Erziehungsverhalten zusammen. Diese Erziehungshaltung bestimmt sich wesentlich aus der sozialen Position der Mutter. STOLZ (1967) konnte hohe Korrelationen nachweisen zwischen den Werten, die die Mutter (nicht der Vater!) für ihre Erziehungspraxis akzeptiert und der sozialen Position: je niedriger die soziale Stellung, um so eher sind die Mütter der Meinung, daß sie einerseits für die emotionale Sicherheit ihrer Kinder zu sorgen, aber andererseits die Kinder auch strikt zu kontrollieren haben, d. h. sie werden die Kinder nicht zu selbständigen Leistungen ermuntern, sondern eher zu angepaßtem und diszipliniertem Verhalten. MCCLELLAND (1966) macht einen derartigen Erziehungsstil für die mangelnde Entwicklung der kindlichen Leistungsmotivation verantwortlich und versucht damit die Tatsache mitzuerklären, daß Unterschichtkinder in der Schule weniger erfolgreich sind als Mittel- und Oberschichtkinder. Auf der anderen Seite hat sich jedoch gezeigt, daß gar nicht gesichert ist, ob hohe Leistungsmotivation entscheidend förderlich ist für die Schulleistung (siehe unseren Abschnitt über die Leistungsmotivation). Die bereits zitierte Untersuchung von AMELANG & VAGT sowie auch die einschlägigen Untersuchungen zum Zusammenhang zwischen Kreativität und Schulleistung (GRETZEL & JACKSON 1962, HASELOFF 1966, FELDHUSEN, TREFFINGER & ELIAS 1970, KEMMLER 1967) haben erbracht, daß sehr entscheidend für das Erlangen guter Schulzensuren Disziplin und Ordentlichkeit, also Unterschicht-Erziehungsziele, sind. Da aber die Kinder niederer sozialer Schichten geringeren Schulerfolg haben als Kinder höherer sozialer Schichten, können die Erziehungsziele

„Konformität, Disziplin, Ordnung" versus „Selbständigkeit, Selbstkontrolle, Unabhängigkeit, Wißbegierde" entweder doch nicht entscheidend sein für die Schulleistung, oder sie werden nicht so klar voneinander getrennt in den höheren und niederen Sozialschichten praktiziert, wie es den meisten Untersuchungsergebnissen zufolge der Fall sein müßte. Der interessante Untersuchungsansatz von KNIEL & MITZLAFF (1972) verdeutlicht, daß nicht allein das Erziehungsverhalten (in diesem Fall Belohnung für Leistungsverhalten) ausschlaggebend ist für das Schulinteresse der Kinder, sondern vielmehr, wie dieses Verhalten von den Kindern empfunden wird. Gleiche Verstärkerabsicht hat herkunftsspezifisch verschiedene Verstärkerwirkung, d. h. konkret: die positive Mutterreaktion auf Leistungsverhalten des Kindes wird von Kindern der Unterschicht (hier Obdachlosenviertel) als unangenehmer empfunden als von Kindern höherer sozialer Schichten. KNIEL & MITZLAFF erklären dieses Resultat damit, daß die Freunde der Obdachlosenkinder weniger positiv zur Schule eingestellt sind als andere Kinder. Da sie ebenfalls feststellen konnten, daß der Einfluß der Peergroups auf die Einstellung zur Schule größer ist als der der Mütter, erscheint diese Erklärung durchaus schlüssig. Interessant an diesen Ergebnissen ist weiterhin, daß sich der Einfluß der Schuleinstellung von Müttern und Peergroup unterschiedlich auf die verschiedenen Unterrichtsfächer, am meisten allerdings auf die Mitarbeitsnote, auswirkt.

Wichtiger noch als das leistungsbezogene Erziehungsverhalten scheint für das Fortkommen der Kinder in der Schullaufbahn das Bildungsinteresse der Eltern zu sein sowie ihre Möglichkeit, die Lernleistungen ihrer Kinder zu fördern. OEVERMANN (1966) beschreibt dazu die vielzitierte Analyse von KAHL (1961): „Erziehungsinteresse der Eltern ist wahrscheinlich einer der wichtigsten ursächlichen Faktoren für den positiven Zusammenhang der Ausbildungschancen mit dem sozio-ökonomischen Status. J. KAHL hat das bei Schülern der High Schools aus der oberen Unterschicht, für die nach IQ und Herkunftsstatus mit einer Wahrscheinlichkeit von .05 mit einem späteren College-Besuch gerechnet werden konnte, eindrucksvoll nachgewiesen. Die Eltern der Jungen, die das College besuchen wollten, waren typischerweise an den Werten, Symbolen und dem Lebensstil der etablierten Mittelschicht orientiert und an einer weiterführenden Ausbildung ihrer Söhne interessiert, während die Eltern der anderen Jungen sich mehr resignierend in ihr Schicksal fügten und für sich und ihre Kinder keine Chancen sahen, die Aufstiegsbarriere zu einer höheren Statusgruppe zu durchbrechen" (S. 105). Die Verhältnisse in der BRD sehen nach der großangelegten Untersuchung von BAUR (1972) entsprechend aus. Sie unterscheidet drei Gruppen: eine Gruppe von Eltern, für die der Übergang der Kinder auf eine weiterführende Schule und speziell auf ein Gymnasium unbedingtes Ziel ist, für die es keine Alternative in der Schulwahl gibt und die auch zum aller-

größten Teil dieses Ziel erreichen konnte, weil ihre Kinder die entsprechenden Leistungen erbracht haben (ob mit oder ohne Nachhilfeunterricht, wird nicht angegeben). Das sind die Eltern in den höheren Berufsstatusgruppen (leitende Angestellte, Beamte des höheren Dienstes, freie Berufe). Der zweiten Gruppe von Eltern, die ebenfalls fast durchweg ihre Einstellung erkennen ließ, daß sie an einer weiterführenden Schulbildung für ihre Kinder interessiert ist, und die ihr Ziel auch schon recht entschieden anstrebte, ist es nur zum Teil gelungen, dieses Ziel zu erreichen. Es sind Eltern aus den mittleren Berufsstatusgruppen wie mittlere Angestellte, mittlere Beamte, mittlere und kleinere selbständige Gewerbetreibende. Von den höheren Berufsstatusgruppen unterscheidet sich diese Gruppe wohl kaum in ihrem Streben nach guter Schulbildung für die Kinder, sondern eher darin, daß sie offenbar weniger in der Lage ist, ihren Kindern die entsprechenden Entfaltungsmöglichkeiten zu bieten. Ursachen dafür konnten nicht festgestellt werden. Die dritte Gruppe, die den größten Teil der Arbeiter, der untergeordneten Angestellten und der Beamten sowie Landwirte (diese fallen allerdings etwas aus der Reihe) umfaßt, weist sowohl ein Defizit an Bildungsstreben als auch an Entfaltungsbedingungen für die Kinder auf. Für diese Gruppe ist ein Syndrom von relativer Zufriedenheit bzw. Bereitschaft bezeichnend, sich mit seiner eigenen sozialen Lage abzufinden, Anspruchslosigkeit als eigenständigem Ziel, einem geringeren Interessenhintergrund, geringem Informationsniveau bzw. Informationsbedürfnis und ambivalentem oder abwehrendem Verhältnis zur Bildung. Als weiteres Hemmnis für diese Gruppe im Bildungsstreben kann man allerdings auch das geringere Selbstwertgefühl und die größere Unsicherheit bei der Schulwahl betrachten. Wenn man jedoch auf irgendwelche Weise dieses Gefälle der Bildungsinteressen und -möglichkeiten ausgleichen könnte, bliebe immer noch ein wesentliches Hindernis vor der Erreichung der Chancengleichheit im Bereich der Schulbildung bzw. der Unabhängigkeit der Schulleistung von der Sozialschicht: die Art der geforderten Schulleistung und dementsprechend die Bewertung der tatsächlich erbrachten Leistung. Zahlreiche Analysen berichten davon, daß die Schule und ihre Meßmethoden (Schulleistungstests, Zensuren, Lehrerbeurteilung) gekennzeichnet ist durch Standards wie ein der Konkurrenzsituation entsprechendes Leistungsverhalten, individualistische Wertorientierungen, personenorientiertes Sozialverhalten, Zukunftsdenken und einen „elaborierten Sprachcode", d. h. durch Standards der Mittelschicht (Bernstein 1959, Oevermann 1966, Spitzmüller 1969, Mollenhauer 1971, Hielscher 1972). Das Unterschichtkind hat sich jedoch durch Lernprozesse, z. B. Identifikationslernen, andere als für unsere mittelständisch geprägte Schule günstige Verhaltensweisen angeeignet. Es steht also beim Eintritt in die Schule vor einer ungleich größeren Aufgabe als Mittel- oder Oberschichtkinder. Es muß nicht nur den Unterrichtsstoff verstehen und auf-

nehmen, sondern es muß zudem seine Verhaltens- und Denkweisen an die der Mittelschicht anpassen, um gleiches leisten zu können bzw. korrekter: um gleich bewertet zu werden. Das Ausmaß der tatsächlich erbrachten Leistung mag durchaus dem von Mittelschichtkindern entsprechend sein, die Art der Leistung sowie die Leistungsbewertung aber ist unterschiedlich.

2.2.3. Schlußbemerkungen

Alle in unserem Beitrag zusammengestellten Untersuchungsbefunde verdeutlichen den Zusammenhang einer oder mehrerer Variablen mit Schulleistungen. Die von uns gewählte Art der Darstellung dieser Befunde, nämlich die Aneinanderreihung, soll dabei keinesfalls bedeuten — wie es vielleicht den Anschein erweckt haben mag —, daß die jeweiligen Determinanten additiv die Schulleistung beeinflussen. Es sei an dieser Stelle vielmehr noch einmal darauf hingewiesen, daß ein vielfältiges Bedingungsgeflecht besteht zwischen den verschiedenen Variablen und auch den verschiedenen Schulleistungen. Alle bisherigen Untersuchungen zu diesem Fragenbereich stützen sich bestenfalls auf Korrelationsrechnungen, so daß bis jetzt ausschließlich Aussagen über Zusammenhänge zwischen diversen Variablen und Schulleistungen gemacht werden können, nicht jedoch zu Ursache-Wirkungs-Beziehungen. Wir wissen also, daß verschiedene Intelligenzfaktoren, Motivation, verschiedene Persönlichkeitsvariablen, Lehrerverhalten und das sozio-kulturelle Milieu des Kindes im Zusammenhang mit seinen Schulleistungen stehen, wir wissen jedoch nichts über die Richtung des Zusammenhangs und schon gar nichts über die Ursachen. Das gemeinsame Auftreten mehrerer Variablen bedeutet keinen ursächlichen Zusammenhang zwischen diesen. Hier liegt noch ein weites Feld für entsprechende Forschungen vor uns.

Die Frage ist, ob eine derartige Forschung erfolgversprechend im Sinne der späteren Schulleistungsförderung ist (vorausgesetzt, ein hohes Maß an Schulleistung stellt auch weiterhin einen Wert dar). Es scheint mir müßig, die mehr oder weniger schulleistungsfördernde oder -hemmende Wirkung von Variablen festzustellen, auf die wir im Rahmen der Schule doch nur einen geringen Einfluß ausüben können (z. B. Ausmaß der Intelligenz, Milieu u. ä.). Es sei denn, wir schicken ein ungenügend intelligentes Kind zu einem Intelligenzförderungstraining, ein ängstliches zu einer Desensitivierung, ein Arbeiterkind lassen wir von einer Oberschicht-Familie adoptieren und wir selbst begeben uns in die Hände eines Verhaltenstherapeuten, der alle schulleistungshemmenden Verhaltensweisen löscht und fördernde aufbaut. Den Wert derartiger Untersuchungen für die Sozial- und Bildungspolitik — im Sinne der Erreichung von Chancengleichheit in bezug auf die Bildungsförderung — wollen wir nicht übersehen. Der wesentliche Beitrag der pädagogischen Psychologie sollte jedoch im Bereich der Curriculumentwicklung und der Lernforschung liegen. Mögliche Störvariablen wie die

in diesem Beitrag aufgeführten, verlieren m. E. für die Schulleistung an Bedeutung, wenn exakte Lernziele definiert sind und wir detaillierte Kenntnis über Lernvorgänge haben, so daß eine (oder für verschiedene Personenkreise verschiedene) optimale Strategie(n) für die Erreichung dieser Lernziele entwickelt werden kann (können). Die pädagogische Psychologie sollte von der bloßen Beschreibung (z. T. von Banalitäten) allmählich zur aktiven Einflußnahme auf die Vermittlung von Lernmaterial gelangen.

2.2.4. Literaturverzeichnis

Amelang, M. u. *Kühn, R.:* Ursachen für die bei Jungen und Mädchen unterschiedlichen Korrelationen zwischen Schulnoten und Leistungstests. Hamburg 1972, unveröffentlicht.

Amelang, M. u. *Vagt, G.:* Warum sind Schulnoten von Mädchen durch Leistungstests besser vorherzusagen als diejenigen von Jungen? Zschr. f. Entwickl.psych. u. Pädag. Psych. 1970, 2.

Allport, G. W.: Persönlichkeit. Heidelberg 1960 [2].

Amthauer, R.: Über das „Spezifische" und das „Kompensatorische" beim Zustandekommen von Leistungen, aufgezeigt an Ergebnissen einer Untersuchung zur Frage der Leistungsunterschiede der Geschlechter. Psychol. Rundschau 1963, 14.

Anastasi, A.: Länge des Schulbesuchs und Intelligenz. In: NWB, Band 16, Hrsg. *F. Weinert.*

Aschersleben, K.: Untersuchungen zur Reliabilität von Schulnoten. Schule u. Psych., 1971, 18, 5.

Bäumler, G. u. *Breitenbach, W.:* Zusammenhänge zwischen Intelligenz, Konzentration, Angst und Leistungsmotivation bei einer studentischen Stichprobe. Psychologie u. Praxis 1970, 14.

Bandura, A.: Social-learning theory of identification processes. In: *Goslin, D. A.,* Handbook of socialization theory and research. Chicago 1969.

Bandura, A.: Principles of Behavior Modifikation. New York 1969.

Bandura, A. u. *Mischel, W.:* Modification of self-imposed delay of reward thought exposure to life and symbolic models. J. Pers. soc. Psych. 1965, 2.

Bastine, R., Charlton, M., Grässner, D. u. *Schwarzel, W.:* Ein Fragebogen zur direktiven Einstellung von Lehrern (FDE). Hamburg 1969.

Baur, R.: Bildung in Zahlen; Modell, Prognosen, Alternativen; Bd. 2: Quantitative Trends im Schulwesen. Weinheim 1972.

Beech, H. R.: Changing Mans Behavior. New York 1969.

Beer, F., Kutalek, N. u. *Schnell, H.:* Der Einfluß von Intelligenz und Milieu auf die Schulleistung. Wien 1968.

Bergius, R.: Analyse der Begabung: die Bedingungen des intelligenten Verhaltens. In: *Roth, H.:* Begabung und Lernen, Stuttgart 1971.

Berlyne, D. E.: Conflict arousal and curiosity. New York 1960.

Bernstein, B.: Sozio-kulturelle Determinanten des Lernens. In: NWB, Band 16, Hrsg. *F. Weinert.*

Bernstein, L.: The effect of variation in handling upon learning and retention. J. comp. physiol. Psychol. 1957, 50.

Bloom, B.: Stability and change in human characteristics. New York 1964.

Bottenberg, E. H. u. *Wehner, E. G.:* Schulleistung in Abhängigkeit von Intelligenz und kognitiven Einzelfunktionen. Schule u. Psych. 1970.

Bracken, H. von: Probleme der „verdeckten" Begabungsreserven. Schule u. Psych. 1967, 5.

Brandstätter, H., Franke, H. u. *Rosenstil, L. V.:* Zur persönlichkeitsspezifischen Vorhersagbarkeit von Leistungsdaten. Zschr. f. exper. u. angew. Psych. 1966, 13.

Brickenkamp, R.: Der Aufmerksamkeits-Belastungs-Test (d2). Göttingen 1970 [3].

Buggle, F., Gerlicher, K. u. *Baumgärtel, F.:* Entwicklung und Analyse eines Fragebogens zur Erfassung von Neurotizismus und Extraversion bei Kindern und Jugendlichen. Diagnostica 1968, 19.

Burt, C.: The Differentiation of Mental Ability. Brit. J. of Educ. Psych. 1954, 24.

Carmical, L.: Characteristics of achievers and underachievers of a large Senior High School. In: *Simons, H.:* Kognitive Bedingungen über- und untererwartungsgemäßer Schulleistungen. Zschr. f. Entwicklungspsych. u. Päd. Psych. 1969, 1,1.

Carter, R. S.: How invalid are marks assigned by teachers. In: J. Educ. Psych. 19, Vol. 43, 4.

Cattell, R. B. u. *Scheier, I. H.:* Handbook for the IPAT Anxiety Scale Questionnaire. Illinois 1963.

Clostermann, G.: Studien zur Testwissenschaft; der Mann-Zeichen-Test in formtypischer Auswertung. Münster 1959.

Cogan, M. L.: The behavoir of teachers and the productive behavior of their pupils. J. exper. Educ. 1958, 27.

Coleman, I. C. u. *Rasof, B.:* Intellectual factors in learning disorder. In: *Simons, H.:* loc. cit.

Cowen, E. L., Zax, M., Klein, R., Izzo, D. u. *Trost, M. A.:* The relation of anxiety in school children to school record, achievement and behavioral measures. Child Devel. 1965, 36.

Cox, F. N.: Educational, streaming and test anxiety. Child Devel. 1962, 33.

Dann, H. D. u. *Müller-Fohrbrodt, G.:* Einige Ursachen und Folgen der Benotungspraxis an Gymnasien. Zschr. f. Entwicklungspsych. u. Pädag. Psych. 1972, 4, 4.

Engels, F.: Herrn Eugen Dührings Umwälzung der Wissenschaft. Berlin 1952, in: *Rosenfeld, G.,* Theorie und Praxis der Lernmotivation, Berlin 1966.

Ewert, O.: Untersuchungen zum Zusammenhang von Phantasie und Intelligenz bei Jugendlichen. Ber. üb. 22. Kongreß d. Dt. Ges. f. Psych., Göttingen 1960.

Ewert, O.: Experimentelle Untersuchungen über Zusammenhang und Entwicklung von gerichteter Phantasie und Intelligenz bei Jugendlichen. Mainzer Habilitationsschrift, 1962.

Eysenck, Sybil, B.: Manual of the Junior Eysenck Personality Inventory. London 1965.

Eysenck, H. J. u. *Rachman, S.:* Neurosen, Ursachen und Heilmethoden. Berlin 1970.

Feldhusen, J. F. u. *Klausmeier, H. J.:* Anxiety, intelligence and achievement in children of low, average and high intelligence. Child Devel. 1962, 33.

Feldhusen, J. F., Treffinger, D. J. u. *Elias, R. M.:* Prediction of academic achievement with divergent and convergent thinking and personality variables. Psych. in the Schools 1970, 7.

Festinger, L.: A theory of cognitive dissonance. In: *Heckhausen, H.,* 1971, loc. cit.

Fippinger, F.: Intelligenz und Schulleistung. Eine experimentelle Untersuchung bei 9- bis 10jährigen Schülern. Erzieh. u. Psych., Beihefte d. Zschr. Schule u. Psych. 41, München, Basel 1966.

Flechsing, K. K. u. a.: Die Steuerung und Steigerung der Lernleistung durch die Schule. In: *Roth, M.*, Begabung und Lernen, Stuttgart 1971.
Flitner, A.: Das Schulzeugnis im Lichte neuer Untersuchungen. Zschr. f. Pädag. 1966, 6.
Foppa, K.: Lernen, Gedächtnis, Verhalten — Ergebnisse aus Problemen der Lernpsychologie. Köln 1966 ².
Frankel, E. A.: A comperative study of achieving and underachieving high school boys of high intellect ability. In: *Simons, H.*, loc. cit.
French, J. R. P., Israel, J. u. *As, D.:* An experiment on participation in a norwegean factory. Human Relations 1960, 13.
Fröhlich, W. D. u. *Drever, J.:* Wörterbuch zur Psychologie. München 1968.
Gärtner-Harnach, V.: Angst und Leistung. Weinheim 1972.
Gaude, P. u. *Teschner,* W.: Objektivierte Leistungsmessung in der Schule. Frankfurt/M., 1970.
Getzel, J. W. u. *Jackson, W.:* Creativity and Intelligence. Explorations with Gifted Students. London, New York 1962.
Glidewell, J. C. (Ed.): Parental attitudes an child behavior. Springfield, Ill. 1961.
Guilford, J. P.: Fundamental Statistics in Psychology and Education. New York 1965.
Hahn, E.: Divergenzen von Intelligenz und Schulleistung. Empirische Untersuchung über einige nicht-intellektuelle Bedingungen schulischer Leistungen. (Manuskript).
Heckhausen, H.: Leistungsmotivation. In: *Thomae, H.*, (Hrsg.), Handbuch der Psychologie, Band 2, 2, Göttingen 1965.
Heckhausen, H.: Förderung der Lernmotivierung und der intellektuellen Tüchtigkeit. In: *Roth, H.*, Begabung und Lernen, Stuttgart 1971.
Heller, K.: Aktivierung der Bildungsreserven. Bern, Stuttgart 1970.
Heller, K. u. *Schirmer, B.:* Wortschatztests für Sehbehinderte, WST (Sb). Weinheim 1973.
Herrmann, Th.: Erziehungsstile. Göttingen 1970 ².
Hess, R. D. u. *Shipman, V.:* Early experience and the socialization of cognitive modes in children. Child Develpm. 1965, 36.
Hielscher, H. (Hrsg.): Die Schule als Ort sozialer Selektion. Heidelberg 1972.
Hill, K. T. u. *Sarason, S. B.:* The relations of test anxiety and defensiveness to test and school performance over the elementary school years. Monogr. Society for Research in Child. Devel. 1967, 38.
Hitpass, J.: Vergleichende Untersuchungen über den Voraussagewert von Aufnahmeprüfungen und Testprüfungen zur Erfassung der Eignung für die weiterführende Schule. Schule u. Psych. 1961, 8.
Hoppe, F.: Erfolg und Mißerfolg. In: *Wasna, M.*, loc. cit.
Horn, W.: Das Begabungs-Test-System, Handanweisungen. Göttingen 1956.
Hyman, H. H.: The value system of different classes. In: Class, Status and power, Ed. *Bendix, R.* u. *Lipset, S. M.*, New York 1966.
Ingenkamp, K.-H.: Zur Problematik der Zensurengebung. Die dt. Schule, 1970, 62.
Janssen, J. P.: Kritische Bemerkungen zu Validitätsstudien mit den Prädiktoren „Schulnoten" und „Intelligenztests". Diagnostica 1972, 18.
Kagan, J. u. a.: Personality and IQ change. In: *Heckhausen, H.*, loc. cit.
Kagan, J. u. a.: Information processing in the child: significance of analytic and reflective attitudes. Psych. Monogr. 1964, 78.
Kagan, J. u. *Moss, H. A.:* Stability and validity of achievement fantasy. In: *Heckhausen, H.*, 1971, loc. cit.

Kahl, J. A.: „Common Man" Boys. In: *Halsey, A. H., Floud, J.* u. *Anderson, C. A.* (Ed.), Education, Economy and Society, New York 1961.
Kemmler, L. u. *Langheinrich, D.:* Eine deutsche Fassung der „Primary Mental Abilities" für 8- bis 10jährige Kinder. In: *Merz,* F. (Hrsg.), Ber. 25. Kongr. Dt. Ges. Psych. Göttingen 1967.
Kemmler, L.: Erfolg und Versagen in der Grundschule. Göttingen 1967.
Klausmeier, H. J. u. *Goodwin, W.:* Learning and human abilities. In: *Thiersch, H.,* loc. cit.
Kleinschmidt, G.: Ergebnisse im Bereich der kompensatorischen Erziehung. Schule und Psych. 1971, 18. 7.
Kniel, A. u. *Mitzlaff, S.:* Eltern-, Lehrer- und Peereinflüsse auf Schulleistungen. Eine Untersuchung bei Hauptschülern im Obdachlosengebiet Mannheim. Zschr. f. Sozialpsych. 1972, 3, 4.
Kornmann, R.: Minimalisieren Schulreifetests die Zahl der Fehlentscheidungen? — Kommentar zum Bericht von *Mandl* u. *Krapp,* loc. cit. Zschr. f. Entwicklungspsych. u. Pädag. Psych. 1972, 4, 4.
Koschat, H.: Analyse von extremen Diskrepanzen zwischen Schulleistung und Intelligenzmessung. Psychol. Rundsch. 1966, 17, 2.
Kruse, L. u. *Rogge, K.-E.:* Motivation. In: Steckbrief der Psychologie, Hrsg. Heidelberger Autorengruppe, Heidelberg 1971.
Latscha, F.: Der Einfluß des Primar-Lehrers. In: *Hess, F., Latscha, F.* u. *Schneider, W.,* Die Ungleichheit der Bildungschancen, Freiburg/Olten 1966.
Liebel, M. u. *Wellendorf, F.:* Schülerselbstbefreiung — Voraussetzungen und Chancen der Schülerrebellion. Frankfurt 1969.
Lienert, G. A.: Die Faktorenstruktur der Intelligenz als Funktion des Intelligenzniveaus. Ber. üb. d. 22. Kongr. d. Dtsch. Ges. f. Psych. Göttingen 1960.
Lienert, G. A.: Testaufbau und Testanalyse. Weinheim 1969 [3].
Lippit, R. u. *White, R. K.:* An experimental study of leadership and group life. In: *Swanson, Newcomb* u. *Hartley,* New York 1952.
Ljung, B.-O.: Skolmotivation, ängstigkeit och prestation. In: *Johanesson, J.,* Effects of praise and blame. Results of teachers incentive upon achievement and attitudes of school children. Stockholm studies in Educat. Psych., Stockholm 1967.
Löschenkohl, E.: Gibt es einen allgemein faßbaren Zusammenhang zwischen Schulleistung und Intelligenz? Psychologie in Erziehung und Unterricht, 1973, 3.
Mandl, H. u. *Krapp, A.:* Ist die Zahl selektiver Fehlentscheidungen in der pädagogischen Diagnostik von Bedeutung? — Gedanken zum Diskussionsbeitrag von *Kornmann, R.,* loc. cit. Zschr. f. Entwicklungspsych. u. Pädag. Psych. 1972, 4, 4.
Mandl, H. u. *Krapp, A.:* Zum Problem der Punktwertgrenzen bei der Interpretation von Schulreifetestergebnissen. Zschr. f. Entwicklungspsych. u. Pädag. Psych. 1972, 2.
Mandler, G. u. *Sarason, S. B.:* A study of anxiety and learning. J. of abnormal soc. Psych. 1952, 47.
Marshall, J. C.: Composition errors and essay examination grades re-examined. American Educ. Research Journal 1967, 4, 1.
McClelland, D. C.: Die Definition eines spezifischen Motivs. In: *Thomae, H.* (Hrsg.), Die Motivation menschlichen Handelns. Köln 1965.
McClelland, D. C.: Longitudinal trends in the relation of thought to action. J. consult. Psych. 1966 b, 30.

McClelland, D. C., Atkinson, J. W., Clark, R. A. and *Lowell, E. L.:* The achievement motive. New York 1953.
McKeachie, W. J.: Motivation, Lehrmethoden und Lernen in Hochschulen. In: *Weinert, F.*, NWB, Bd. 16.
Meili, R.: Lehrbuch der psychologischen Diagnostik. Stuttgart 1961 [4].
Meili, R.: Der Analytische Intelligenz-Test. Stuttgart 1966.
Meyer, W. V., Heckhausen, H., Kemmler, L.: Validierungskorrelate der inhaltsanalytisch erfaßten Leistungsmotivation guter und schwacher Schüler des 3. Schuljahrs. Psychol. Forsch. 1964/65, 28.
Mollenhauer, K.: Sozialisation und Schulerfolg. In: *Roth, H.* (Hrsg.) 1971 [6], loc. cit.
Mühle, G.: Definitions- und Methodenprobleme der Begabungsforschung. In: *Roth, H.* (Hrsg.) 1971 [6], loc. cit.
Nickel, H. u. *Schlüter, P.:* Angstwerte bei Hauptschülern und ihr Zusammenhang mit Leistungs- sowie Verhaltensmerkmalen, Lehrerurteil und Unterrichtsziel. Zschr. f. Entwicklungspsych. u. Pädag. Psych. 1970, 2, 2.
Nickel, H., Schlüter, P. u. *Fenner, H.-J.:* Angstwerte, Intelligenztest- und Schulleistungen sowie der Einfluß der Lehrerpersönlichkeit bei Schülern versch. Schularten. Psychol. in Erzieh. u. Unterr., 1973, 1.
Notz, I.: Übereinstimmung zwischen Lehrerurteil und einer Begabungsmessung von Schülern in der 4. Klasse einer Berliner Grundschule. Schule u. Psych. 1965, 12, 10.
Oevermann, U.: Schichtenspezifische Formen des Sprachverhaltens und ihr Einfluß auf die kognitiven Prozesse. In: *Roth* (Hrsg.) 1971 [6], loc. cit.
Oevermann, U.: Soziale Schichtung und Begabung. Zschr. f. Pädag., Beiheft 6, 1966.
Ohlhoff, G.: Begabung und Schulerfolg. Schule u. Psych. 1963, 10, 1.
Orlik, P.: Ein Beitrag zu den Problemen der Metrik und der diagnostischen Valenz schulischer Leistungsbeurteilung. Zschr. f. exp. u. angew. Psych. 1961, 8.
Peel, E. A.: Present Trends in Selection for Secondary Education in England. In: *Schultze, W.*, Über den Voraussagewert der Auslese-Kriterien für den Schulerfolg am Gymnasium. Frankfurt 1964.
Pidgeon, D. A.: Reserarch into Secondary School Selection in England. In: *Schultze, W.*, Über den Voraussagewert der Auslese-Kriterien für den Schul- am Gymnasium, Frankfurt 1964.
Porter, R. B. u. *Cattell, R. B.:* Childrens Personality Questionnaire (CPQ). Champaign/Illinois, Institute for Personality and Ability Testing, 1963.
Priester, H. J.: Die Standardisierung des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK). Bern 1961.
Rank, T.: Schulleistung und Persönlichkeit, München 1962.
Rösler, H.-D.: Leistungshemmende Faktoren in der Umwelt des Kindes, Leipzig 1967 [2].
Roloff, E.-A.: Intelligenz und Schulleistung. Schule u. Psych. 1957, 4, 10.
Rosen, B. C.: The Achievement Syndrom. In: *Rolf, H.-G.*, Sozialisation und Auslese durch die Schule. Heidelberg 1967.
Rosen, B. C. u. *d'Andrade, R.:* The psychological origins of achievement motivation. Sociometry 1959, 22.
Rosenfeld, G.: Theorie und Praxis der Lernmotivation. Berlin 1966.
Rosenthal, R.: Experimenter expectancy and the reasuring nature of the null hypothesis decision procedure. Psychol. Bulletin Monogr. Supplement 70, 1968.
Roth, H.: Begabung und Schule. Schweiz. Lehrerzeitung, 1964, 109, 5.
Roth, H. (Hrsg.): Begabung und Lernen. Stuttgart 1971 [6].

Ryans, D. G.: Charakteristics of teachers. Washington 1960.
Sander, A.: Begabung, Intelligenz, Leistung. Schule u. Psych. 1967, 14, 7.
Sarason, S. B.: The measurement of anxiety in children: some questions and problems. In: *Spiegelberger, Ch. D.*, Anxiety and Behavior, New York 1966.
Sarason, S. B., Hill, K. T. u. *Zimbardo, P. G.:* Eine Längsschnittuntersuchung über den Zusammenhang zwischen Prüfungsangst und dem Verhalten bei Intelligenz- und Schulleistungstests. In: *Weinert, F.*, NWB, Bd. 16.
Schell, H.: Angst und Schulleistung. Göttingen 1972.
Schmitz, G. F.: Grundschulleistung, Intelligenz und Übertrittsauslese. Erziehung u. Psych., Beihefte d. Zschr. Schule u. Psych. 29, München 1964.
Schröder, H.: Zur Problematik der Fähigkeitsdiagnose in der Schülerbeurteilung. Zschr. f. exper. u. angew. Psych. 1971.
Schwarz, E.: Schulreife, Intelligenz u. Schulleistung im 1. Schuljahr. Schule u. Psych. 1967, 14, 8.
Seitz, W.: Über den Zusammenhang von Persönlichkeitseigenarten, Schulnoten und HAWIK-Leistungen bei Volksschülern. Psychol. Beiträge 1970 a.
Seitz, W.: Über den Zusammenhang von Persönlichkeitseigenarten mit diversen Aufmerksamkeitsleistungen und der Farb-Wort-Interferenz bei Volksschülern, Arch. f. Psych. 1970 b, 122.
Seitz, W.: Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztest-Leistungen bei Volksschülern. Zschr. f. exper. u. angew. Psych. 1971, 18.
Seitz, W.: Über den Zusammenhang der Brengelmann-Skalen Nr, Do und IA mit diversen intellektuell-kognitiven Leistungen. Psychologie und Praxis 1971, 15, 1.
Seitz, W. und *Löser, G.:* Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztest-Leistungen bei Gymnasialschülern. Zschr. f. exper. u. angew. Psych. 1969, 16.
Seitz, W. u. *Metzelder, L.:* Empirischer Beitrag zum Zusammenhang zwischen HAWIK-Leistungen mit Persönlichkeitsveriablen. Psych. Beiträge 1970, 12.
Sherif, M. B. u. a.: Status in Experimentally Produced Groups. Am. J. Sociol. 1955, 60.
Simons, H.: Kognitive Bedingungen über- und untererwartungsgemäßer Schulleistungen. Zschr. f. Entwicklungspsych. u. Päd. Psych. 1969, 1.
Skolnik, A.: Motivational imagery and behavior over twenty years. In: *Heckhausen, H.*, 1971, loc. cit.
Sontag, L. W., Barker, C. T. u. *Nelson, V. L.:* Mental growth and personality development: a longitudinal study. In: *Heckhausen, H.*, 1971, loc. cit.
Starch, D. u. *Elliot, E. C.:* Reliability of grading work in history. School Review 1913, 21.
Steinkamp, G.: Die Rolle des Volksschullehrers im schulischen Selektionsprozeß. In: *Hielscher, H.*, loc. cit.
Stelzl, W.: Die Korrelation zwischen Intelligenz, Gedächtnis und Schulleistungen zu Beginn der Reifezeit. Diss. d. Phil. Fak. d. Univ. Graz 1949.
Stolz, L. M.: Influence on parent behavior. Stanford University Press. 1967.
Suermann, V.: Zur Entwicklung von Schulleistung und Intelligenz. Schule u. Psych. 1971.
Tausch, R.: Erziehungspsychologie. Göttingen 1970 [5].
Taylor, H. C. u. *Russell, J. T.:* The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. J. appl. Psychol. 1939, 23.

Tent, L.: Die Auslese von Schülern für weiterführende Schulen. Göttingen, 1969.
Tewes, U.: Emotionalität u. Schulleistung: Einige Angaben zur Validität der HANES (KJ). Diagnostica 1973, 1, 19.
Thiersch, H.: Lehrerverhalten und kognitive Leistungen. In: *Roth, H.* (Hrsg.), 1971 [6], loc. cit.
Todt, E.: Untersuchungen zur Vorhersage von Schulnoten. In: NWB, Band 16, *F. Weinert* (Hrsg.).
Todt, E.: Differentieller Kenntnistest (DKT). Deutsche Ges. f. Personalwesen.
Undeutsch, U.: Zum Problem der begabungsgerechten Auslese beim Eintritt in die höhere Schule und während der Schulzeit. In: *Roth, H.* (Hrsg.) 1971 [6], loc. cit.
Vontobel, I.: Leistungsbedürfnis und soziale Umwelt. Bern 1970.
Wasna, M.: Motivation. Intelligenz und Lernerfolg. München 1972.
Weber, A.: Intelligenz und Schulleistung. Schule u. Psych. 1966, 13, 12.
Weinert, F.: Schülerpersönlichkeit und Schulleistung. In: *Ingenkamp, K.-H.* (Hrsg.), Schulkonflikt und Schülerhilfe. Weinheim 1965.
Weinert, F. (Hrsg.): Pädagogische Psychologie. (= NWB, Bd. 16). Köln/Berlin 1972 [7].
Weiss, R.: Das Verhältnis von Schulleistung und Intelligenz. In: *Lückert, H.-R.* (Hrsg.), Begabungsforschung und Bildungsförderung als Gegenwartsaufgabe, München, Basel 1969.
Weiss, R.: Über den Zusammenhang zwischen Schulleistung und Intelligenz, Schule u. Psych. 1964, 11.
Weiss, R.: Zensur und Zeugnis. Linz 1965.
Wendeler, J.: Standardarbeiten. Weinheim 1972.
Wendeler, J.: Intelligenztests in Schulen. Weinheim 1972.
Wenzl, A.: Theorie der Begabung. Leipzig 1934.
Wewetzer, K. H.: Intelligenz und Intelligenzmessung. Darmstadt 1970.
Wewetzer, K. H.: Zur Differenzierung der Leistungsstrukturen bei verschiedenen Intelligenzgraden. Ber. 21. Kongr. Dtsch. Ges. f. Psychol. Göttingen 1958.
Wieczerkowski, W., Nickel, H. u. *Rosenberg, L.*: Einige Bedingungen der unterschiedlichen Bewertung von Schüleraufsätzen. Psychol. Rundschau 1968, 4.
Winterbottom, M.: The relation of need for achievement to learning experiences in independency and mastery. In: *Atkinson, J. W.* (Ed.), Motives in Fantasy, Action and Society. Princeton, N. J. 1964 [2].
Wolpe, J.: The Practice of Behavior Therapy. New York 1970.
Yates, A. u. *Pidgeon, D. A.*: Admission to Grammar School. In: *Schultze, W.*, Über den Voraussagewert der Auslesekriterien für den Schulerfolg am Gymnasium. Frankfurt 1964.
Zielinski, W.: Beziehungen zwischen Schulleistung, Intelligenz und Familienkonstellation. Schule u. Psych. 1966, 16, 10.
Zielinski, W.: Beziehungen zwischen Ängstlichkeit, schulischer Aktivität, Intelligenz und Schulleistung bei 9- bis 11jährigen Volksschülern. Schule u. Psych. 1967, 14, 9.
Zielinski, W.: Macht und Ohnmacht der Zensuren. Pädag. Rundsch. 1961, 15.
Zigler, E. u. *Butterfield, E. C.*: Motivational aspects of change in Intelligenztest performance of culturally deprived nursery school children. Child. Devel. 1968, 39.
Ziler, H.: Der Mann-Zeichen-Test in detail-statistischer Auswertung. Münster 1959.
Zöchbauer, W.: Die Aufnahmeprüfung in der Mittelschule. Salzburg 1962.

3. Testtheoretische Ansätze der Schulleistungsmessung

Einleitender Kommentar

Die klassische Testtheorie (Meßfehlertheorie) bildet nach wie vor die Grundlage für objektive, zuverlässige und gültige Schülerbeurteilungen. Diese Feststellung gilt trotz einer Reihe von Problemen, die sich im Hinblick auf die Leistungsbewertung (z. B. durch sog. standardisierte Schulleistungstests), erst recht aber unter dem Aspekt der Instruktions- oder Unterrichtshilfen (z. B. durch sog. kriterienbezogene Tests) stellen. Die Kenntnis testtheoretischer Axiome ist aber auch für die Zensurengebung von Belang, wenigstens solange man sich um eine objektivere und damit gerechtere Urteilspraxis (z. B. mit Hilfe von informellen Tests) bemüht.

Nach der Explikation des Begriffspaares ‚Messen' und ‚Testen' werden in dem ersten Beitrag von LANGFELDT die Grundannahmen der Meßfehlertheorie in knapper und doch sehr verständlicher Form dargelegt. Im Mittelpunkt seiner Ausführungen steht der Bezug zur Leistungsmessung, d. h. die Anwendung der klassischen Testtheorie auf die *standardisierten* Schulleistungstests. Neben der Erörterung der sog. Testgütekriterien und einschlägiger Konstruktionsprinzipien verdient hier die abschließende Kritik an der Meßfehlertheorie im Hinblick auf die Schulleistungsbeurteilung besondere Beachtung.

Der folgende Beitrag von BÜSCHER beschäftigt sich mit testtheoretischen Problemen unter dem Aspekt *kriterienbezogener* Leistungsmessung. Damit ist — in Abhebung von den ‚klassischen' (standardisierten) Schulleistungstests — jene Art von Schulleistungstests angesprochen, die noch kaum ein Jahrzehnt alt sind und gewöhnlich unter Begriffe wie ‚lernzielorientierte', ‚lehrzielorientierte', ‚kriteriumsbezogene' u. ä. Tests zusammengefaßt werden.

Der Kriterienbegriff unterscheidet sich hier nicht nur von dem nämlichen Begriff der klassischen Testtheorie (s. Abschn. 3.1.4.2. oben), er ist auch im Kontext moderner testtheoretischer Ansätze keineswegs so eindeutig, wie es auf den ersten Blick erscheinen mag. Der — unumgänglichen — Begriffsdiskussion folgt eine ausführliche Erörterung einschlägiger Fragen der Konstruktion sog. Lernsteuerungstests (siehe unten). Dabei werden spezielle Probleme der Aufgabenanalyse herausgestellt sowie die wichtigsten Verfahrensansätze zur Item- bzw. Testanalyse kriterienbezogener Leistungsmessung besprochen. BÜSCHER gelangt schließlich zu der Ansicht, daß die Diskussion um die kriterienbezogenen Leistungstests erst am Anfang stehe. „Wenn sich das Konzept der kriterienbezogenen Messung in der pädagogischen Praxis durchsetzen soll, dann müssen bestimmte Meßaspekte (Kon-

zepte der Test- und Aufgabenanalyse) neu durchdacht werden." Bei den vorliegenden Konzepten handelt es sich seiner Meinung nach keinesfalls schon um abgesicherte (neue) theoretische Modelle, allenfalls um „erste Ansätze" hierzu.

Der dritte Beitrag in diesem Hauptkapitel schlägt quasi eine Brücke zwischen den diskutierten theoretischen Grundsatzfragen einerseits und praktischen Notwendigkeiten andererseits. Ausgehend von den unterschiedlichen testtheoretischen Ansätzen sowie der weithin fehlenden Stringenz konventioneller Termini entwickelte ROSEMANN seine „Gedanken zu einer pädagogisch begründeten Klassifikation der Schultests". Quintessenz dieser Bemühung ist die Klassifizierung der Tests unter dem Gesichtspunkt der Funktion von Leistungs*feststellung* (Lernsteuerungstests) und Leistungs*bewertung* (Lernkontrolltests).

Der Klasse der *Lernsteuerungstests* wären demnach folgende Verfahren herkömmlicher Benennung zuzuordnen: die (informellen) kriterienbezogenen Tests ohne Benotung (Zensurierung) der Schülerleistung sowie — seltener — normbezogene Tests, z. B. für kleinere Lerneinheiten. Der Klasse der *Lernkontrolltests* wären zuzurechnen: alle formellen oder sog. standardisierten Schulleistungstests, die informellen normbezogenen Tests sowie die (informellen) kriteriumsbezogenen Tests mit einhergehender Benotung der Schülerleistung.

3.1. Die klassische Testtheorie als Grundlage standardisierter Schulleistungstests

Hans-Peter Langfeldt

3.1.0. Vorbemerkung

Wenn dem Lehrer empfohlen wird, in seiner täglichen Berufspraxis standardisierte Schulleistungstests anzuwenden oder informelle Tests selbst zu konstruieren, so muß ihm der Erwerb eines Mindestmaßes an Kenntnissen über die theoretischen Grundlagen der verwendeten Tests zugemutet werden. Dieses Mindestmaß an notwendigen Kenntnissen versucht der folgende Beitrag zu vermitteln.

Grundlage der publizierten Schultests und besonders der informellen Tests ist die „klassische Testtheorie". Es wird daher von dieser die Rede sein. „Moderne" Testtheorien werden nur im Rahmen der abschließenden Diskussion angesprochen (Abschn. 3.1.7.1.). Die klassische Testtheorie ist unabhängig vom Inhalt des einzelnen Tests. Sie wird deshalb zunächst allgemein dargestellt werden, wobei allerdings die Aspekte, die für Schulleistungstests von besonderer Bedeutung sind hervorgehoben werden. Erst an späterer Stelle (Abschn. 3.1.6.) werden die theoretisch erarbeiteten Begriffe anhand eines Beispiels noch einmal verdeutlicht. Demjenigen Leser, der noch nie mit einem Test konfrontiert wurde, wird daher empfohlen, den Abschnitt 3.1.6.3. parallel zu den Abschnitten 3.1.3. bis 3.1.5. zu lesen, damit die relativ abstrakten Darstellungen mit konkreten Vorstellungen verbunden werden können.

Eine „Einführung in die klassische Testtheorie" ist im allgemeinen eher Gegenstand von Lehrbüchern mit mehreren hundert Seiten (z. B. ANASTASI 1963; CRONBACH 1960; DRENTH 1969; GULLIKSEN 1950; HORST 1971; LIENERT 1969; MAGNUSSON 1969). Dieser Beitrag will und kann nicht das Studium eines dieser Lehrbücher ersetzen, falls der Leser sich genauer einarbeiten möchte. Es kann hier nur darum gehen, einen relativ vereinfachten und groben Überblick über die Problematik der klassischen Testtheorie und der entsprechenden Testkonstruktion zu geben. Auf mathematische Ableitungen oder Beweise bestimmter statistischer Formeln wird bewußt verzichtet. Beim Leser wird allerdings mindestens die Kenntnis folgender Begriffe vorausgesetzt: Arithmetisches Mittel, Streuung, Varianz, Korrelationskoeffizient und Normalverteilung. Einschlägige Statistiklehrbücher informieren darüber (z. B. CLAUSS & EBNER 1970; HASELOFF & HOFFMANN 1970; HELLER et al. 1974; MITTENECKER 1970; WALKER 1970).

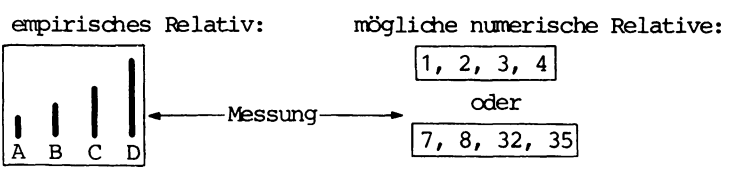
3.1.1. Messen und Testen

3.1.1.1. Definition von Messen

Testen heißt nichts anderes, als ein Individuum oder eine Gruppe von Individuen hinsichtlich eines oder mehrerer definierter Merkmale zu messen (z. B. Intelligenz, Schulleistung, Ängstlichkeit usw.). Wenn also irgendeiner Versuchsperson (Vp) nach einem bestimmten Test ein Testpunktwert zugewiesen wird, so liegt ein Meßwert vor. Der Test war in diesem Falle das Meßinstrument. Das Verständnis der klassischen Testtheorie wird erleichtert, wenn zunächst der Begriff des Messens oder der Messung erläutert wird.

„Das Wesen der Messung ist die Abbildung einer Menge vorgegebener Objekte (oder Meßgegenstände) und der zwischen ihnen empirisch feststellbaren Relationen auf eine Menge von Zahlen und den Relationen zwischen ihnen." (FISCHER 1968 b, S. 54). Man geht davon aus, daß ein beobachtbares *empirisches Relativ* besteht, das durch eine Menge von Objekten und den beobachtbaren Beziehungen zwischen diesen definiert ist. Ein empirisches Relativ ist immer nur auf einen Ausschnitt der gesamten Wirklichkeit beschränkt. Ausschnitt bedeutet in diesem Zusammenhang, daß nicht alle möglichen Objekte und auch nicht alle möglichen Beziehungen zwischen ihnen beobachtet werden. Vielmehr wird nur eine Stichprobe der Objekte und ihrer Beziehungen beobachtet. Das empirische Relativ ist unabhängig von jeglicher Messung: Es besteht schon vor der Messung und wird durch sie nicht verändert — oder sollte durch sie nicht verändert werden.

Schematisches Beispiel zur Messung:



Anzahl der Objekte: 4
Art der Beziehungen
größer-kleiner

Anzahl der Zahlen: 4
Art der Beziehungen:
größer-kleiner

Es ergibt sich hier eine Übereinstimmung des empirischen und des (der) numerischen Relativs. Die Anzahl der Objekte und die Anzahl der Zahlen ist identisch. Die kleinste Zahl entspricht dem kleinsten Objekt, die größte Zahl dem größten Objekt. Die Beziehungen zwischen den Zahlen ist in beiden numerischen Relativen gleich der Beziehung zwischen den Objekten.

Wenn nun dieses empirische Relativ durch ein sogenanntes *numerisches Relativ* abgebildet werden kann, dann liegt eine Messung vor. Ein numerisches Relativ ist durch eine Menge reeller Zahlen und den Beziehungen zwischen diesen definiert. Wenn die Abbildung des empirischen Relativs durch das numerische Relativ angemessen ist, dann kann man annehmen, daß die Beziehungen zwischen den Zahlen eine Repräsentation der Beziehungen zwischen den Objekten darstellen.

Wie das Beispiel zeigt, ist die Auswahl der Zahlenwerte des numerischen Relativ willkürlich. Sie wird durch das empirische Relativ nicht vorgegeben. Entscheidend ist nur, daß die Beziehungen zwischen den Zahlen den Beziehungen zwischen den Objekten entsprechen. Im Beispiel handelt es sich um Größer-Kleiner-Beziehungen. Diese werden durch die aufgeführten numerischen Relative repräsentiert. Es ist zu beachten, daß zunächst nichts darüber ausgesagt ist, wie groß die Größenunterschiede zwischen den Objekten sind.

Die Messung in diesem Beispiel war *isomorph,* d. h. jedem Objekt A, B, C, D entsprach genau eine Zahl. Umgekehrt entsprach jede Zahl 1, 2, 3, 4 genau einem Objekt. Psychologische Messungen sind im allgemeinen *homomorph,* d. h. zwar entspricht jedem Objekt eindeutig eine Zahl, aber nicht jeder Zahl entspricht eindeutig nur einem Objekt. In einem Schultest erhält jeder Schüler einen Testwert, der für ihn charakteristisch ist. Andererseits erhalten meistens mehrere Schüler denselben Testwert, so daß man nur aus Kenntnis des Testwertes nicht feststellen kann, von welchem Schüler der Test stammt.

Es wird deutlich, daß Messen immer ein Feststellen von Unterschieden bedeutet, die durch den Vergleich der Objekte miteinander ermittelt werden. Nicht jede Zuordnung von Skalen zu Objekten kann als Messung interpretiert werden. Man kann nicht von Messung sprechen, wenn die Zuordnung zufällig geschieht oder so fehlerhaft ist, daß die Beziehungen zwischen den Zahlen nichts über die Beziehungen zwischen den Objekten aussagen.

3.1.1.2. Skalen für Meßwerte

Meßwerte liegen auf einer Skala. In der Physik beispielsweise auf einer Meter-Skala oder einer Volt-Skala. Hinsichtlich der Aussagekraft eines Meßwertes ist jedoch die Charakteristik der unterlegten Skala bedeutsam. Skalen unterscheiden sich durch ihr Niveau. Das Niveau wird dabei durch die Art der Informationen definiert, die der Skala entnommen werden können. Eine Skala mit höherem Niveau erlaubt weitgehendere Interpretationen als eine Skala mit niedrigem Niveau. Folgende Skalen können in einer hierarchisch gegliederten Ordnung dargestellt werden.

Die Nominalskala:

Messungen auf dem Niveau einer Nominalskala erbringen nur Informationen über Identität bzw. Nicht-Identität der verglichenen (gemessenen) Objekte. Identität wird dabei durch die gemeinsame Zugehörigkeit der Objekte zur selben Klasse definiert. Bäcker und Metzger gehören etwa zur Klasse der „Handwerker", Sänger und Maler zur Klasse der „Künstler". Messungen auf dem nominalen Niveau liegen immer dann vor, wenn durch sie die Objekte in Klassen eingeteilt werden können (etwa: männliche Personen, weibliche Personen, Gruppe A oder Gruppe B, Erwachsene und Kinder). Auf einer Nominalskala werden also qualitative Unterschiede festgestellt.

Die Ordinalskala (oder Rangskala):

Die Rangskala nimmt auf dem untersten Niveau eine quantitative Unterscheidung vor. Die Meßobjekte werden hinsichtlich von Größer-Kleiner-Beziehungen geordnet. Die Schüler einer Schulklasse könnten entsprechend ihrer Körpergröße geordnet werden. Dem Größten würde der Zahlenwert 1 zugewiesen, dem Zweitgrößten der Zahlenwert 2, usw., oder der beste Rechner der Klasse könnte den Wert 1 erhalten, der Zweitbeste 2, usw. Durch die Skala erhält man nur Informationen über die Rangfolge, nicht aber über den Abstand zwischen den Meßobjekten. Man weiß also nicht, um wieviel 1 größer (besser) als 2 ist, oder ob der Unterschied zwischen 1 und 2 gleich groß ist wie der zwischen 2 und 3. Da die Abstände zwischen den einzelnen Zahlenwerten nicht bekannt sind, können keine direkten arithmetischen Operationen vorgenommen werden. Ein typisches Beispiel für eine Rangskala sind die Schulnoten, wobei 1 die beste und 6 die schlechteste Schulleistung bedeutet. Im strengen statistischen Sinne sind die Berechnungen von Notendurchschnitten nicht statthaft.

Die Intervallskala:

Wenn durch die Skala zusätzlich noch die Abstände zwischen den Skalenpunkten definiert werden, so erhält man eine Intervallskala: Die Intervalle sind bekannt und konstant. Gleiche Differenzen zwischen je zwei Meßwerten bedeuten gleiche Abstände. Ein physikalisches Beispiel ist die Celsius-Skala, ein psychologisches die Skala des Intelligenzquotienten (IQ). Da die Intervalle bekannt sind, sind Operationen des Addierens und Subtrahierens erlaubt. Man kann beispielsweise feststellen, daß ein Schüler mit IQ = 110 gegenüber einem Schüler mit IQ = 100 um das gleiche intelligenter ist wie ein Schüler mit IQ = 105 gegenüber einem mit IQ = 95. Die Skalen im Schultest haben, nach bestimmten statistischen Operationen, fast ausschließlich Intervallniveau (s. Abschn. 3.1.5.3.).

Die Verhältnisskala:

Die Verhältnisskala bildet das höchste Niveau der Skalenhierarchie. Sie unterscheidet sich von der Intervallskala durch ihren echten Nullpunkt. Dadurch wird es möglich, Verhältnisse zu bilden. Man erhält Aussagen in der Art: 2 kg ist halb so schwer wie 4 kg oder 8 kg ist doppelt so schwer wie 4 kg. Bei psychologischen Messungen wird nur äußerst selten Verhältnisniveau erreicht, da ein absoluter Nullpunkt meistens nicht definiert werden kann. (Was sollte es z. B. bedeuten: IQ oder Schulleistung gleich Null?) Verhältnisskalen erlauben Aussagen über die Gleichheit von Summen, Produkten und Quotienten.

Zusammenstellung der Skalen (in Anlehnung an SIXTL 1967, S. 12):

Skala:	Information:	Beispiel:
Nominal:	qualitative Information: gleich oder ungleich	männlich/weiblich
Ordinal:	quantitative Information: größer oder kleiner	Schulnoten
Intervall:	quantitative Information: Gleichheit von Intervallen und Unterschieden	Celsius-Skala, IQ-Skala, Wertpunkte-Skala in Schultests
Verhältnis:	quantitative Information: Gleichheit von Summen, Produkten und Quotienten	physikalische Skalen wie etwa cm, kg, m^2 m^3

Unter bestimmten Voraussetzungen können die Skalen ineinander überführt werden. Diese Überführung nennt man Skalentransformation. Nominalskalen können in den Fällen, in denen die Klassengrenzen fließend sind, in Rangskalen umgewandelt werden. Durch immer differenziertere Klassenbildung entstehen mehrere Klassen, die schließlich in eine Rangfolge gebracht werden können. Die beiden qualitativen Klassen „gute Schüler" und „schlechte Schüler" lassen sich über die Ausdifferenzierung in sehr gute, gute, mittelmäßige, schlechte, sehr schlechte Schüler auf einer Rangskala abbilden. Umgekehrt können sogar Ordnungen auf Verhältnisskalen zu qualitativen Klassen reduziert werden. Etwa beim Lebensalter die Aufteilung in „Kinder" und „Erwachsene". Bei festen Klassengrenzen, wie etwa beim Geschlecht (männlich/weiblich), besteht diese Möglichkeit natürlich nicht.

Allgemein steht und fällt pädagogisch-psychologische Diagnostik mit der Möglichkeit, Skalen transformieren oder neu definieren zu können. So kann beispielsweise eine Rangskala nach Einführung bestimmter Grundannahmen in eine Intervallskala überführt werden, sofern die tatsächlichen empirischen

Verhältnisse dabei nicht verändert werden. Es soll hier nur auf die prinzipielle Möglichkeit der Transformation hingewiesen werden. Von praktischer Bedeutung ist die Skalentransformation bei der Konstruktion von Schultests, wo es darum geht, Testergebnisse objektiv interpretieren zu können (s. Abschn. 3.1.5.2. und 3.1.5.3.).

3.1.1.3. Gütekriterien von Messungen

Bisher wurde festgestellt, daß die Informationen, die einem Meßwert zu entnehmen sind, vom Niveau der Skala abhängen, auf der er repräsentiert ist. Bei der Interpretation eines Meßwertes interessiert nun — unabhängig vom Skalenproblem — die Frage nach der Richtigkeit, der Güte des Meßwertes. Inwieweit kann man sich auf einen gemessenen Zahlenwert verlassen?

Zur Bewertung der Güte stehen drei Kriterien zur Verfügung:

Die *Reliabilität (Zuverlässigkeit):* Unter Reliabilität versteht man den Grad der Genauigkeit der Messung. Messungen sind dann reliabel (zuverlässig), wenn die Werte unabhängig vom Zeitpunkt der Messung sind. Verschiedene Messungen desselben Objekts mit demselben Meßinstrument sollten stets dasselbe Ergebnis erbringen.

Die *Validität (Gültigkeit):* Unter Validität versteht man das Ausmaß, mit dem ein Meßinstrument tatsächlich das mißt, was es messen soll. Meßwerte sind also dann valide (gültig), wenn sie das zu messende Merkmal repräsentieren. Für das Körpergewicht einer Person sind Meßwerte der Kilogramm-Skala valide Werte. Meßwerte der Zentimeter-Skala sind bezüglich des Körpergewichts nur wenig valide. Da große Personen im allgemeinen schwerer sind als kleine, besteht immerhin noch eine geringe Validität der cm-Skala für das Körpergewicht.

Die *Objektivität:* Messungen sind dann objektiv, wenn der Meßwert unabhängig vom Messenden ist. Wenn zwei oder mehrere Personen mit dem gleichen Meßinstrument (z. B. Waage) den gleichen Gegenstand messen, sollten sie zum gleichen Ergebnis kommen.

Optimale Messungen zeichnen sich durch hohe Reliabilität, hohe Validität und hohe Objektivität aus. Optimale Messungen werden allenfalls mit naturwissenschaftlichen Instrumenten erreicht. Psychologische oder pädagogische Messungen (z. B. Schulleistungsmessungen) erfüllen die drei genannten Gütekriterien nur annäherungsweise.

3.1.1.4. Zusammenfassung

Es wurde dargestellt, daß Testen in der Praxis ein Messen von Personen oder Personengruppen hinsichtlich bestimmter Merkmale bedeutet. Nach

der Definition von Messen als Zuordnung eines numerischen Relativs zu einem vorgefundenen empirischen Relativ wurde gezeigt, daß Meßwerte auf Skalen mit unterschiedlichem Niveau repräsentiert sind: Nominal-, Ordinal-, Intervall- und Verhältnisskalen. Je höher das Skalenniveau, desto aussagekräftiger sind die einzelnen Meßwerte hinsichtlich der Beziehungen zwischen den Objekten.

Unabhängig vom Skalenniveau, d. h. also auch unabhängig von den Beziehungen zwischen den Objekten ist zu fragen, wie gut ein Meßwert die Wirklichkeit abbildet. Es wurden die Begriffe Reliabilität, Validität und Objektivität als Gütekriterien von Messungen eingeführt. Je reliabler, valider und objektiver eine Messung ist, desto größer ist die Wahrscheinlichkeit, daß ein Meßwert der empirischen Wirklichkeit entspricht.

3.1.2. Grundlagen der klassischen Testtheorie

3.1.2.1. Fragestellung

Trotz der praktischen Übereinstimmung von Messen und Testen ist die klassische Testtheorie keine Meßtheorie. Sie fragt nicht danach, wie ein Meßwert entsteht oder entstehen sollte, vielmehr betrachtet sie vorgegebene Meßwerte. Konsequenterweise ist dann das Hauptproblem der klassischen Testtheorie die Frage nach der Güte einer schon vorhandenen Messung.

Um ein konkretes Beispiel zu nennen: In einem Rechtschreibtest, in dem 40 falsch geschriebene Wörter herauszufinden sind, findet ein Schüler 26. Bei einer maximalen Punktzahl von 40 erreicht er also 26 Punkte. Die Frage nach der Güte dieser Messung (dieses Tests) differenziert sich in drei Einzelfragen:

1. Kam der Testwert durch zufällige Einflüsse oder aufgrund einer überdauernden Fähigkeit des Schülers zustande? (Frage nach der Reliabilität).
2. Sagt der Testwert etwas über die Rechtschreib-Fähigkeit des Schülers aus? (Frage nach der Validität).
3. Kam der Testwert aufgrund vergleichbarer, kontrollierter Umstände zustande? (Frage nach der Objektivität).

Die klassische Testtheorie versucht, diese Fragen mit Hilfe einiger Grundannahmen zu beantworten.

3.1.2.2. Axiome der klassischen Testtheorie

Wenn die Messung überhaupt sinnvoll sein soll, so muß das zu messende Merkmal in einem gewissen Ausprägungsgrad vorhanden sein. Man muß

daher annehmen, daß zu jedem Meßwert X_{ij} einer Vp i im Test j (z. B. die 26 Punkte des Schülers im Rechtschreibtest) ein sogenannter *„wahrer Wert"* T_{ij} (T für engl. true) existiert.

(A 1) zu X_{ij} existiert ein T_{ij}

Nun ist jede Messung wahrscheinlich mit einem mehr oder minder großen Fehler behaftet. Der Meßwert X_{ij} ist also nicht gleich dem wahren Wert T_{ij}, vielmehr setzt er sich aus dem wahren Wert und einem Fehlerwert E_{ij} (E für engl. error) zusammen.

(A 2) $$X_{ij} = T_{ij} + E_{ij}$$

Bei sorgfältig kontrollierter Messung kann man annehmen, daß der Fehlerwert E_{ij} zufällig entstanden ist. Eine nicht sorgfältige Messung in diesem Sinne läge vor, wenn konstante und systematische Fehler auftreten (z. B. Wiegen mit einer vorgehenden Waage). Wiederholt man eine Messung beliebig oft, so wäre jede einzelne Messung mit einem Fehler behaftet, der mal größer oder kleiner wäre, mal in die eine oder die andere Richtung ginge. Insgesamt gesehen heben sich die Fehler dann gegenseitig auf, d. h. der Mittelwert aller Fehlerwerte $\bar{E}_{ij}$ einer Vp i aus beliebig vielen Wiederholungen des Tests j ist gleich Null.

(A 3) $$\bar{E}_{ij} = 0$$

Aus den genannten Axiomen lassen sich mehrere Hilfssätze ableiten, die eine praktische Testkonstruktion erst ermöglichen: Da die Fehlerwerte zufällig zustande kommen und in keinerlei systematischen Zusammenhang stehen, sind sie auch völlig unabhängig von den wahren Werten. Formal ausgedrückt: Die Korrelation r_{TE} der wahren Werte T und der Fehlerwerte E ist gleich Null.

(S 1) $$r_{TE} = 0$$

Aufgrund der Zufallsbedingtheit der Fehlerwerte kann man annehmen, daß die Fehlerwerte E_1 und E_2 zweier Meßreihen mit demselben Meßinstrument ebenfalls nicht miteinander korrelieren. Die Fehlerwerte stehen also auch untereinander in keinerlei Zusammenhang. Ihre Korrelation $r_{E1\,E2}$ ist ebenfalls gleich Null.

(S 2) $$r_{E1\,E2} = 0$$

Die bisherigen Ableitungen beziehen sich auf den hypothetischen Fall beliebig vieler Meßwertwiederholungen am selben Objekt (Schüler, Vp) mit demselben Meßinstrument (Test). Wenn dasselbe Objekt nun nicht mehrmals mit einem Instrument, sondern mit verschiedenen Instrumenten, die aber das gleiche erfassen, gemessen werden, so spricht man von parallelen Messungen. (Beispiel: Bestimmung des Gewichts eines Gegenstandes mit verschiedenen Waagen). Auch bei psychologischen Tests oder bei Schultests besteht die Möglichkeit paralleler Messungen. Man kann zwei Tests konstruieren, die inhaltlich genau das gleiche Merkmal erfassen und die in derselben Stichprobe den gleichen Mittelwert und die gleiche Streuung erreichen.

Für den Fall von parallelen Messungen kann man ableiten: Wenn die Messungen tatsächlich parallel sind, so müssen die wahren Werte T_{i1} und T_{i2} der Vp i in den beiden Tests 1 und 2 gleich sein.

(S 3) $$T_{i1} = T_{i2}$$

Auch die Fehler in parallelen Messungen stehen in keinem Zusammenhang zueinander. In beiden Messungen werden also die Mittelwerte der Fehler beides Mal Null sein. Die Streuungen der Fehler s_{E1} und s_{E2} in den beiden Messungen müssen ebenfalls gleich sein.

(S 4) $$s_{E1} = s_{E2}$$

Bei der Darstellung der Axiome und den daraus abgeleiteten Sätzen wurde deutlich, daß die klassische Testtheorie Annahmen über die Fehlerhaftigkeit von Messungen in den Mittelpunkt rückt. Man kann sie daher als „Fehlertheorie" charakterisieren. Dies wird noch einmal im nächsten Abschnitt deutlich.

3.1.2.3. Das Reliabilitätskonzept

„Unter der Reliabilität eines Tests versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes ... Verhaltensmerkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht". (LIENERT 1969, S. 14).

Bei beliebig vielen Meßwiederholungen (oder beliebig vielen Parallelmessungen) ergibt sich ein Mittelwert von

$$\bar{X}_{ij} = \bar{T}_{ij} + \bar{E}_{ij}; \text{ (siehe A 2)}$$

da

$$\bar{E}_{ij} = 0 \text{ ist, (siehe A 3)}$$

ergibt sich

$$\bar{X}_{ij} = \bar{T}_{ij}$$

$\bar{T}_{ij}$ ist der Mittelwert der wahren Werte. Der wahre Wert einer Vp i im Test j bleibt, bei Testwiederholungen jedoch gleich, so daß der Mittelwert $\bar{T}_{ij}$ gleich dem einzelnen wahren Wert T_{ij} ist.

Er ist also:

(S 5) $$\bar{X}_{ij} = T_{ij}$$

d. h. der Mittelwert einer Menge von Meßwiederholungen oder (Parallelmessungen) ist gleich dem wahren Wert der Vp.

Nun kann ein Test bei einer Vp nicht beliebig oft wiederholt werden, da sich dadurch das zu messende Merkmal ändern kann (z. B. Lernen bei Rechtschreibtests). Um eine einzige Messung bezüglich ihrer Reliabilität beurteilen zu können, sind Informationen darüber notwendig, wie sich Testwerte, wahre Werte und Fehlerwerte in einer repräsentativen Stichprobe von Vpn verteilen. Es läßt sich zeigen, daß die Varianz der Testwerte X

(Testvarianz s^2_X) sich aus der Varianz der wahren Werte T (wahre Varianz s^2_T) und der Varianz der Fehlerwerte E (Fehlervarianz s^2_E) zusammensetzt.

(S 6) $$s^2_X = s^2_T + s^2_E$$

Je größer nun der Anteil der wahren Varianz s^2_T und je kleiner der Anteil der Fehlervarianz s^2_E an der Testvarianz s^2_X ist, desto reliabler ist die Messung. Die Reliabilität eines Tests kann demnach definiert werden durch den Quotienten s^2_T/s^2_X.

Bei einem geringen Anteil der wahren Varianz geht der Quotient gegen Null, bei einem hohen Anteil gegen 1. Der Quotient aus zwei Varianzen ist statistisch definiert als Korrelationskoeffizient. Dementsprechend ist das Relabilitätsmaß eines Tests ein Korrelationskoeffizient, der üblicherweise mit r_{tt} (t = test) bezeichnet wird. Für die Reliabilität eines Tests gilt also

(S 7) $$r_{tt} = \frac{s^2_T}{s^2_X}$$

Diese Korrelation ist empirisch feststellbar, wie im Abschnitt 3.1.3. gezeigt wird.

3.1.2.4. Validitätskonzepte

„Die Validität eines Testes gibt den Grad der Genauigkeit an, mit dem dieser Test ... diejenige Verhaltensweise, die er messen soll oder zu messen vorgibt, tatsächlich mißt" (LIENERT 1969, S. 16).

Ein Punktwert in einem Test ist immer Ergebnis des spezifischen Testverhaltens einer Vp. Das Problem der Validität eines Tests besteht nun in der Frage, inwieweit dieses spezifische Verhalten auf anderes Verhalten verallgemeinert werden kann. Beim Schluß vom Testverhalten auf das allgemeine Verhalten außerhalb der Testsituation ergeben sich mehrere Möglichkeiten. Daher können verschiedene Validitätskonzepte beschrieben werden.

Die Inhaltsvalidität (content validity):

Inhaltsgültigkeit eines Tests liegt vor, wenn das Testverhalten als repräsentative Verhaltensstichprobe einer Verhaltensgesamtheit betrachtet werden kann. Wenn beispielsweise ein Schultest Additionsaufgaben enthält, so

wird aufgrund des Testverhaltens auf die Fähigkeit „Additionsaufgabenlösen-können" geschlossen. Inhaltsgültigkeit wird also durch Repräsentationsschluß festgestellt.

Bei den gebräuchlichen Schultests wird meistens inhaltliche Validität angenommen. Um überhaupt sinnvoll Schulleistung zu erfassen, müssen die Inhalte eines Schultests den Unterrichtsinhalten entsprechen. So ist es wenig sinnvoll, in einem Rechentest Aufgaben aufzunehmen, die im Unterricht überhaupt noch nicht behandelt wurden. Unterrichtsinhalte sind durch Lehrpläne festgelegt. Dementsprechend müssen die Testinhalte ebenfalls in Übereinstimmung mit den Lehrplänen konstruiert werden. Stimmt ein Test inhaltlich mit den Lehrplänen überein, so spricht man von Lehrplangültigkeit oder curricularer Gültigkeit.

Kriteriumsbezogene Validität: Übereinstimmungsvalidität (concurrent validity) und Vorhersagevalidität (predictive vallidy)

Bei beiden Gültigkeitskonzepten wird vom Testverhalten auf ein definiertes Kriteriumsverhalten geschlossen. Wichtig ist dabei, daß diese Schlußfolgerung nur aufgrund eines empirisch ermittelten Zusammenhanges möglich ist, eines Korrelationsschlusses. Ein Beispiel dafür ist die Gültigkeit von Schulreifetests für das Kriterium „Klassenziel der ersten Klassen erreicht" bzw. „nicht erreicht". In diesem Falle handelt es sich um ein Beispiel der Vorhersagegültigkeit. Der Schulreifetest wird im allgemeinen vor der Einschulung durchgeführt und nach Ablauf des ersten Schuljahres wird überprüft, ob das Kriterium erreicht wird.

Übereinstimmungsgültigkeit liegt dann vor, wenn Test und Kriterium gleichzeitig erhoben werden. Man könnte nach der Zeugnisausgabe am Ende eines Schuljahres sofort Intelligenztests durchführen und überprüfen, wie Intelligenztest und Schulnoten miteinander korrelieren. Diese Korrelation könnte dann als Übereinstimmungsgültigkeit des Intelligenztests interpretiert werden.

Übereinstimmungs- und Vorhersagevalidität werden also durch empirisch ermittelte Korrelationskoeffizienten r_{tc} (t = test, c = criterion) angegeben. Sie unterscheiden sich nur durch die zeitliche Verschiebung von Test und Kriterium.

Konstruktvalidität:

Bei der Konstruktvalidität (Cronbach & Meehl 1955) wird vom Testverhalten auf theoretische Konzepte (Konstrukte) geschlossen, die „hinter" dem Verhalten stehen, die es erklären. Aufgrund des Konstrukts werden Hypothesen über das zu erwartende Verhalten aufgestellt. Eine Verhaltensmessung durch einen Test muß dann in Richtung der Hypothesen ausfallen.

Ein Beispiel aus der pädagogisch-psychologischen Forschung: Aus dem Konstrukt „Leistungsmotivation" ließe sich folgende Hypothese ableiten: Bei gleicher Intelligenz erreichen hoch leistungsmotivierte Schüler bessere Schulleistungen als wenig leistungsmotivierte Schüler. Um nun die Konstruktvalidität eines Leistungsmotivationstests zu überprüfen, würde man folgende Untersuchung machen. Man testet eine genügend große Stichprobe von gleichintelligenten Schülern und teilt sie dann in zwei Gruppen ein. Eine erste Gruppe der Schüler, die in diesem Leistungsmotivationstest unter dem Durchschnitt liegen und eine zweite Gruppe jener Schüler, die über dem Durchschnitt liegen. Entsprechend der Hypothese müßten die Schüler der zweiten Gruppe über bessere Schulleistungen verfügen.

Fallen die Ergebnisse im Sinne der Hypothese aus, so gilt in diesem Fall die Validität des Tests als gegeben. Wird die Hypothese nicht bestätigt, so bestehen zwei Interpretationsmöglichkeiten: 1. Der Test ist nicht valide im Sinne des Konstrukts, 2. Die abgeleitete Hypothese ist falsch. Es wird deutlich, daß zur Bestimmung der Konstruktvalidität eines Tests mindestens ebenso viele Untersuchungen notwendig sind wie Hypothesen gebildet werden können. Somit kann die Konstruktvalidierung eines Tests eigentlich nie abgeschlossen werden. Man ist statt dessen auf eine Beschreibung möglichst vieler Situationen angewiesen, in denen das Verfahren sich bewährte oder nicht bewährte.

3.1.2.5. Objektivitätsarten

Unter Objektivität eines Tests versteht man „den Grad, in dem die Ergebnisse eines Testes unabhängig vom Untersucher sind. Ein Test wäre demnach vollkommen objektiv, wenn verschiedene Untersucher bei demselben Probanden zu gleichen Ergebnissen gelangten" (LIENERT 1969, S. 13).

Bei der Anwendung von Tests gibt es mehrere kritische Stellen, an denen die Unabhängigkeit des Testergebnisses vom Untersucher gefährdet ist. Entsprechend können formal unterschiedliche Objektivitätsarten festgestellt werden.

Die Durchführungsobjektivität:

Unterschiedliches Verhalten der Testleiter kann zu unterschiedlichem Testverhalten der Probanden und damit zu unterschiedlichen Testergebnissen führen. Eine einigermaßen hohe Durchführungsobjektivität ist nur dann gewährleistet, wenn der Test genaue Verhaltensregeln für den Testleiter angibt. Dazu gehören Anweisungen, wie in welchen Situationen zu reagieren ist und welche Hilfen erlaubt oder unerlaubt sind. Dazu gehört weiterhin, daß die Instruktionen zu den einzelnen Testaufgaben schriftlich fixiert sind

und vom Testleiter wörtlich wiedergegeben werden. Hohe Durchführungsobjektivität ist nur durch eine drastische Einschränkung der sozialen Interaktionen zwischen Testleiter und Testperson während der Testdurchführung erreichbar.

Die Auswertungsobjektivität:

Um ein hohes Maß an Auswertungsobjektivität zu erreichen, muß eindeutig festgelegt sein, wie das Testverhalten zu bewerten ist. Bei Aufgaben, die eine eindeutig richtige Lösung haben, ist selbstverständlich die Auswertungsobjektivität höher als bei Aufgaben, die mehrdeutige Lösungen zulassen. Es kann z. B. objektiv festgestellt werden, ob eine Rechenaufgabe richtig gelöst wurde. Es ist dagegen weniger eindeutig, ob eine freie Antwort auf die Frage: „Was ist Gerechtigkeit?" richtig beantwortet wurde.

Die Interpretationsobjektivität:

In dem Maße, in dem die Interpretation eines Testergebnisses vom Interpreten unabhängig ist, kann man von Interpretationsobjektivität sprechen. Daher ist zu fordern, daß ein Test genaue Anweisungen enthält, wie ein bestimmtes numerisches Testergebnis zu interpretieren ist. Die Interpretation richtet sich aus am theoretischen Konzept, das dem Test zugrunde liegt und am Vergleich zwischen dem Testergebnis des einzelnen Probanden und den Ergebnissen einer Bezugsgruppe.

Erinnern wir uns an das Beispiel des Schülers mit 26 Punkten in einem Rechtschreibtest. Um dieses Ergebnis einigermaßen objektiv interpretieren zu können, muß der Test angeben, welches Konzept von Rechtschreibleistung ihm zugrunde gelegt wurde (z. B. ist die Groß-Klein-Schreibung, die Schreibweise von Dehnlauten u. ä. betroffen). Der Test muß auch angeben, ob 26 Punkte im Vergleich mit anderen Schülern eine gute oder schlechte Leistung bedeuten. Die Angabe des theoretischen Konzeptes fehlt im Schultest häufig. Der Vergleich mit einer Bezugsgruppe wird durch die Normierung des Tests (Abschn. 3.1.5.3.) geleistet.

Allgemein kann festgehalten werden, daß Testergebnisse verschiedener Probanden nur miteinander vergleichbar sind, wenn sie unter vergleichbaren Bedingungen zustande gekommen sind. Nur Tests, bei denen die Testdurchführung, Testauswertung und Testinterpretation exakt vorgeschrieben sind, werden als *standardisierte* Tests bezeichnet. Nur solche Tests können die Objektivität einigermaßen gewährleisten.

3.1.2.6. Zusammenhänge zwischen den Gütekriterien

Es dürfte leicht nachzuvollziehen sein, daß ein Test, der nicht objektiv ist, nicht reliabel und nicht valide sein kann. Seine Ergebnisse entstehen eher

zufällig. Objektivität eines Tests ist also notwendige Voraussetzung für seine Reliabilität und Validität. Sie ist allerdings keine hinreichende Voraussetzung: Ein Test kann durchaus objektiv sein, er muß dann aber noch nicht unbedingt reliabel und valide sein.

Nur wenn ein Test reliabel ist, kann er auch valide sein. Geringe Reliabilität eines Tests bedeutet ja hohe Zufälligkeit der Ergebnisse. Zufällige Ergebnisse können aber in bezug auf ein definiertes Merkmal nicht valide sein. Andererseits kann ein Test sehr wohl reliabel sein, er braucht deshalb aber nicht valide zu sein. Er kann zwar mit sehr hoher Präzision messen; er muß dabei aber nicht unbedingt das zu messende Merkmal tatsächlich erfassen. Objektivität und Reliabilität eines Tests sind also notwendige, aber nicht hinreichende Voraussetzungen für eine hohe Validität.

Umgekehrt kann man aus einer nachgewiesenen Validität folgern, daß der Test auch einigermaßen reliabel und objektiv sein muß.

3.1.2.7. Zusammenfassung

Die klassische Testtheorie ist eine Meßfehlertheorie. Sie fragt danach, wie fehlerhaft ein vorgegebener Testwert ist. Es wurde auf die Axiomatik eingegangen, die im wesentlichen besagt, daß ein gemessener Wert (Testwert) sich aus einem „wahren Wert" und einem „Fehlerwert" zusammensetzt. Wahre Werte und Fehlerwerte stehen ebenso wie die Fehlerwerte untereinander in keinerlei systematischem Zusammenhang. Fehlerwerte sind rein zufällig. Die Konzepte der Gütekriterien von Tests (Reliabilität, Validität und Objektivität) wurden auf dem Hintergrund der Axiomatik dargestellt. Ihre Zusammenhänge wurden kurz erwähnt. In den nachfolgenden Abschnitten soll gezeigt werden, wie die klassische Testtheorie bei der Konstruktion standardisierter Schultests angewendet wird.

3.1.3. Reliabilität von Schultests

Zur Bestimmung der Reliabilität eines Tests stehen mehrere Methoden zur Verfügung. Ihre Anwendung hängt zum großen Teil von den praktischen Möglichkeiten während der Testkonstruktion und dem Anwendungsbereich des Tests ab.

3.1.3.1. Retest-Reliabilität

Bei dieser Methode wird ein Test an einer Stichprobe durchgeführt und nach einer gewissen Zeit an derselben Stichprobe wiederholt: Test und Retest. Die Ergebnisse der beiden Testungen werden miteinander korreliert. Der so ermittelte Korrelationskoeffizient r_{tt} ist ein direktes Maß für die

Reliabilität des Tests. Eine Korrelation von $r_{tt} = 1$ kann nur zustande kommen, wenn die Reihenfolge der Versuchspersonen hinsichtlich ihrer Testwerte bei beiden Testungen gleich ist. Es ist also nicht notwendig, daß die Versuchsperson im zweiten Test genau den Wert des ersten Tests wieder erreichen. Die Korrelation von $r_{tt} = 1$ kommt auch zustande, wenn alle Versuchspersonen um die gleiche Punktzahl besser (schlechter) geworden sind. Ein Übungs- und Lerneffekt durch den ersten Test muß also nicht unbedingt die Reliabilität verringern. Die Reliabilitätsminderung durch Übung ist jedoch sehr wahrscheinlich, da wohl nur in den seltensten Fällen ein gleichmäßiger Übungseffekt für alle Versuchspersonen auftritt.

Eine zweite Fehlerquelle, die zur Reliabilitätsminderung führen kann, liegt einfach in der verstreichenden Zeit. Obwohl unter Umständen ein Test ein Merkmal zuverlässig erfaßt, kann eine niedrigere Retest-Reliabilität auftreten und zwar dann, wenn sich das Merkmal in der Zwischenzeit verändert hat. Bei veränderten Merkmalen müssen die Versuchspersonen im zweiten Test andere Werte erhalten; die Test-Retest-Korrelation sinkt ab.

Auf diese Fehlerquellen ist es zurückzuführen, daß die Methode der Retest-Reliabilität bei Schultests, im Gegensatz zu anderen psychometrischen Tests, nur eine geringe Rolle spielt. Die Retest-Reliabilität eines Schultests muß eigentlich niedrig ausfallen, da in der Zwischenzeit Unterricht stattfindet, in dem alle Schüler optimal gefördert werden sollen. Nach einer gewissen Zeit sollten alle Schüler eine bestimmte unterrichtsbedingte Leistung zeigen (z. B. Kenntnisse über historische Daten). Würde man zu dieser Zeit einen Schultest zum zweiten Mal durchführen, müßten nahezu alle Schüler den möglichen Höchstpunktwert erreichen. Die Test-Retest-Korrelation ginge gegen Null. Eine hohe Retest-Reliabilität eines Schultests wäre also ein Indiz für wenig effektiven Unterricht oder für Unterricht, der alle Schüler gleich fördert, aber dafür sorgt, daß die guten Schüler die guten und die schlechten Schüler die schlechten bleiben.

3.1.3.2. Paralleltest-Reliabilität

Man konstruiert zu einem bestehenden Test A einen optimalen Paralleltest A'. Parallel in diesem Sinne heißt: Die einzelnen Testaufgaben und der gesamte Test haben in derselben Stichprobe den gleichen Mittelwert und die gleiche Streuung (statistische Parallelität).

Weiterhin ist zu fordern, daß die beiden Tests A und A' inhaltlich ähnlich sind und das gleiche Merkmal messen (inhaltliche Parallelität). Nach der Konstruktion zweier Paralleltests führt man beide in einer Stichprobe durch und korreliert die Ergebnisse miteinander. Eine hohe Korrelation ist dabei nur zu erreichen, wenn beide Tests das gleiche Merkmal zuverlässig erfassen.

Für Schultests ist die Paralleltest-Reliabilität aus zwei Gründen sehr bedeutsam. Zum ersten fällt bei dieser Methode die Fehlerquelle des zeitlichen Ablaufs aus, da die beiden Paralleltests gleichzeitig durchgeführt werden. Der zweite Grund ist praktischer: Für Schultests empfiehlt sich grundsätzlich die Anwendung von Paralleltests, da das Abschreiben der Schüler voneinander verhindert werden kann, wenn Banknachbarn zwar ähnliche aber doch nicht identische Tests bearbeiten. Sofern ein Schultest über Parallelformen verfügt, wird meistens die Paralleltestreliabilität angegeben. Sie liegt im allgemeinen bei $r_{tt} = 0.9$.

Das strengste Reliabilitätskriterium liegt in der Kombination von Retest- und Paralleltest-Reliabilität. Die Testwiederholung geschieht dabei nicht mit demselben Test, sondern mit einem Paralleltest. Obwohl eigentlich für alle guten Tests eine strenge Überprüfung der Reliabilität wünschenswert ist, scheidet diese Methode für Schultests ebenso aus, wie die einfache Retest-Reliabilitätskontrolle.

3.1.3.3. Halbierungs-Reliabilität

Steht zur Reliabilitätskontrolle nur ein Test zur Verfügung und ist die Bestimmung der Retest-Reliabilität nicht sehr sinnvoll, so besteht die Möglichkeit, den vorhandenen Test in zwei Hälften aufzuteilen und die Ergebnisse einer Stichprobe zu korrelieren. Man stellt dann eigentlich eine Paralleltest-Reliabilität fest, wobei die beiden Testhälften als Paralleltest definiert sind. Da jedoch die beiden Paralleltests (Testhälften) nur halb so lang sind wie der ursprüngliche Test, muß der ermittelte Korrelationskoeffizient statistisch aufgewertet werden. Dies ist ohne weiteres möglich (für Details: LIENERT 1969, S. 219—225). Die Halbierungs-Reliabilität sagt etwas über den inneren Zusammenhang des Tests aus.

Besondere Bedingungen während der einzigen Testdurchführung, beispielsweise mangelhafte Durchführungsobjektivität oder Aufgaben, die eigentlich nicht in den Test hineingehörten, können die Halbierungs-Reliabilität beeinflussen, ohne daß die Fehlerquelle bemerkt wird. Die Halbierungs-Reliabilität wird bei Schultests vereinzelt angegeben. Die Koeffizienten liegen im allgemeinen ebenfalls bei $r_{tt} = 0.9$.

3.1.3.4. Die Konsistenz-Reliabilität

Die Konsistenz-Reliabilität (oder auch innere Konsistenz) kann als Weiterentwicklung der Halbierungs-Reliabilität angesehen werden. Man kann jeden Test ja nicht nur in zwei Hälften teilen, sondern man kann ihn dritteln, vierteln usw. Schließlich kommt man zu soviel Teilen, wie der Test Aufgaben hat. Die durchschnittliche Korrelation jedes Teiles (jeder Aufgabe) mit

jedem (jeder) ergibt ein Maß für den inneren Zusammenhang des Tests. Ein hoher Korrelationskoeffizient besagt, daß alle Aufgaben dasselbe erfassen. Wenn alle Aufgaben dasselbe erfassen, so sind zufällige Messungen ausgeschlossen, der Test ist also sehr reliabel. Bezüglich der Fehler, denen die Konsistenz-Reliabilität unterliegt, gilt das gleiche wie bei der Halbierungs-Reliabilität.

Da zur Bestimmung der Konsistenz-Reliabilität (zum rechnerischen Verfahren vgl. LIENERT 1969, S. 225 ff.) nur eine einzige Testdurchführung notwendig ist, wird sie im Laufe der praktischen Testkonstruktion fast immer berechnet. Nahezu alle publizierten Schultests geben ihre Konsistenz-Reliabilität an (r_{tt} um 0.9).

3.1.3.5. Das Konzept des Standardmeßfehlers

Mit der Angabe der Reliabilität eines Tests ist natürlich die Frage, wie zuverlässig ein individueller Punktwert eines Schülers ist, noch nicht beantwortet. Um diese Frage beantworten zu können, braucht man außer der Kenntnis der Reliabilität des Tests noch Informationen über die Verteilung der Fehlerwerte.

In Abschnitt 3.1.2.3. wurde festgestellt, daß

$$r_{tt} = \frac{s_T^2}{s_X^2} \qquad \text{oder } s_T^2 = r_{tt} \cdot s_X^2 \qquad (1)$$

und

$$s_X^2 = s_T^2 + s_E^2 \text{ oder } s_T^2 = s_X^2 - s_E^2 \qquad (2)$$

Man hat hier zwei Gleichungen (1) und (2), in denen die notwendigen Informationen enthalten sind. Durch Gleichsetzen erhält man:

$$s_X^2 - s_E^2 = r_{tt} \cdot s_X^2$$

Durch weiteres Umformen kommt man zu

$$s_E = s_X \cdot \sqrt{1 - r_{tt}}$$

Alle Glieder der rechten Seite sind empirisch feststellbar. Man kann daher die Fehlerstreuung s_E berechnen. Die Fehlerstreuung wird als *Standardmeßfehler* bezeichnet.

Der Standardmeßfehler ist ein Schätzverfahren für die Genauigkeit eines Testwertes. Unter Hinzunahme von Wahrscheinlichkeitsüberlegungen kann eine *obere* und *untere Vertrauensgrenze* CL (confidential limit) eines Meßwertes X bestimmt werden. Die beiden Grenzen geben den Bereich an, in

dem sich der wahre Wert T, der zu einem bestimmten Testwert X gehört, bewegen kann. Dazu wird der Standardmeßfehler mit dem Wahrscheinlichkeitsmaß z multipliziert.

Die obere und untere Vertrauensgrenze CL eines Testwertes X sind

$$CL = X \pm z \cdot s_E$$

Im allgemeinen werden für das Wahrscheinlichkeitsmaß zwei Wahrscheinlichkeiten angenommen: 95 % und 99 %; die zugehörigen z-Werte sind 1,96 bzw. 2,54.

Rechenbeispiel: Nehmen wir an, ein Schüler erreiche in einem allgemeinen Schultest einen Punktwert von 80. Die Reliabilität des Tests sei $r_{tt} = 0{,}84$; die Streuung sei $s_x = 10$. Die Frage ist nun, in welchem Bereich liegt der wahre Wert des Schülers mit 95%iger (bzw. 99%iger) Wahrscheinlichkeit?

Der Standardmeßfehler ist: $s_E = 10 \cdot \sqrt{1 - 0{,}84}$

$$s_E = 10 \cdot 0{,}4 = 4$$

Die Vertrauensgrenzen für 95%ige Wahrscheinlichkeit sind:

$$CL_{95\%} = 80 \pm 1{,}96 \cdot 4$$

$$CL_{95\%} = 80 \pm 7{,}84$$

Man kann also sagen: Der wahre Wert des Schülers liegt mit 95%iger Wahrscheinlichkeit im Bereich zwischen ~ 72,2 Punkten und ~ 87,8 Punkten.

Für die Vertrauensgrenzen mit 99%iger Wahrscheinlichkeit gilt analog:

$$CL_{99\%} = 80 \pm 2{,}54 \cdot 4$$

$$CL_{99\%} = 80 \pm 10{,}18$$

Bei der Interpretation von Testergebnissen sind der Standardmeßfehler und die Vertrauensgrenzen unbedingt zu beachten. Sie bieten den großen Vorteil, die Fehlerhaftigkeit jedes einzelnen Testergebnisses abschätzen zu können. Soweit der Standardmeßfehler und die Vertrauensgrenzen in den Testmanualen nicht angegeben sind, sei es dem Benutzer von Schultests dringend empfohlen, sie selbst auszurechnen.

3.1.3.6. Zusammenfassung

In diesem Abschnitt wurden vier Methoden zur Bestimmung der Reliabilität von Tests vorgestellt. Allgemein kann festgestellt werden, daß die Re-

liabilität von Schultests befriedigend hoch ist (siehe auch SÜLLWOLD 1964). Am Konzept des Standardmeßfehlers wurde gezeigt, daß man abschätzen kann, wie fehlerhaft ein einzelner Testwert eines Schülers sein kann. Hieraus ergab sich die Forderung, den Standardmeßfehler bei der Interpretation einzelner Testwerte hinzuzuziehen.

3.1.4. Validität von Schultests

Das Validitätsproblem von Tests ist allgemein am schlechtesten gelöst. Dementsprechend sind Validitätsuntersuchungen von Tests häufig unbefriedigend. Schultests machen darin keine Ausnahme. Im folgenden soll dargestellt werden, wie versucht wird, Validitätsprobleme von Schultests zu lösen.

3.1.4.1. Curriculare Validität

Bei Schultests zieht man sich häufig auf die „curriculare Validität" zurück, die als Spezialfall der inhaltlichen Validität angesehen werden kann (s. Abschn. 3.1.2.4.).

Zunächst legt man fest, welches Fach oder welche Fächer durch einen Schultest erfaßt werden sollen. Nach dieser Definition des Zieles werden die Lehr- und Bildungspläne aller Bundesländer analysiert. Beispielsweise wird gefragt, welche Inhalte des Faches Rechnen laut Bildungsplänen nach der ersten Hälfte des vierten Schuljahres unterrichtet worden sein müssen. Für einen entsprechenden Rechentest, der in der zweiten Hälfte des vierten Schuljahres angewendet werden soll, werden dann solche Aufgaben konstruiert, die den Anforderungen der Bildungspläne entsprechen. Prinzipiell besteht die Schwierigkeit, daß man nur durch Augenschein feststellen kann, ob ein Schultest auch den Bildungsplänen entspricht. Eine im statistischen Sinne objektive Prüfung, inwieweit eine curriculare Validität gegeben ist, kann nicht durchgeführt werden. Außerdem kann man nicht sicher sein, daß die Absichten der Bildungspläne auch von allen Lehrern genau genug eingehalten werden. Selbst wenn man allen Grund hat, eine curriculare Validität des Tests anzunehmen, kann der Fall eintreten, daß der Test für eine oder mehrere Klassen nicht valide ist, weil die Klassenlehrer den einen oder anderen Aspekt des Lehrplans über- oder unterbetont haben. Die Lehrplangültigkeit des Tests scheitert dann an der relativen Unterrichtsfreiheit des Lehrers. Die üblichen Standardformeln in den Schultests wie: „Die Gültigkeit ist logisch evident" oder „Gültigkeit in Übereinstimmung mit den Lehrplänen aller Bundesländer (einschl. Berlin-West)" (siehe Beltz-Verlag, 1968) sind dann wohl mit Vorsicht zu interpretieren. Dies um so mehr, da andere Validitätsangaben nur sehr selten gegeben werden.

3.1.4.2. Kriteriumsbezogene Validität

Die kriteriumsbezogenen Konzepte der Übereinstimmungs- und Vorhersagevalidität sind bei den Schultests mit unterschiedlichem Erfolg anwendbar. Die Hauptschwierigkeit liegt in der Bestimmung des Kriteriums, mit dem der Test übereinstimmen oder das er vorhersagen soll.

Zunächst zur Übereinstimmungsvalidität: In einem ersten Ansatz kann man natürlich die Korrelation eines Schultests mit der oder den entsprechenden Schulnoten bestimmen. Die Schulnote wäre in diesem Falle das Kriterium. Wenn ein Schultest irgend etwas mit der Schulleistung eines Schülers zu tun haben soll, so muß er in gewissem Umfange mit seinen Schulnoten zusammenhängen (korrelieren). Erinnert man sich allerdings einer der Absichten, die zur Einführung von Schultests führten, nämlich eine „gerechtere" Beurteilung der Schüler zu erreichen, so ist eine hohe Korrelation zwischen Schultestergebnissen und Schulnoten auch bedenklich: Ist nun der Test ein ebenso schlechtes Instrument wie eine Schulnote? Oder sind Schulnoten schließlich doch so gut wie Schultests? In beiden Fällen erübrigte sich die Anwendung von Schultests.

Bei dieser Art der Validitätsbestimmung erklärt man die Beurteilungen und Noten von Lehrern für schlecht und ungeeignet und begründet damit die Konstruktion von Tests. Anschließend wird die Validität der Tests überprüft, indem man ihre Übereinstimmung mit den Noten feststellt. Bei hoher Übereinstimmung werden die Tests für gut erklärt und den Lehrern mit dem Argument verkauft, ihre Noten seien ja so schlecht (WECHSLER 1956). Etwas ernsthafter: Die Validierung von Schultests an Schulnoten läuft Gefahr, in einen Zirkelschluß zu verfallen, da ihre Korrelation sowohl als Maß für die Validität der Tests als auch der Schulnoten interpretiert werden kann. Diese Gefahr besteht um so mehr, da die instrumentellen Eigenschaften der Schulnoten im allgemeinen schlechter zu veranschlagen sind als die der Schultests. Trotzdem muß man wohl davon ausgehen, daß ein bestimmtes Maß an Übereinstimmung notwendig ist. Die Schultests korrelieren zu etwa 0,6 mit den entsprechenden Schulnoten, das gleiche gilt für amerikanische Schultests (SÜLLWOLD 1964).

Bei der Bestimmung einer Vorhersagegültigkeit von Schultests ist es etwas leichter. Als vorherzusagende Kriterien lassen sich etwa heranziehen: Versetzung / Nicht-Versetzung in die nächste Klasse oder Schulstufe, Bestehen / Nicht-Bestehen einer Prüfung, erfolgreicher / nicht-erfolgreicher Schulabschluß usw. So werden Schulreifetests häufig am Kriterium der Versetzung in die zweite oder gar dritte Klasse validiert und erreichen beachtliche Validitätskoeffizienten. Aufgrund der empirisch ermittelten Korrelation zwischen Schulreifetestergebnis und dem Kriterium läßt sich voraussagen, mit welcher Wahrscheinlichkeit ein Schüler mit einem bestimmten Testergeb-

nis das Klassenziel erreichen wird oder nicht. Aufgrund ähnlicher Überlegungen wie beim Standardmeßfehler (s. Abschn. 3.1.3.5.) läßt sich auch ein *Standardschätzfehler* ableiten, mit dessen Hilfe sich ein Bereich angehen läßt, in dem der vorherzusagende Kriteriumswert mit 99 % oder 95 % Wahrscheinlichkeit liegt. Da eigentlich nur die Schulreifetests Angaben über eine Vorhersagevalidität machen, wird auf die Darstellung der Standardschätzfehler an dieser Stelle verzichtet. (Näheres: LIENERT 1969, S. 476 bis 481). Die Diskussion von MANDL & KRAPP (1972 a), KORNMANN (1972) und MANDL & KRAPP (1972 b) zeigt in beispielhafter Weise die Schwierigkeiten der Vorhersage bestimmter Kriteriumsleistungen bei Schulreifetests.

Die Bestimmungen einer Vorhersagegültigkeit wird praktisch immer schwierig, wenn der Zeitraum zwischen Testerhebung und Kriteriumserfassung lang ist. Bei einem Schultest, der zu Entscheidungen bei der Übertrittsauslese zum Gymnasium herangezogen werden soll, müßte eigentlich die Vorhersagevalidität bezüglich des Abiturs oder der Mittleren Reife bestimmt werden. Dazu müßte man während der Testkonstruktion eine repräsentative Stichprobe von Schülern des vierten Schuljahres testen und dann ihre schulische Laufbahn etwa ein Jahrzehnt lang verfolgen, um festzustellen, wer von ihnen Hauptschulabschluß, Mittlere Reife oder Abitur gemacht hat. Inzwischen hätten sich wahrscheinlich die Lehrpläne geändert, so daß die curriculare Gültigkeit verloren wäre. Der Test könnte dann nicht mehr angewandt werden. Schultests enthalten deshalb auch keine Angaben über die Vorhersagegültigkeit über längere Zeit hinweg. Gerade die Tests, die im vierten Schuljahr angewendet werden und sich als Ausleseinstrumente zum Übertritt aufs Gymnasium empfehlen, machen hierin keine Ausnahme. Zum Problem der Schuleignungsermittlung bzw. Klassifikation von Schulbegabungen siehe ausführlicher TENT (1969) und HELLER (1970, 1973).

3.1.4.3. Konstruktvalidität

Das theoretisch anspruchsvollste Gültigkeitskonzept, das der Konstruktvalidität, wäre bei den Schultests in der Anwendung am fruchtbarsten, wenn ein Konstrukt „Schulleistung" ähnlich explizit definiert wäre wie etwa das Konstrukt „Intelligenzleistung".

Die Schulzeugnisse und die einzelnen Noten sind logischerweise Repräsentanten der Schulleistung. Es konnte jedoch gezeigt werden, daß Noten mehrfach determiniert sind. Schulnoten hängen u. a. zusammen mit der Leistungsfähigkeit des Schülers, seinem Sozialstatus, seinem Geschlecht, seinem Alter usw. (Das Kapitel „Leistungsbeurteilung durch Notengebung" gibt darüber Aufschluß.) Ein Konstrukt „Schulleistung" hätte zu definieren, inwieweit solche Determinanten tatsächlich zur Schulleistung zu zählen sind oder nicht.

Erste Zugänge zum Konstrukt bieten faktorenanalytische Untersuchungen, in denen Schulnoten, Schultestleistungen, Intelligenzleistungen und nicht-kognitive Variablen eingehen (z. B. FINGERHUT & LANGFELDT 1971). Aus diesem Strukturgefüge der genannten Variablen läßt sich eine empirische Beschreibung von Schulleistung ableiten, so wie sie im gegenwärtigen Schulsystem impliziert definiert ist (s. Kap. 2.1.). Der Einsatz von Schultests im Rahmen solcher Untersuchungen ergäbe Aufschluß über ihre Validität im Sinne einer Annäherung an eine Konstruktvalidität. Untersuchungen in dieser Richtung sind spärlich; die wenigen vorhandenen standen zudem meist unter einer anderen Fragestellung, so daß eine Interpretation des jeweils verwendeten Schultests im Sinne einer Konstruktvalidierung oft schwierig ist. In den üblichen Handanweisungen fehlen leider noch Beschreibungen über solche Untersuchungen. Nach den gegenwärtigen praktischen Möglichkeiten bei der Konstruktion von Schultests wäre eine Annäherung an eine Konstruktvalidierung jedoch aussichtsreich und relativ leicht durchzuführen.

3.1.4.4. Zusammenfassung

Zur Validitätsbestimmung bei Schultests kann zusammenfassend folgendes gesagt werden: In den meisten Fällen begnügt man sich mit der Feststellung einer curricularen Validität. Eine Übereinstimmungsvalidität ist bei Schultests mangels geeigneter Kriterien nur selten bestimmbar, falls man nicht mit einer Übereinstimmung von Noten und Testleistungen zufrieden ist. Die Bestimmung einer Vorhersagevalidität scheidet bei Schultests wegen des damit verbundenen Verlustes der curricularen Validität häufig aus. Schulreifetests dagegen treffen häufig Vorhersagen über den Schulerfolg über ein bis zwei Jahre hinweg. Untersuchungen im Sinne einer Konstruktvalidierung von Schultests sind äußerst selten. Insgesamt ist die Frage nach der Validität von Schultests noch relativ unbefriedigend beantwortet.

3.1.5 Objektivität von Schultests

3.1.5.1. Durchführungs- und Auswertungsobjektivität

Bei den gebräuchlichsten Schultests ist die Durchführungsobjektivität im allgemeinen sichergestellt. Die Durchführungsbedingungen sind für jeden Schüler insofern gleich, als das Testleiterverhalten genau vorgeschrieben ist. Der Testleiter muß die Testinstruktion wörtlich vorlesen, die Reihenfolge der Aufgaben einhalten und sich an vorgegebene Zeitgrenzen halten. Die Durchführungsobjektivität ist demnach bei solchen Tests am höchsten, in denen die Durchführung durch ein Tonband gesteuert ist. Dies zeigt, daß Durch-

führungsobjektivität nur durch eine radikale Einschränkung der sozialen Interaktion zwischen Testleiter (Lehrer) und Probanden (Schüler) erreichbar ist. Dies Vorgehen ist unter pädagogischen Gesichtspunkten gerade bei jungen Schülern bedenklich. Das Fehlen der sozialen Interaktion kann hier gerade die Leistungsfähigkeit der Schüler verfälschen.

Obwohl die Testbedingungen objektiv einigermaßen konstant sind, muß man davon ausgehen, daß sie subjektiv als unterschiedlich erlebt werden. Diese subjektive Unterschiedlichkeit wird beispielsweise davon beeinflußt, ob der einzelne Schüler dem Test ängstlich oder aufgeschlossen entgegensieht, ob er gerade verärgert, ob er müde oder munter ist. Obwohl es also schwierig ist, die Bedingung gleich zu gestalten (SADER & KEIL 1966) und obwohl die stereotype Einhaltung dieser Bedingungen einzelne Schüler oder Schülergruppen benachteiligen kann (KORNMANN, ENDRIGKEIT & SANDER 1972) kann auf eine Standardisierung und Reglementierung der Testsituation nicht verzichtet werden, da nur so wenigstens annäherungsweise eine Bedingungskonstanz und damit Chancengleichheit erreicht werden kann.

Es sei nur noch kurz erwähnt, daß selbst in dem Falle, in dem die Durchführungsobjektivität vom Testautor her gesichert ist und die Testleiter sich an die Testinstruktion halten, noch Testleitereinflüsse feststellbar sind, die die Durchführungsobjektivität gefährden können. Schulkinder, die von geübten Testleitern getestet werden, können unter Umständen bessere Testleistungen erbringen als solche, die von weniger geübten Testleitern getestet werden (FINGERHUT & LANGFELDT 1973).

Das Problem der Auswertungsobjektivität ist gerade bei den Schultests noch am leichtesten zu lösen. Ihre Aufgaben beziehen sich immer auf eindeutige Lösungen. Es bleibt zum Beispiel kein Zweifel, daß *Rythmus* einen Schreibfehler enthält oder daß die Lösung der Aufgabe 36 537 : 641 = 57 ist. Fehler oder richtige Lösungen sind bei Schultests also einwandfrei zu identifizieren. Durch die Verwendung von durchsichtigen Schablonen, auf denen die zu wertenden Lösungen markiert sind, kann die Auswertung objektiv und ökonomisch gestaltet werden, solange der Auswerter die nötige Sorgfalt walten läßt. Etwas eingeschränkt kann die Auswertungsobjektivität bei Schulreifetests sein, in denen unter Umständen Muster, Figuren oder ein Mann gekennzeichnet werden müssen. Dabei entstehen häufig Zweifel, ob ein bestimmtes Detail der Zeichnung noch als richtig zu bewerten ist. Die Vorgabe von möglichst vielen Auswertungsbeispielen durch den Testautor ist dann nützlich und erforderlich.

3.1.5.2. Interpretationsobjektivität

Die Frage ob das Erreichen von 26 Punkten in einem Rechentest bei 40 möglichen eine gute oder schlechte Rechenleistung bedeutet, ist eine Frage der

Interpretation. Die Objektivität der Interpretation wird bei der Testkonstruktion durch die sogenannte Normierung der Testleistung gelöst. Jede Testleistung wird auf dem Hintergrund einer mittleren Norm interpretiert. Es geht dabei um folgende Fragen: Wie ist die durchschnittliche Leistung aller Schüler in diesem Test? Wieweit liegt ein Schüler über oder unter diesem Durchschnitt? — Erst die Differenz zwischen dem allgemeinen Mittelwert und dem individuellen Punktwert legt die Grundlage für die Interpretation einer Testleistung.

Wenn nun im Rahmen der Interpretation Vergleiche zwischen verschiedenen Testleistungen notwendig werden, steht man vor einem weiteren Problem. Ein Schüler erreicht etwa in einem Rechtschreibtest einen Punktwert von 14 bei 20 möglichen; in einem Heimatkundetest erreicht er 29 von 35 Punkten. Ist seine Heimatkundeleistung nun vergleichsweise besser als seine Rechtschreibleistung? — Es wird deutlich, daß aufgrund der sogenannten Rohpunkte, d. h. der Anzahl der richtigen Lösungen, solche Interpretationen nicht möglich sind. Auch diese Schwierigkeit kann durch die Normierung der Testergebnisse gelöst werden. Normierte Ergebnisse gestatten Vergleiche zwischen den Schülern hinsichtlich eines Tests und hinsichtlich sehr verschiedener Tests. Im nachfolgenden Abschnitt werden die gebräuchlichsten Normierungsarten bei Schultests dargestellt. Die Normierung der Testergebnisse wird dabei als Mittel betrachtet, ihre Interpretation zu objektivieren.

3.1.5.3. Normierung von Testergebnissen

Erfahrungsgemäß zeigt sich, daß Testrohpunkte meist nicht normal verteilt sind. Man kann daher nicht davon ausgehen, daß die Leistungsunterschiede von Punktwert zu Punktwert gleich sind. Hier hilft folgende Annahme: An sich verteilen sich die erfaßten Leistungen normal. Wenn bei den empirischen Rohwerten nun nicht-normale Verteilungen auftreten, so liegt dies in den Eigenarten der Aufgaben begründet, welche die Rohpunktskala verzerren. Wenn es nun gelingt, diese Verzerrung rückgängig zu machen, so ergibt sich wieder eine Normalverteilung. Dieses Rückgängigmachen der Verzerrung erreicht man durch die Transformation der Rohpunkte (auf einer Ordinalskala) in sogenannte normalverteilte Normpunkte (auf einer Intervallskala). Diese Umwandlung der Rohpunkte in Normpunkte wird mit *Normierung* (oder *Standardisierung* oder *Eichung)* eines Tests bezeichnet. Die Normierung des Tests geschieht anhand der Ergebnisse einer repräsentativen Stichprobe, der Normierungs- oder Eichstichprobe. Die Art der Normalverteilung, in die transformiert wurde, wird im allgemeinen kurz nur mit Norm des Tests benannt.

Grundsätzlich könnte jede definierte Normalverteilung zur Normierung herangezogen werden. Es ist jedoch günstig, sich nur auf wenige zu be-

schränken, da dadurch eine Einheitlichkeit der Testnormen erreicht wird, die den Vergleich zwischen verschiedenen Tests erleichtert. Aus den zur Verfügung stehenden Normierungsmöglichkeiten werden in Schultests Standardnormen, Standard-Äquivalent-Normen und Prozentrangnormen benützt. Rechnerische Verfahren zur Gewinnung der Normen werden hier nicht angegeben. Es muß auf die entsprechenden Kapitel der einschlägigen Lehrbücher verwiesen werden (siehe auch ASCHERSLEBEN 1973).

Die Standardnormen:

Die Standardnormen sind diejenigen, die auf die z-Werte der sogenannten Standardnormalverteilung zurückgehen. Die z-Werte sind zur Normierung selbst sehr unhandlich, da man wegen des Mittelwertes von $\bar{z} = 0$ und einer Streuung von $s_z = 1$ mit positiven und negativen Dezimalzahlen arbeiten muß. Durch eine lineare Transformation werden Z-Werte oder Standardwerte (SW) eingeführt. Die SW haben einen Mittelwert von $\overline{SW} = 100$ und eine Streuung von $s_{sw} = 10$. Die SW entstehen aus den z-Werten durch die Umrechnung $SW = 100 + 10\,z$. Es entsteht so eine Skala, die im allgemeinen von 70 bis 130 reicht.

Zu den Standardnormen gehören auch die Standardschulnoten (SN), die ebenfalls aus der z-Werte-Verteilung abgeleitet werden und zwar nach der Umrechnung: $SN = 3 - z$. Es ergibt sich somit eine Verteilung mit Mittelwert $\overline{SN} = 3$ und Streuung $s_{SN} = 1$.

In einzelnen Schultests werden SN angegeben, wohl unter der Annahme, daß diese Normen für Lehrer am anschaulichsten sind. Von der Verwendung ist jedoch abzuraten, da sie gerade wegen ihrer Anschaulichkeit den Lehrer verführen könnten, sie zusammen mit seinen subjektiven Schulnoten zu verarbeiten (etwa Berechnung eines allgemeinen Notendurchschnittes). Diese Praxis widerspräche auch dem Sinn einer Testanwendung in der Schule.

Die bei den Standardnormen durchgeführten Transformationen sind linear, d. h. die Verteilungsform bleibt an sich unverändert, sie wird nur gedehnt oder gepreßt und verschoben. Eine schiefe Verteilung bleibt auch nach der Transformation schief. Dies bedeutet, daß Standardnormen nur anwendbar sind, wenn die Rohpunkteverteilung schon (annähernd) normal verteilt war. Rohpunktverteilungen in Schultests sind jedoch meistens nicht normal verteilt. Standardnormen sollten daher nur angewendet werden, wenn der Testautor die Normalverteilung der Rohpunkte belegen kann.

Standard-Äquivalent-Normen:

Durch die Flächentransformation wird das Problem der Transformation nicht-normaler Verteilungen in normale Verteilungen statistisch befriedi-

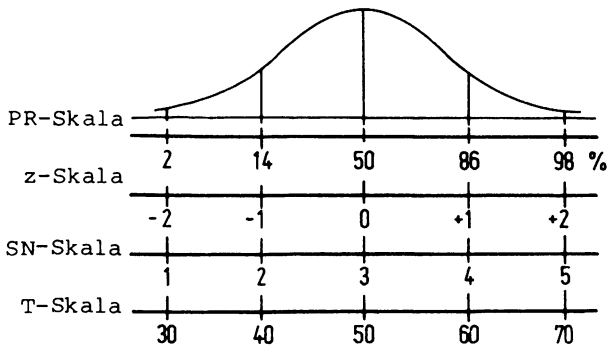
gend gelöst. Die Unterschiede zwischen den einzelnen Punktewerten werden vergrößert oder verkleinert, je nachdem ob sie sich im mittleren oder extremen Verteilungsbereich befinden. Dabei bleibt der Flächenabschnitt zwischen zwei Rohpunkten gleich groß wie der entsprechende Flächenabschnitt zwischen den transformierten Normpunkten. Die gebräuchlichste Skala dieser Art ist die T-Wert-Skala. Sie hat einen Mittelwert von $\overline{T} = 50$ und eine Streuung von $s_T = 10$. Sie wird fast in jedem Schultest angewendet.

Prozentrangnormen:

Die Prozentrangnormen sind am leichtesten zu ermitteln. Im Prinzip wird nur danach gefragt, wieviel Prozent aller Schüler eine gleich gute oder schlechtere Leistung erreicht haben als der zu beurteilende Schüler. Hat ein Schüler einen Rohpunktwert von 26 erreicht, so kann etwa festgestellt werden, daß 80 % aller Schüler der Eichstichprobe einen Punktwert zwischen Null und 26 einschließlich erreichen. Der Schüler erhält damit den Prozentrang PR = 80 (d. h. auch: nur 20 % der Schüler der Eichstichprobe erbrachten eine noch bessere Leistung). Diese Anschaulichkeit in der Interpretation ist ein Grund für die häufige Anwendung der Prozentrangnormen bei Schultests. Ein zweiter Vorteil liegt darin, daß keinerlei Voraussetzungen bezüglich der Verteilungsform der Rohpunkte gemacht werden müssen.

Die folgende Abbildung faßt die gebräuchlichsten Normen bei Schultests zusammen und gestattet Normenvergleiche:

Abb. 1: Vergleichende Darstellung von Normskalen. In Anlehnung an MICHEL (1964, S. 29). (Abkürzungen: PR = Prozentrang, SN = Standard-Schulnoten)



3.1.5.4. Kriterien zur Anwendung unterschiedlicher Normen

Falls ein Schultest mehrere unterschiedliche Normen zur Verfügung stellt, sollte sich die Auswahl der Normen nach den Bedürfnissen des Interpretierenden richten. Wenn Standardnormen angegeben werden, sollten diese nur angewendet werden, wenn die Normalverteilung der Rohpunkte nachgewiesen werden kann. Standardschulnoten dienen allenfalls einer ganz groben, eher pauschalen Orientierung und sollten nicht mit den subjektiven Schulnoten von Lehrern in Verbindung gebracht werden.

Als die günstigste Normenart ist die T-Wert-Skala anzusehen, bei der die Transformation von nicht-normalen in normale Verteilungen gelöst ist. Durch diese Transformation ist ebenso wie bei den Standardnormen, das Niveau einer Intervallskala (s. Abschn. 3.1.1.2.) erreicht. T-Werte können daher statistisch weiter verarbeitet werden, siehe z. B. die Berechnung einer mittleren Schulleistung aus einem Rechtschreib-, einem Rechen- und einem Fremdsprachentest. T-Werte sind außerdem differenziert genug, um eine sinnvolle Klassifikation der Schüler, etwa im Rahmen der Binnendifferenzierung einer Klasse, zu erlauben.

Prozentrangnormen bewegen sich auf dem Niveau einer Rang- oder Ordinalskala. Entsprechend der Darstellung in Abschn. 3.1.1.2. können sie statistisch nicht weiter verarbeitet werden. Die Differenzen zwischen den Leistungen verschiedener Schüler in verschiedenen Tests sind nicht miteinander vergleichbar. Demnach eignen sie sich nur zur groben Klassifikation. Prozentrangnormen sind deskriptiv; ihr Vorteil liegt in ihrer Anschaulichkeit. Sie empfehlen sich daher besonders bei Beratungen von Eltern, denen die Angabe eines T-Wertes recht nichtssagend vorkommen muß.

Nachzutragen wäre, daß manche Schultests noch Äquivalentnormen in Form der Verteilung des Intelligenzquotienten angeben. Aus theoretischen Gründen sind Äquivalentnormen abzulehnen (Michel 1964). Auch aus praktischen Gründen sollten Schultestergebnisse nicht in Form von Intelligenzquotienten angegeben werden. Sie suggerieren eine Identität von Schulleistung und Intelligenz, die nicht gegeben ist.

3.1.5.5. Zusammenfassung

Schultests verfügen über eine recht hohe Durchführungs- und Auswertungsobjektivität, soweit diese überhaupt zu erreichen sind. Testergebnisse können aufgrund ihrer Normen objektiv interpretiert werden. Für den praktischen Gebrauch werden T-Wert-Normen und Prozentrangnormen empfohlen.

3.1.6 Formaler Aufbau von Schultests

Bisher war immer von Tests die Rede. Jeder Test besteht nun aber aus Aufgaben (Items), die für diesen Test mehr oder weniger geeignet sein können. Die Aufgaben- oder Itemanalyse bietet nun Möglichkeiten, die Brauchbarkeit jeder einzelnen Aufgabe abzuschätzen. Zunächst soll allerdings kurz auf unterschiedliche Aufgabenarten eingegangen werden.

3.1.6.1. Aufgaben und Aufgabenanalyse

Als grobe Zweiteilung der Aufgaben bietet sich die Unterscheidung an in Aufgaben mit freien und Aufgaben mit gebundenen Antworten. *Freie* Antworten in diesem Sinne sind etwa: einen angefangenen Satz oder eine Geschichte zu Ende zu bringen, einen Aufsatz zu schreiben, ein Muster zu entwerfen oder ein Bild zu malen. Freie Antworten lassen der Vp großen Spielraum für individuelle Aussagen. Bei ihnen gibt es nicht immer ein objektives Kriterium dafür, ob eine Aufgabe richtig oder falsch gelöst wurde. Freie Antworten erfordern die Interpretation eines Fachmannes, was einen gewissen Mangel an Objektivität nach sich zieht. Aufgaben mit freien Antworten finden in fast allen Schulreifetests Verwendung.

In den Schultests werden fast ausschließlich Aufgaben mit *gebundener* Antwort verwendet. Bei Aufgaben diesen Typs ist vorher eindeutig festgelegt, was als richtig oder falsch zu werten ist. Gebundene Aufgabenantworten können so konstruiert sein, daß die richtige Lösung angegeben werden muß (z. B. bei Rechenaufgaben, oder: in welchem Jahr hat Columbus Amerika entdeckt?), daß die richtige Lösung aus mehreren herausgefunden werden muß, etwa in Form sog. Mehrfach-Wahlantworten (z. B. Der längste Fluß der Erde ist: a) Mississippi, b) Amazonas, d) Donau, e) Wolga), oder daß Antworten einander zugeordnet werden müssen (z. B. Welche Stadt gehört zu welchem Land?: Paris, Salzburg, London, München — Deutschland, Österreich, Spanien, Frankreich, USA, England). Ein Vorteil der gebundenen Antworten liegt in der objektiven Überprüfbarkeit der Ergebnisse. Für eine freie Interpretation bleibt kein Raum. Als Nachteil könnte angeführt werden, daß keine dieser Aufgaben ungewöhnliche oder kreative Bearbeitungen durch den Schüler zulassen.

Im Laufe einer Testkonstruktion muß überprüft werden, inwieweit jede einzelne Aufgabe den Zielen des Tests dienlich ist. Diese Überprüfung wird mit *Aufgaben-* oder *Itemanalyse* bezeichnet. Da, wie schon erwähnt, die gebräuchlichen Schultests vornehmlich Aufgaben mit gebundenen Antworten verwenden, kann sich die Beschreibung der Aufgabenanalyse auf diesen Aufgabentyp beschränken. Auf Schwierigkeiten, die bei freien Antworten auftreten, wird nicht eingegangen werden.

Im ersten Schritt der Aufgabenanalyse wird die Schwierigkeit jeder einzelnen Aufgabe festgestellt. Die *Aufgabenschwierigkeit* ist empirisch definiert durch den Prozentanteil P der Vpn, die diese Aufgabe richtig lösten. Eine Aufgabe mit der Schwierigkeit P = 20 ist also relativ schwierig, da sie nur von 20 % aller Vpn gelöst werden konnte; eine Aufgabe mit P = 90 ist dagegen leicht. Für einen Test sind im allgemeinen Aufgaben mit extremer Schwierigkeit (sehr leicht oder sehr schwierig) vergleichsweise nutzlos, da sie kaum Informationen über Leistungsunterschiede zwischen den Vpn bieten. Eine Aufgabe, die von allen Vpn gelöst wird (P = 100), sagt ebensowenig etwas über Unterschiede von Vpn aus wie eine Aufgabe, die so schwierig ist, daß niemand sie lösen kann (P = 0). Falls keine besondere Fragestellung für den Test vorliegt, wird man also nur Aufgaben mit mittlerer Schwierigkeit in den endgültigen Test aufnehmen.

Im nächsten Schritt wird die sogenannte *Trennschärfe* jeder Aufgabe bestimmt. Man geht dabei von der Überlegung aus, daß jede Aufgabe die Stichprobe der Vpn in zwei Gruppen trennt: Vpn mit richtiger Lösung und Vpn mit falscher Lösung. Es sollte nun nicht zufällig sein, ob eine Vp nun gerade in die eine oder andere Gruppe kommt. Vielmehr sollten in der Gruppe mit den richtigen Lösungen die „guten" Vpn sein, in der Gruppe mit den falschen Lösungen die „schlechten". Gut oder schlecht wird in diesem Zusammenhang durch das Ergebnis aus allen Aufgaben definiert; gute Vpn erreichen eine hohe Punktzahl, schlechte nur eine niedrige. Eine Aufgabe, die diese Trennung in gute und schlechte Vpn optimal leistet, wird als trennscharf bezeichnet. Als Maß der Trennschärfe dient der Trennschärfeindex. Es handelt sich dabei um die Korrelation r_{xg} der einzelnen Aufgabe g mit dem Gesamttestergebnis x. Je höher diese Korrelation ist, desto mehr gute Vpn, d. h. Vpn mit hohem Punktwert, konnten die Aufgabe g lösen. Für einen guten Test ist zu fordern, daß die Trennschärfeindizes für alle Aufgaben möglichst hoch, mindestens aber signifikant sind. Trennschärfeindizes guter Aufgaben liegen in der Regel bei 0,4 und 0,5, selten höher.

Falls für eine Aufgabe ein negativer Trennschärfeindex ermittelt wird, bedeutet dies, daß gute Vpn gerade diese Aufgabe eher falsch und schlechte Vpn eher richtig lösten. Ein negativer Trennschärfeindex kann beispielsweise auftreten, wenn die Aufgabe durch ihre Formulierung die guten Vpn auf falsche Lösungswege lockt, während die schlechten Vpn sie richtig beantworten, weil sie nicht über die Aufgabe nachdenken und sie nur rein mechanisch beantworten.

In weiteren Schritten bietet die Aufgabenanalyse noch die Möglichkeit, die Reliabilität und — falls ein geeignetes Kriterium vorhanden ist — auch Validität jeder einzelnen Aufgabe zu bestimmen. Eine nähere Beschreibung darüber sei hier erspart: Aussagen über die Reliabilität und Validität der

einzelnen Aufgaben fehlen ohnehin bei den Schultests. Eine gründliche Analyse und Kontrolle der Aufgaben ist im Rahmen einer Testentwicklung absolut notwendig, da ein Test schließlich nur so gut sein kann wie seine Aufgaben.

3.1.6.2. Untertests und Testbatterien

Tests erfassen im allgemeinen mehrere (unterschiedliche) Aspekte einer Leistung. In einem Rechentest können beispielsweise Aufgaben der mündlichen und schriftlichen Addition, Subtraktion, Multiplikation und Division enthalten sein. In diesem Falle würden die einzelnen Aufgaben zu entsprechenden Untertests, wie etwa „Addieren und Subtrahieren", „Multiplizieren" und „Dividieren" zusammengefaßt werden. Für jeden einzelnen Untertest erhielten die Schüler einen Punktwert. In einem anderen Fall könnte ein Schultest etwa bestehen aus Untertests wie: Lesen, Rechtschreiben, Rechnen und heimatkundliches Wissen.

Sowohl in theoretischer als auch in praktischer Hinsicht besteht zwischen Test und Untertest kein Unterschied: Untertests sind (kleinere) Tests im Test. Werden dagegen mehrere komplette Tests (u. U. mit einzelnen Untertests) im Rahmen einer Untersuchung zusammengefaßt, so spricht man von einer Testbatterie.

3.1.6.3. Ein praktisches Beispiel: Der AST 4

Bisher wurde die klassische Testtheorie und ihre Anwendung auf die Konstruktion von Schultests relativ abstrakt dargestellt. Um nun diese Darstellung durch ein konkretes Beispiel zu ergänzen und zu vertiefen, soll ein Schultest näher beschrieben werden. Zu diesem Zweck wurde der „Allgemeine Schulleistungstest für vierte Klassen" (AST 4) von FIPPINGER (1967 a) ausgewählt. Literatur über diesen Test liegt vor (u. a. FIPPINGER 1967 b; AMELANG & KÜHN 1972; FINGERHUT & LANGFELDT 1973).

Um den AST 4 zu charakterisieren, werden Ziele und Inhalte des Tests bzw. der Untertests aus dem Testbeiheft (FIPPINGER 1967 a) zitiert. Diesen Beschreibungen wird jeweils eine typische Aufgabe hinzugefügt. Um die Anwendung des Tests nicht zu gefährden, wurden die Aufgaben in Anlehnung an die Originalaufgaben vom Verfasser neu formuliert. Unzulänglichkeiten der Beispielaufgaben gehen daher voll zu Lasten des Verfassers und nicht des Testautors.

„Der AST 4 ... gestattet es, die Schulleistung eines Schülers — sowie der gesamten Klasse — objektiv und vergleichbar in der 2. Hälfte der 4. Klasse festzustellen. Er erfaßt nicht nur die Leistung des Schülers in einem bestimmten Fach,

sondern in allgemeiner Form alle Leistungsanforderungen, die an den Schüler einer 4. Klasse gestellt werden.

Damit erhält der Lehrer einmal einen zuverlässigen Maßstab für sein eigenes Leistungsurteil (Zensuren, Noten) und zum anderen gesicherte Hinweise für die Übertrittsauslese in weiterführende Schulen." (Testbeiheft, S. 3).

Der Test liegt in zwei Parallelformen A und B vor, wodurch eine Anwendung im Klassenraum möglich ist, ohne daß die Schüler voneinander abschreiben können. Der kombinierte Untertest für Heimatkunde (s. u. HW/KV) ist jedoch in beiden Testformen identisch. Die folgende Aufstellung gibt einen Überblick über die sieben Untertests in der Reihenfolge der Darbietung (Testbeiheft, S. 3).

LESEVERSTÄNDNIS (LV)
Paralleltest-Reliabilität: $r_{AB} = 0{,}76$; Darbietungszeit: $t = 12$ Min.

„Dieser Untertest überprüft vor allem das Sinnverständnis beim Lesen. Der Schüler soll zunächst in sich geschlossene Geschichten erlesen und anschließend Fragen beantworten, die sich auf den Inhalt der Geschichte beziehen. Der Antwort kann dann jeweils entnommen werden, ob die Geschichte in ihrem Inhalt bzw. Sinn verstanden wurde."

Umfang: Zwei Geschichten mit je 10 Aufgaben (Mehrfach-Wahlantworten)
Beispielaufgabe:
Wie alt können die Jungen in der Geschichte wohl gewesen sein?
a) 5 Jahre; b) 15 Jahre; c) 21 Jahre; d) 18 Jahre

WORTSCHATZ (WS)
$r_{AB} = 0{,}87$; $t = 8$ Min.

„In mehreren Wortgruppen soll zu einem vorgegebenen Wort aus vier Möglichkeiten jeweils die passendste (dem Wortsinne nach) herausgefunden werden."

Umfang: 30 Aufgaben mit Mehrfach-Wahlantworten
Beispielaufgabe:
Insekt paßt zu: a) Vogel; b) Hund; c) Biene; d) Goldfisch

KOPFRECHNEN (KR)
$r_{AB} = 0{,}93$; $t = 10$ Min.

„In einzelnen Operationen (Addition, Subtraktion, Ergänzen, Multiplikation, Division und einfaches Bruchrechnen), die den Lehrplananforderungen der einzelnen Bundesländer für die 4. Schuljahrgänge entsprechen, wird die Fähigkeit des Kopfrechnens überprüft."

Umfang: 18 Aufgaben mit ansteigender Schwierigkeit
Beispielaufgabe:
1) $656 + 260 = ?$ (leichte Aufgabe)
2) $\frac{9}{2} + 1{,}5 = ?$ (schwere Aufgabe)

SCHRIFTLICHES RECHNEN (SR)
$r_{AB} = 0{,}94$; $t = 15$ Min.

„Die Leistungen in schriftlichen Rechnungen werden mittels Additions-, Subtraktions-, Multiplikations- und Divisionsaufgaben erfaßt."

Umfang: 18 Aufgaben mit ansteigender Schwierigkeit
Beispielaufgaben:
1) Zuzählen:

335
421
\+ 153 (leichte Aufgabe)

2) Teilen: 354 652 : 14 = ? (schwere Aufgabe)
Im Anschluß an den Untertest SR wird eine Pause von 10 Minuten eingelegt.

TEXTAUFGABEN (TA)
$r_{AB} = 0{,}94$; $t = 20$ Min.

„Die Fähigkeit des rechnerischen Denkens wird anhand von Textaufgaben ermittelt."

Umfang: 20 Aufgaben mit ansteigender Schwierigkeit
Beispielaufgabe:
1) Zählt man zwei Zahlen zusammen, dann kommt 153 heraus; die eine Zahl heißt 65. Wie heißt die andere? (leichte Aufgabe)
2) Klaus möchte sich ein Kofferradio für 150 DM kaufen. Er hat 120 DM gespart. Er spart täglich 75 Pfennige. Wie lange muß er noch warten? (schwere Aufgabe)

RECHTSCHREIBEN (RS)
$r_{AB} = 0{,}90$; $t = 10$ Min.

„Es werden dem Schüler Sätze gegeben, die jeweils ein falsch geschriebenes Wort enthalten. Dieses soll herausgefunden und berichtigt werden. (Die Gefahr, sich ein falsch geschriebenes Wortbild einzuprägen, ist mit der Berichtigung ausgeschaltet.)"

Umfang: 20 Sätze
Beispielaufgabe:
Der Bauer pflügt den Aker

HEIMATKUNDLICHES WISSEN / KARTENVERSTÄNDNIS (HW/KV)

In diesem Untertest bestehen keine Parallelformen, es wurde daher die Halbierungsreliabilität berechnet: $r_{tt} = 0{,}86$; $t = 15$ Min.

„Dieser Untertest besteht aus zwei Teilen:

a) dem Heimatkundlichen Wissen (HW):
Unter besonderer Berücksichtigung des geographischen, historischen und biologischen Aspektes der Heimatkunde sind Wissensfragen von überregionalem Charakter zu beantworten.

b) dem Kartenverständnis (KV):
Da die Einführung ins Kartenverständnis eine wesentliche Aufgabe des Unterrichtsfaches Heimatkunde ist, soll der Schüler anhand einer Landkarte sowie eines Hausgrundrisses mit sich daran anschließenden Fragen zeigen, in welchem Maße er diese Fähigkeit besitzt."

Umfang: 20 Aufgaben mit Mehrfach-Wahlantworten zum heimatkundlichen Wissen und je 10 Aufgaben mit Mehrfach-Wahlantworten zum Kartenverständnis (Landkarte und Hausgrundriß).

Beispielaufgabe für HW:

Heidelberg liegt: a) an der Isar c) am Neckar
b) an der Ruhr d) an der Elbe

Beispielaufgabe für KV (Hausgrundriß):

Wie weit muß Herr Müller gehen, wenn er um das ganze Haus will?

a) 32 m; b) 48 m; c) 36 m; d) 44 m

Die Paralleltest-Reliabilität des Gesamttestes beträgt 0,91. Der Testautor beruft sich auf curriculare Validität (zur Übereinstimmungsvalidität mit Schulnoten siehe AMELANG & KÜHN 1972). Bei den einzelnen Aufgaben handelt es sich um gebundene Antworten.

Die drei Untertests LV, WS und RS ergeben zusammengefaßt die Deutschleistung, KR, SR und TA die Rechenleistung des Schülers.

Normen für jeden Untertest und den Gesamttest werden getrennt angegeben; und zwar für wenig gegliederte (weniger als acht Klassen) und für voll gegliederte Schulen (mindestens acht Klassen). Die angegebenen Normen beziehen sich auf Prozentränge, T-Werte (die im Testheft als Standardwerte bezeichnet werden) und Standardschulnoten.

3.1.6.4. Abschließende Definition eines Schultests

Es sei erlaubt, den Abschnitt 3.1.6. damit abzuschließen, womit Lehrbücher zur Testtheorie und Testkonstruktion üblicherweise beginnen: mit einer De-

finition vom (Schul-)Test. Hierbei wird eine Definition von LIENERT in leicht abgewandelter Form herangezogen *.

„Ein Schulleistungstest ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Schulleistungsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Schulleistung".

3.1.7. Diskussion

3.1.7.1. Kritik an der klassischen Testtheorie

Es wurde in Abschnitt 3.1.2. gezeigt, daß die klassische Testtheorie versucht, Aussagen über schon vorgegebene Meßwerte zu machen. Sie fragt nicht danach, wie diese Meßwerte zustande kamen. Auf dieses Versäumnis läßt sich die Kritik an der klassischen Testtheorie zurückführen.

Die klassische Testtheorie macht bestimmte Grundannahmen über die ihr gegebenen Meßwerte. Es ist dann selbstverständlich, daß die Theorie mit der Richtigkeit ihrer Annahmen steht oder fällt. Eines ihrer Axiome besteht in der Annahme, daß Fehlerwerte und wahre Werte unkorreliert sind. Es ist aber durchaus denkbar, daß unterschiedliche wahre Werte auch unterschiedliche Fehlerwerte bedingen.

Die Reliabilität r_{tt} wird nun definiert durch den Quotienten aus wahrer Varianz s^2_T und beobachteter Varianz s^2_X (Abschn. 3.1.2.3.). Diese Varianzen sind aber abhängig von der beobachteten Population, in unterschiedlichen Populationen treten unterschiedliche Varianzen auf. Der Reliabilitätskoeffizient r_{tt} hängt damit ebenfalls von der beobachteten Population ab. „Praktisch bedeutet dies, daß die Reliabilität nicht die Zuverlässigkeit eines Meßinstruments (Test) charakterisiert, sondern seine Zuverlässigkeit in bezug auf eine bestimmte Population" (FISCHER 1968 b, S. 66). Allgemein wäre jedoch zu fordern, daß die Zuverlässigkeit nur ein Merkmal des Instrumentes ist und nicht mit den gemessenen Objekten zusammenhängt. Die Zuverlässigkeit einer Federwaage etwa hängt auch nur von der Beschaffenheit der Waage und ihrer Einzelteile ab und nicht von den gewogenen Gegenständen. Die Zuverlässigkeit eines Schultests dagegen hängt von der Stichprobe der untersuchten Schüler ab.

---

* „Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung." (*Lienert* 1969, S. 7).

Für den Standardmeßfehler (Abschn. 3.1.3.5.) gilt das gleiche, da die Reliabilität r_{tt} in seine Berechnung eingeht. Man kann also feststellen, daß der Standardmeßfehler in verschiedenen Stichproben unterschiedlich ausfällt. Auch hier wäre zu fordern, daß Fehlerhaftigkeit einer Messung nur vom Meßinstrument und nicht auch von den gemessenen Objekten abhängt.

Die gleiche Überlegung gilt ebenso für die kriteriumsbezogene Validität eines Tests, die durch die Korrelation r_{tc} von Test- und Kriteriumswerten bestimmt wird. Wie alle Korrelationskoeffizienten ist auch die Kriteriumskorrelation eines Tests von der beobachteten Population abhängig.

Selbst das Testergebnis einer einzelnen Vp ist abhängig von den Ergebnissen ihrer Bezugsgruppe. Dies fängt bei der Aufgabenschwierigkeit an. Die Schwierigkeit einer Aufgabe variiert von Strichprobe zu Stichprobe. Eine Rechenaufgabe kann z. B. für Schulanfänger so schwierig sein, daß keiner der Schüler sie lösen kann (P = 0), andererseits ist die gleiche Aufgabe für Viertkläßler so leicht, daß alle Schüler sie lösen (P = 100).

Das Ergebnis eines Leistungstests besteht meistens aus der Anzahl der richtigen Lösungen. Diese Testrohpunkte liegen auf einer Ordinalskala und erlauben nur eine Rangordnung der Vpn. Erst auf einer Intervallskala können die Unterschiede zwischen Vpn exakt beschrieben werden. Es ist daher notwendig, die Rohpunkteverteilung in eine Normalverteilung zu transformieren. Das Ergebnis dieser Transformation ist aber wiederum abhängig von der Population der untersuchten Vpn. Der Mittelwert und die Streuung der Rohpunkte, die das Bezugssystem für die Transformation bilden, sind je nach Stichprobe unterschiedlich. Konkret kann dies bedeuten, daß eine Leistung in einer Stichprobe über dem Mittelwert und in einer anderen unter dem Mittelwert liegt. Dementsprechend gilt sie im ersten Fall als „gute Leistung" und im zweiten als „schlechte Leistung". Dieses Problem tritt bei Schultests auf, wenn die Normen z. B. für Stadt- und Landschulen getrennt berechnet werden. Schüler in Landschulen können unter Umständen eine objektiv schlechtere Rohpunkt-Testleistung erbringen als Schüler in Stadtschulen (Fippinger 1967 b). Bei der Normierung wird die Rohpunktzahl mit der Leistung einer Bezugsgruppe verglichen. Ein Schüler einer Landschule erhält dann für einen bestimmten Rohpunktwert etwa den Normpunktwert T = 70, während ein Schüler einer Stadtschule für denselben Rohpunktwert einen T-Wert unter 70 erhält.

Alle wichtigen Aussagen der klassischen Testtheorie über einen Test, wie Reliabilität, Standardmeßfehler, kriteriumsbezogene Validität, Aufgabenschwierigkeit und Testergebnis, sind also von der getesteten Population abhängig. Diese Populationsabhängigkeit ist der Hauptangriffspunkt der Kritik an der klassischen Testtheorie. Wie bei physikalischen Messungen wäre auch bei pädagogisch-psychologischen Messungen Populationsunabhängigkeit zu fordern.

Die neue Entwicklung (FISCHER 1968 c) zeigt, daß es möglich ist, Tests zu entwickeln, die im genannten Sinne populationsunabhängig sind. Diese Tests beruhen auf stochastischen Testmodellen, deren bekanntestes Beispiel das Modell von RASCH darstellt. (Eine Darstellung dieses Modells findet sich u. a. in FRICKE 1972, S. 39—53). Die Anwendung dieses Modells auf Schultests ist prinzipiell möglich, erfordert aber einen dermaßen hohen Aufwand an Forschungskapazität, „daß dieses theoretisch vorteilhafte Modell zwar in größeren Forschungsprojekten, jedoch nicht in einzelnen Schulklassen eingesetzt werden kann" (FRICKE 1972, S. 108). Weiterhin ist zu beachten, daß die Probleme der Reliabilität und Validität solcher Messungen noch nicht ausreichend gelöst sind, falls man sich nicht auf „klassische" Verfahren der Reliabilitäts- und Validitätsbestimmung zurückzieht (STELZL 1972). Die Schulleistungsdiagnostik anhand von Tests wird wohl noch eine geraume Zeit mit der klassischen Testtheorie auskommen müssen.

3.1.7.2. Kritik an der Anwendung klassischer Tests

In unserem gegenwärtigen selektiven Schulsystem wird die Schulleistung eines Schülers hauptsächlich diagnostiziert, um eine Prognose über seinen Schulerfolg stellen zu können. Es ist inzwischen zum Allgemeingut pädagogischen Wissens geworden, daß Zensuren keine Grundlage für sichere Prognosen abgeben können (INGENKAMP 1971). So entstand das Bedürfnis nach Verfahren (Tests), die bessere und gerechtere Entscheidungen ermöglichen als Schulnoten. Es wird also zu diskutieren sein, welche Folgen eine Testanwendung haben kann.

Beginnen wir mit einem Vorteil der Schultests: der Objektivität. Ein Schultest stellt im Idealfall den Leistungsstand jedes einzelnen Schülers objektiv fest. Der Test erscheint so als gerechtes und unbestechliches Instrument. Diese Objektivität ist voll zu begrüßen, wenn die Feststellung des Leistungsstandes der Schüler zur Überprüfung des Unterrichtserfolges des Lehrers eingesetzt wird, um ihm die Entscheidung zu erleichtern, ob er eine bestimmte Unterrichtseinheit weiter ausdehnen muß, sie beenden oder gar abkürzen kann. In der Regel wird die Leistung jedoch als Erfolgskontrolle des Schülers betrachtet, d. h. seine Leistung muß bewertet werden. Bei der Bewertung müssen aber die Umstände berücksichtigt werden, unter denen die Leistung erbracht wurde. Dies mag an zwei Beispielen deutlich werden:

1) Schüler können objektiv schlechte Leistungen erbringen, weil sie von einem Lehrer unterrichtet werden, der nicht in der Lage ist, sie angemessen zu fördern. Aufgrund ihrer objektiv schlechten Leistung verwehrt man ihnen etwa den Zutritt zu einer weiterführenden Schule oder erwägt vielleicht die Umschulung in eine Sonderschule für Lernbehinderte.

2) Zwei Schüler werden mit einem Schultest geprüft, um Entscheidungen über ihre zukünftige Schullaufbahn treffen zu können. Der erste Schüler stamme aus vernachlässigendem Milieu und erreiche eine Leistung knapp unter der Grenze, ab der man im allgemeinen den Besuch einer weiterführenden Schule empfiehlt. Der zweite stamme aus „gut bürgerlichem" Milieu, erhalte seit einiger Zeit gezielt Nachhilfeunterricht und erreiche einen Punktwert über der kritischen Grenze. Der „unbestechliche" Test empfiehlt nur dem zweiten Schüler den Besuch des Gymnasiums. Die Empfehlung des Tests kann bei beiden Kindern pädagogisch unangemessen sein.

Die „gerechte" Leistungsfeststellung eines Tests entspricht einem „naiven" Gerechtigkeitsbegriff, der davon ausgeht, daß gleiche Leistungen auch immer gleich zu werten sind. Um pädagogisch sinnvolle Entscheidungen treffen zu können, muß die individuelle Situation des Schülers beachtet werden. Dies kann sehr wohl dazu führen, daß gleiche Leistungen unterschiedlich bewertet werden. Das Festhalten an der Objektivität des Tests kann zu pädagogisch falschen Entscheidungen führen.

Im Rahmen der Testkonstruktion ist die Bewertung der Leistung ein Normenproblem. In Schultests werden häufig getrennte Normen für Stadt- und Landkinder gegeben, bei denen in der Regel eine niedrigere Rohpunktleistung der Landkinder ausreicht, um die gleiche Normpunktleistung zu erhalten wie die Stadtkinder. Landkinder werden durch die Normen bevorzugt. Durch die Verwendung verschiedener Bezugsgruppen wird die Objektivität des Tests bei der Bewertung eigentlich schon durchbrochen.

Schultests geben im allgemeinen repräsentative Normen für das gesamte Bundesgebiet an. Diese Repräsentativität ist jedoch zweifelhaft. Die Eichstichproben, anhand derer die Normen erstellt werden, umfassen mehr als 2 000 Schüler. Dies ist eine beachtlich hohe Zahl. Es handelt sich aber nicht um Einzelschüler, sondern um Schüler aus schätzungsweise 80 bis 100 Schulklassen. Wenn ein Test für das ganze Bundesgebiet repräsentativ sein soll, werden acht bis zehn Schulklassen (evtl. aus je vier bis fünf Stadt- bzw. Landschulen) pro Bundesland erfaßt. Diese Zahlen zeigen, daß die Repräsentativität der Schultests nicht so eindrucksvoll sein kann, wie es die Gesamtzahl aller Schüler nahelegt. So konnten denn auch FINGERHUT & LANGFELDT (1973) zeigen, daß die Normen eines Schultests für eine Stichprobe von immerhin 1 855 hessischen Grundschülern nicht angemessen waren. Für den Nachweis der Repräsentativität eines Tests sollte nicht die Anzahl aller getesteten Schüler, sondern die Anzahl der untersuchten Klassen dienen, da z. B. 30 einzelne Schüler aus 30 verschiedenen Klassen wohl eine wesentlich heterogenere Stichprobe darstellen als 30 Schüler aus einer Klasse.

Ein weiterer schwacher Punkt der Normen stellt ihre Überalterung dar. Man kann wohl davon ausgehen, daß die Schulleistung im Sinne des Tests

(zumindest regional) im Laufe der Zeit variiert. Einführung neuer Lehrbücher, Änderung der Methoden, geringfügige Lehrplanänderung, Stundenplanverschiebungen (Kurzschuljahre) und vieles mehr können Ursache solcher Schwankungen sein. Die Normen werden jedoch meist nur einmal bei der Konstruktion des Tests erstellt. Sie können notwendigerweise solche Schwankungen nicht berücksichtigen und sind dann relativ schnell überholt.

Wenden wir uns nun der Reliabilität und Validität des Schultests zu. Es ist in Abschn. 3.1.3.1. gezeigt worden, daß die Bestimmung einer Retest-Reliabilität nicht sehr sinnvoll ist, da ein Schultest nur für einen bestimmten Zeitraum im Schuljahr gültig ist. Die jeweils feststellbare Paralleltest-, Halbierungs- und Konsistenz-Reliabilität sagen nichts über die Veränderung oder die Konstanz der objektiven Testleistung aus. Dies wäre aber notwendig zu wissen, um Entscheidungen für Schulmaßnahmen für den einzelnen Schüler treffen zu können.

Immerhin können verschiedene Kriterien aufgestellt werden, an denen ein Schultest validiert werden kann. Es handelt sich dann allerdings nicht mehr um die Überprüfung derselben objektiven Leistung, wie sie durch den Test definiert wird, sondern um die Vorhersage einer Kriteriumsleistung, die mehr oder weniger mit den gemessenen Leistungen übereinstimmen kann. Allgemein ist für jeden Test eine hohe Validität zu fordern. Von einem Schultest wäre demnach zu verlangen, daß er mit einem anderen Schultest (oder einem anderen Kriterium) hoch korreliert, der einige Zeit später durchgeführt wird. Es wurde schon darauf hingewiesen, daß eine hohe Korrelation nur erreichbar ist, wenn die schlechten Schüler relativ schlecht und die guten Schüler relativ gut bleiben. Die Forderung nach einer Vorhersage-Validität widerspricht daher den pädagogischen Bemühungen, die dahin gehen sollten, alle Schüler so zu fördern, daß sie möglichst alle ein gesetztes Leistungskriterium erreichen. Die konsequente Anwendung der Reliabilitäts- und Validitätskonzepte der klassischen Testtheorie stabilisiert die schulischen Verhältnisse. So gesehen können gerade die Schultests die Chancengleichheit und das Recht aller Schüler auf optimale Förderung behindern.

Die Bestimmung der Vorhersage-Validität eines Schultests kann daher nur den Zweck haben, sie ungültig zu machen. Dies soll an einem Beispiel klargemacht werden. Einem Schüler wird aufgrund eines validen Tests eine schlechte Prognose gestellt. Nehmen wir an, das Testergebnis sage aus, daß er mit 90%iger Wahrscheinlichkeit das Klassenziel (oder ein anderes Kriterium) nicht erreichen wird. Es muß nun alles getan werden, damit diese Prognose nicht eintrifft. Falls dies bei diesem Schüler und bei allen anderen gelänge, bestünde keine Korrelation mehr zwischen Test- und Kriteriumsleistungen.

In der praktischen Schularbeit ist die Situation jedoch komplizierter. Man kann ja nicht davon ausgehen, daß alle Unterrichtsmaßnahmen

bei jedem Schüler gleich gut ansprechen. Unterricht ist vielmehr bei verschiedenen Schülern auch unterschiedlich erfolgreich. Im Idealfall müßte ein Test daher noch eine Prognose darüber abgeben, welche Förderungsmaßnahmen bei welchem Schüler am erfolgversprechendsten sind. Erforderlich sind Aussagen etwa der folgenden Art:

Falls keine besonderen Maßnahmen ergriffen werden, hat ein Schüler im Augenblick nur mit einer Wahrscheinlichkeit von 25 % Aussicht auf Erfolg (z. B. Versetzung). Betreut man ihn mit Förderungsmaßnahme A (z. B. programmierte Kurse), steigt seine Erfolgswahrscheinlichkeit auf 55 %, bei Maßnahme B (z. B. Kleingruppenunterricht durch gezielt ausgebildete Lehrer) auf 85 % und bei C (z. B. Einzelunterricht) auf 60 %. Man würde sich bei diesem Schüler für Maßnahme B entschließen. Bei anderen Schülern könnte eher Maßnahme A oder C angezeigt sein. Solche Entscheidungsstrategien, wie sie von CRONBACH & GLESER (1965) formuliert wurden (zur Einführung, siehe MICHEL & MAI 1968), sollten eigentlich von Schultests geleistet werden. Dies wird allerdings erst möglich sein, wenn Schultests laufend einer Vielzahl von Validitätsuntersuchungen unterworfen werden. Angesichts dieser Problematik kann der Rückzug der meisten Schultestautoren auf die curriculare Validität ihrer Tests nur deprimieren.

Es sollte nach dem Gesagten klargeworden sein, daß Schultests nicht kritiklos und rigide angewendet werden dürfen. Sie können eine, aber nicht die einzige Grundlage für fundiertere Entscheidungen in der Schule legen. Sie können dem Lehrer eine Entscheidungs*hilfe* sein, sie können ihm die Entscheidung selbst nicht abnehmen. Die Anwendung eines Schultests kann dem einzelnen Lehrer auch nichts von der Verantwortung abnehmen, die er für jeden einzelnen seiner Schüler trägt.

3.1.8. Zusammenfassung

In seinem Gutachten für den Deutschen Bildungsrat fordert INGENKAMP u. a. „Publikationen zur Information der Lehrer über die Anwendung und Interpretation von Schultests" (INGENKAMP 1970, S. 428). Der vorliegende Beitrag hatte die Absicht, dieser Forderung nachzukommen.

Zunächst wurde das System der klassischen Testtheorie allgemein dargestellt und erst später auf die Schultests angewendet, obwohl dadurch die Lektüre gelitten haben mag. Es konnte jedoch nicht darum gehen, die Konzepte der Schultests nur pragmatisch zu schildern.

Die klassische Testtheorie legt am Test die Gütekriterien Objektivität, Reliabilität und Validität an. Es wurde gezeigt, wie schwierig es ist, diese Kriterien im Rahmen von Schultests zu verwirklichen.

Selbst gute Tests können die Verantwortung des Lehrers für sein pädagogisches Handeln grundsätzlich nicht übernehmen. Aus dieser Sicht sind

bestimmte Abneigungen oder Befürchtungen von Lehrern gegenüber Schultests nicht gerechtfertigt. Vielmehr verlangt die Anwendung von Tests in erhöhtem Maße entscheidungssichere, verantwortungsbewußte und qualifizierte Lehrer.

3.1.9. Literaturverzeichnis

Amelang, M. u. *Kühn, R.:* Ursachen für die bei Jungen und Mädchen unterschiedlichen Korrelationen zwischen Schulnoten und Leistungstests. 28. Kongreß der Deutschen Gesellschaft für Psychologie, Saarbrücken 1972. (Kongreßbericht in Vorbereitung).

Anastasi, Anne: Psychological Testing. McMillan, London (Canada), 1969 [3].

Aschersleben, K.: Transformation und Normierung in Pädagogik und Psychologie. Psychol. in Erz. u. Unterricht, 1973, *20*, 77—88.

Beltz-Verlag: Deutsche Schultests: Informationsbroschüre und Gesamtverzeichnis. Beltz, Weinheim, 1968.

Clauss, G. u. *Ebner, H.:* Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Deutsch, Frankfurt/M., 1970.

Cronbach, L. J.: Essentials of Psychological Testing. Harper & Row, London, 1970 [3].

Cronbach, L. J. u. *Gleser, Goldine:* Psychological Tests and Personnel Decisions. University of Illinois Press. Urbana, 1965 [2].

Cronbach, L. J. u. *Meehl, P. E.:* Construct Validity in Psychological Tests. Psychol. Bull., 1955, *52*, 281—302.

Drenth, P. J. D.: Der psychologische Test. Barth, München, 1969.

Fingerhut, W. u. *Langfeldt, H. P.:* Schülermerkmale, Lehrermerkmale und ihre Beziehungen zu Schulnoten. Unveröffentl. Diplomarbeit, Marburg, 1971.

Fingerhut, W. u. *Langfeldt, H. P.:* Erfahrungen mit dem Allgemeinen Schulleistungstest für 4. Klassen (AST 4). Psychol. in Erz. u. Unterricht, 1973, *20*, 249 bis 257.

Fippinger, F.: Allgemeiner Schulleistungstest AST 4. Beltz, Weinheim, 1967 (a).

Fippinger, F.: Empirische Untersuchung zur Leistung von Schülern an voll und wenig gegliederten Schulen. Schule u. Psychol., 1967 (b), *14*, 96—103.

Fischer, G. H. (Hrsg): Psychologische Testtheorie. Huber, Bern, 1968 (a).

Fischer, G. H.: Kritik der klassischen Testtheorie, 1968 (b). In: *Fischer, G. H.* (Hrsg.) 1968 (a), 54—77.

Fischer, G. H.: Stochastische Testmodelle, 1968 (c). In: *Fischer, G. H.* (Hrsg.) 1968 (a), 78—132.

Fricke, R.: Über Meßmodelle in der Schulleistungsdiagnostik. Schwann, Düsseldorf, 1972.

Gulliksen, H.: Theory of Mental Tests. *Wiley*, New York, 1950.

Haseloff, O. W. u. *Hoffmann, H. J.:* Kleines Lehrbuch der Statistik. De Gruyter, Berlin, 1970 [4].

Heiss, R. (Hrsg.): Handbuch der Psychologie: Band 6, Psychologische Diagnostik. Hogrefe, Göttingen, 1964.

Heller, K.: Aktivierung der Bildungsreserven. Huber/Klett, Bern/Stuttgart, 1970.

Heller, K.: Intelligenzmessung. Neckar-Verlag, Villingen, 1973.

Heller, K. et al.: Planung und Auswertung empirischer Untersuchungen. Klett, Stuttgart, 1974.

Horst, P.: Messung und Vorhersage. Eine Einführung in die psychologische Testtheorie. Beltz, Weinheim, 1971.
Ingenkamp, K.: Möglichkeiten und Grenzen des Lehrerurteils und der Schultests. In *Roth, H.* (Hrsg.) 1970 [5], 407—431.
Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Beltz, Weinheim, 1971.
Kornmann, R.: Minimalisieren Schulreifetests die Zahl der Fehlentscheidungen? Z. f. Entw. Psychol. u. Päd. Psychol., 1972, *4*, 282—286.
Kornmann, R., Endrigkeit, F. u. *Sander, H.:* Sind lernbehinderte Schüler in Gruppen-Intelligenztests benachteiligt? Diagnostica, 1972, *18*, 111—121.
Lienert, G. A.: Testaufbau und Testanalyse. Beltz, Weinheim, 1969 [3].
Magnusson, D.: Testtheorie. Deuticke, Wien, 1969.
Mandl, H. u. *Krapp, A.:* Zum Problem der Punktwertgrenzen bei der Interpretation von Schulreifetestergebnissen. Z. f. Entw. Psychol. u. Päd. Psychol., 1972 (a), *4*, 140—146.
Mandl, H. u. *Krapp, A.:* Ist die Zahl selektiver Fehlentscheidungen in der pädagogischen Diagnostik von Bedeutung? Z. f. Entw. Psychol. u. Päd. Psychol., 1972 (b), *4*, 287—290.
Michel, L.: Allgemeine Grundlagen psychometrischer Tests. In: *Heiss, R.* (Hrsg.) 1964, 19—70.
Michel, L. u. *Mai, N.:* Entscheidungstheorie und Probleme der Diagnostik bei Cronbach & Gleser. Diagnostica, 1968, *14*, 99—121.
Mittenecker, E.: Planung und statistische Auswertung von Experimenten. Deuticke, Wien, 1970 [8].
Roth, H. (Hrsg.): Begabung und Lernen. Klett, Stuttgart, 1970 [5].
Sader, M. u. *Keil, W.:* Bedingungskonstanz in der psychologischen Diagnostik. Arch. Ges. Psychol., 1966, *118*, 279—308.
Sixtl, F.: Meßmethoden der Psychologie. Beltz, Weinheim, 1967.
Süllwold, F.: Schultests. In: *Heiss, R.* (Hrsg.) 1964, 352—384.
Stelzl, Ingeborg: Was bringt das *Rasch*-Modell in der Praxis? Psychol. Beitr., 1972, *15*, 298—310.
Tent, L.: Die Auslese von Schülern für weiterführende Schulen. Hogrefe, Göttingen, 1969.
Walker, Helen M.: Statistische Methoden für Psychologen und Pädagogen. Beltz, Weinheim, 1970 [10].
Wechsler, D.: Die Messung der Intelligenz Erwachsener. Textband zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE). Huber, Bern, 1956.

3.2. Einige testtheoretische Aspekte kriterienbezogener Leistungsmessung

Peter Büscher

Der kriterienbezogene Meßaspekt hat in der pädagogischen Diagnostik in jüngster Zeit sehr starke Beachtung gefunden. So plausibel dieser Ansatz ist, so schwierig sind die damit verbundenen theoretischen und praktischen Probleme.

Einige dieser Schwierigkeiten sollen im folgenden dargestellt werden.

3.2.1. Terminologische Probleme

Vor der Eröterung testtheoretischer Fragen erscheint es notwendig, die Bedeutung einiger Begriffe zu klären.

3.2.1.1. Lernzielorientierte und standardisierte Tests

In psychologischen Publikationen zum Thema ‚Leistungsmessung in der Schule' findet man gelegentlich eine Klassifizierung der Schulleistungstests in sogenannte standardisierte und lernzielorientierte Tests. Eine Abgrenzung dieser Art dürfte jedoch nicht ganz logisch sein, da alle Meßverfahren, die man unter der Kategorie ‚Schulleistungstests' zusammenfassen kann, stets in irgendeiner Weise an Lernzielen orientiert bzw. auf Lernziele bezogen sind. Der Lernzielbezug ist eine notwendige Bedingung jeder schulischen Leistungsmessung. Art des Lernzielbezugs und Grad der Lernzielspezifizierung können allerdings bei den einzelnen Meßinstrumenten recht unterschiedlich sein. Neben Tests, die sich auf eine mehr globale Überprüfung allgemeiner Lernzielbereiche beziehen, liegen Verfahren vor, die der Messung eng umschriebener, operational definierter Unterrichtsziele dienen.

3.2.1.2. Lernzielorientierte und lehrzielorientierte Tests

Von KLAUER (1972) stammt der Vorschlag, den Begriff ‚lernzielorientiert' durch ‚lehrzielorientiert' zu ersetzen. „Im folgenden wird von lehrzielorientierten Tests gesprochen, denn es handelt sich um Tests, mit deren Hilfe geprüft werden soll, ob beziehungsweise wie gut die Schüler das Lehrziel erreicht haben, d. h. das Ziel, das der Lehrer angestrebt hat ... Die Rede von Lernzielen, von lernzielorientierten Tests und ähnlichen Ausdrükken ist oft nichts anderes als eine gedankenlose, eilfertige Anpassung an Modeerscheinungen" (KLAUER 1972). Obwohl der Terminus ‚lehrzielorientiert' ohne Zweifel eine treffendere Beschreibung des Sachverhaltes darstellt,

wird man den bisher gebräuchlichen Ausdruck ‚lernzielorientiert' nicht als völlig falsch beurteilen können, da das vom Lehrenden festgelegte Lehrziel gleichzeitig auch als Ziel des im Lernenden internal ablaufenden Lernprozesses betrachtet werden kann. Dennoch erscheint es sinnvoll, der Bezeichnung ‚lehrzielorientiert' den Vorrang zu geben.

3.2.1.3. Normbezogene und kriterienbezogene Tests

Die Unterscheidung einer ‚normbezogenen Messung' von einer ‚kriterienbezogenen Messung' geht auf GLASER (1963) zurück. Sie hat ziemlich rasch Eingang in die Diskussion um die Probleme der Schulleistungsmessung gefunden und wird heute allgemein als brauchbares Konzept betrachtet.

Der Hauptunterschied zwischen einer normbezogenen und einer kriterienbezogenen Messung liegt in der Art der Information, die man mit Hilfe der Testwerte gewinnen möchte.

Interpretiert man die individuelle Testleistung im Vergleich zur Leistung anderer Individuen, dann liegt eine normbezogene Messung vor. Aufgrund der unterschiedlichen Merkmalsausprägungen, die von den einzelnen Probanden erreicht werden, kann eine Rangordnung aufgestellt und der relative Standort einer Person in bezug zu einer genau definierten Normgruppe angegeben werden. Man spricht deshalb auch von einem normativen oder relativen Standard. Die Meßskala ist hierbei im Mittelbereich verankert, d. h. im Bereich des durchschnittlichen Leistungsniveaus einer bestimmten Personengruppe. Die Skaleneinheiten sind Funktionen der Leistungsverteilung oberhalb und unterhalb dieses Durchschnittswertes. Da normbezogene Messungen nicht notwendigerweise operational definierte Lernziele implizieren, enthalten die gewonnenen Testwerte meist nur wenig Informationen über Umfang und Art des vom Schüler beherrschten Lehrstoffes.

Interpretiert man hingegen die individuelle Testleistung im Vergleich zu einem spezifischen, fest definierten Leistungsstandard (Kriterium), so liegt eine kriterienbezogene Messung vor. Nach WANG (1969) ist ein kriterienbezogener Test „... ein Leistungstest, der entwickelt wurde, um das Vorhandensein oder Fehlen eines spezifischen, durch das Lehrziel beschriebenen Kriteriumsverhaltens zu messen". Eine etwas allgemeinere Formulierung wird von GLASER (1971) vorgeschlagen: „A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." Jede individuelle Merkmalsausprägung wird also an einem festgelegten Verhaltensniveau gemessen. Man spricht deshalb auch von einem absoluten Standard. Im Gegensatz zur normbezogenen Messung ist die Meßskala hier an den Enden verankert. Das eine Skalenende zeigt die perfekte Beherrschung, das andere Ende das völlige Fehlen der betreffenden Fähigkeit an (EBEL 1971).

3.2.1.4. Kriterien

In diesem Zusammenhang ist es notwendig, auf die unterschiedlichen Bedeutungen des Begriffes ‚Kriterium' einzugehen. In der traditionellen Testtheorie bezeichnet man als Kriterium eine Variable, an der ein Meßinstrument validiert werden kann. „Die kriterienbezogene Validität wird ... durch den Bezug der Testpunktwerte zu Kriterienpunktwerten definiert. In der Regel geschieht dies mit Hilfe der Korrelation" (LIENERT 1967). So kann man beispielsweise die Gültigkeit eines Eignungstests für irgendein Schulfach mit Hilfe der Kriterien ‚Lehrerurteil', ‚Zensuren', ‚Prüfungserfolg' etc. bestimmen.

Aus der Lernpsychologie kommt der Begriff des Lernkriteriums (criterion level). Hierbei ist das Kriterium ein mehr oder minder willkürlich definierter Lernleistungsstandard (cut-off-point), der eine Klassifizierung in erfolgreiche und nichterfolgreiche Probanden ermöglicht.

Spricht man jedoch von kriterienbezogener Messung, dann versteht man unter Kriterium bestimmte fest umschriebene Verhaltensweisen, die sich der Lernende im Hinblick auf eine Lehreinheit aneignen soll. GARVIN (1971) weist darauf hin, daß es verschiedene Arten von Kriterien gibt, deren jeweils besondere pädagogische Bedeutung zu beachten ist.

So gibt es Aufgaben, die in jeder vorstellbaren Situation auf einem hohen Niveau geleistet werden müssen. Man denke etwa an die Durchführung einer Operation durch den Arzt, an das Starten oder Landen eines Flugzeuges usw. Jede Aufgabe, die die öffentliche Sicherheit betrifft gehört dazu. Hier sind kriterienbezogene Messungen angebracht. Kriterien sind dann die Zulassungsstandards bestimmter Berufe oder Tätigkeiten.

Andere Aufgaben sind dadurch gekennzeichnet, daß bei ihnen — obgleich ein Kriterienniveau ohne Schwierigkeiten definiert werden kann — eine gewisse Leistungsbreite tolerierbar ist. Es entsteht kein ernsthafter Schaden, wenn das Kriterium nicht völlig erreicht wird. Verhaltenskorrekturen können, sofern sie pädagogisch notwendig erscheinen, im allgemeinen nachgeholt werden. Eine Tätigkeit dieser Art ist beispielsweise das Kochenlernen.

Des weiteren sind dann Aufgaben zu nennen, bei denen durchaus verschiedene Leistungsniveaus akzeptabel sind. So gibt es sicherlich Berufsmöglichkeiten für Stenotypistinnen mit unterschiedlichem Schreibtempo.

Schließlich kennt man Aufgaben, die überhaupt nicht nach einem bestimmten Leistungsstandard ausgeführt werden müssen (z. B. sportliche Leistungen). Sinnvolle Kriterien lassen sich hierbei ohne weiteres definieren, ihre Bedeutung ist jedoch eingeschränkt. Nach GARVIN ist es also notwendig, sinnvolle von weniger sinnvollen Kriterien zu unterscheiden. Die in einem sequentiell geordneten Lernbereich geforderte Eingangsleistung für die nächste Lehreinheit ist ein sinnvolles Kriterium. Der von einem Lehrer willkür-

lich festgesetzte Leistungsstandard 90/90 stellt kein sinnvolles Kriterium dar. „Unless at least one of the instructional objectives of a unit envisions a task that must subsequently be performed at a specified level of competence in at least some situation, criterion-referenced measurement is irrelevant because there is no criterion" (GARVIN 1971).

Aus dieser Beschreibung der normbezogenen und kriterienbezogenen Messung lassen sich folgende Feststellungen ableiten:

— Testwerte können normbezogen oder kriterienbezogen interpretiert werden.

  Man kann die Ergebnisse eines Schulleistungstests dazu verwenden, die Schüler einer Klasse in eine Rangordnung zu bringen, um beispielsweise Zensuren zu verteilen (normbezogene Messung). Eine Leistungsdifferenzierung erfolgt dann mit Hilfe von Standardwerten oder Prozenträngen. Als Bezugsnorm bietet sich die Leistungsverteilung der entsprechenden Klassenstufe an.

  Die Ergebnisse desselben Schulleistungstests können aber auch dazu dienen, festzustellen, ob die Lernenden einem festdefinierten Leistungsstandard genügen oder nicht (kriterienbezogene Messung). Der Unterschied zwischen normbezogener und kriterienbezogener Messung ist demnach nicht so sehr in den Meßinstrumenten als vielmehr in der Meßfunktion bei pädagogischen Entscheidungen zu suchen. Dennoch bestehen zwischen beiden Testarten entscheidende Konstruktionsunterschiede. Der Konstrukteur eines normbezogenen Tests wünscht Variabilität der Testwerte. Diese Einstellung veranlaßt ihn, alles zu unternehmen, um eine Streuung der Ergebnisse zu erreichen. Er eliminiert deshalb zu leichte oder zu schwere Items, er versucht bei Mehrfachwahl-Aufgaben die Attraktivität der Antwortalternativen zu erhöhen u. ä. m. Der Konstrukteur eines kriterienbezogenen Tests hingegen wird von einem anderen Ziel geleitet. Sein Hauptinteresse besteht darin, sicherzustellen, daß die entwickelten Items das definierte Lehrziel repräsentieren, gleichgültig ob sie leicht oder schwer, trennscharf oder nicht trennscharf sind. Die unterschiedlichen Intentionen bei der Entwicklung norm- bzw. kriterienbezogener Meßinstrumente können zur Formulierung recht unterschiedlicher Testaufgaben führen.

— Es gibt ganz bestimmte Entscheidungssituationen, in denen normbezogene Messungen angebracht sind. Müssen Selektions- oder Rangordnungsentscheidungen getroffen werden, so sind normbezogene Messungen geeignet. Sollen Leistungsstärken oder -schwächen von Schülern im Hinblick auf bestimmte Lehreinheiten diagnostiziert werden, dann empfiehlt sich eine kriterienbezogene Messung.

Normbezogene Tests sind im Rahmen der schulischen Leistungsmessung wesentlich weiter verbreitet als kriterienbezogene Tests. Das mag seine

Gründe darin haben, daß sich Testtheoretiker und -praktiker bisher fast ausschließlich mit Eignungs-, Selektions- und Vorhersageproblemen beschäftigt haben. Erst in der letzten Zeit ist eine verstärkte Beschäftigung mit kriterienbezogenen Messungen festzustellen.

3.2.2. Testtheoretische Probleme

Wie bereits ausgeführt, ist einer der wichtigsten Unterschiede zwischen normbezogener und kriterienbezogener Messung im unterschiedlichen Ausmaß der Testwertevarianz zu suchen. Während normbezogene Tests notwendigerweise eine große Streuung der Testergebnisse aufweisen, muß bei kriterienbezogenen Tests mit einer mehr oder weniger eingeschränkten Variabilität der Werte gerechnet werden. In Extremfällen kann die Streuung sogar Null werden, nämlich dann, wenn alle Schüler das durch das Lehrziel definierte Kriterium erreichen.

Abb. 1: Streuung der Testwerte bei normbezogenen Tests

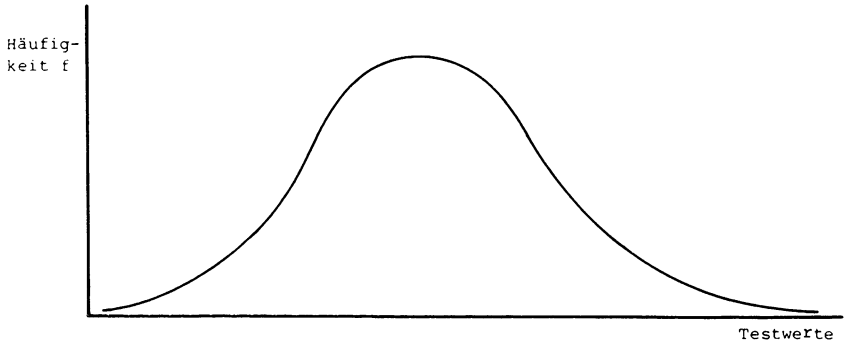
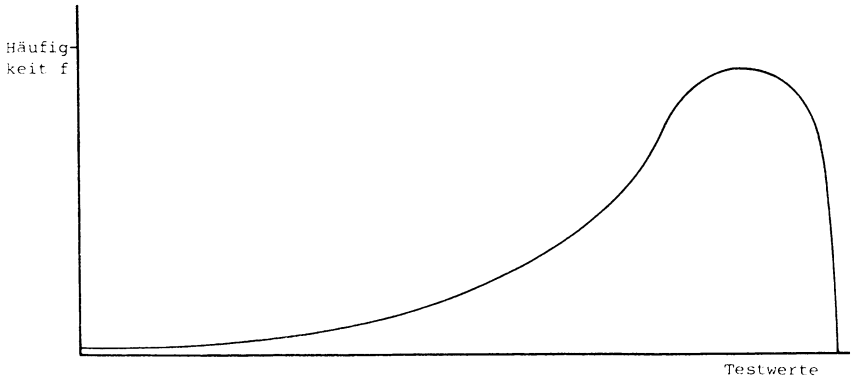


Abb. 2: Stark reduzierte Testwertestreuung bei kriterienbezogenen Tests



3.2.2.1. Unzulänglichkeit der klassischen Testtheorie bei kriterienbezogenen Tests

Die verringerte Testwertevarianz hat erhebliche Konsequenzen für einige Meßaspekte. So beruhen beispielsweise die gebräuchlichen Rechenverfahren zur Bestimmung der Reliabilität und Validität eines Tests sowie die Prozeduren der Itemanalyse fast ausschließlich auf dem Vergleich von Varianzverhältnissen (Korrelationen). Deshalb können die meisten klassischen testtheoretischen Methoden bei kriterienbezogenen Messungen kaum oder überhaupt nicht angewandt werden (POPHAM & HUSEK 1969). „Wenn man versucht, die bekannten Formeln aus der klassischen Testtheorie ... zu benutzen, wird man im Falle der lehrzielorientierten Messung Schwierigkeiten bekommen. Denn in fast allen Formeln zur Quantifizierung der ... Testgütekriterien ist der Produkt-Moment-Korrelationskoeffizient enthalten, der dann nicht definiert ist, wenn mindestens eine Variable keine Streuung mehr aufweist" (FRICKE 1972 a). Neben der reduzierten Varianz der Testwerte und den damit verbundenen meßtheoretischen Schwierigkeiten bei kriterienbezogenen Tests ergeben sich weitere Probleme durch die unterschiedlichen Konzepte bei der Testentwicklung normbezogener bzw. kriterienbezogener Meßinstrumente. Die Brauchbarkeit von Aufgaben bei normbezogenen Tests wird mit Hilfe einer sogenannten Itemanalyse überprüft. Dabei werden für jede Aufgabe bestimmte Kennwerte berechnet. Die gebräuchlichsten Itemkennwerte sind die Schwierigkeit, die Trennschärfe und die Gültigkeit.

Im folgenden soll gezeigt werden, warum die Anwendung dieser traditionellen Analysetechniken auf kriterienbezogene Testaufgaben zu unbefriedigenden Ergebnissen führen muß.

3.2.2.1.1. Analyse der Aufgabenschwierigkeit

Bekanntlich ist der Schwierigkeitsindex (P) einer Aufgabe definiert als der Prozentsatz der richtigen Lösungen. Bei der Auswahl von Items für einen normbezogenen Test wird man eine möglichst große Streuung der Schwierigkeitsindizes innerhalb des geplanten Meßbereichs anstreben. Soll beispielsweise der Test über die gesamte Meßskala differenzieren, so sind Aufgaben zu selegieren, deren Schwierigkeitskennwerte etwa zwischen $P = 20$ und $P = 80$ liegen. Bei einem Test zur Leistungsbeurteilung Hochbegabter würde man lediglich Items mit hoher und sehr hoher Schwierigkeit ($P = 5$ bis $P = 25$) auswählen.

Aufgaben, die von allen oder von keinem Probanden gelöst werden, sind für normbezogene Messungen in jedem Falle unbrauchbar und werden deshalb bei der Itemanalyse eliminiert. Ganz anders ist die Situation bei kriterienbezogenen Tests. Hier gibt der Lösungsprozentsatz keinen Hinweis dar-

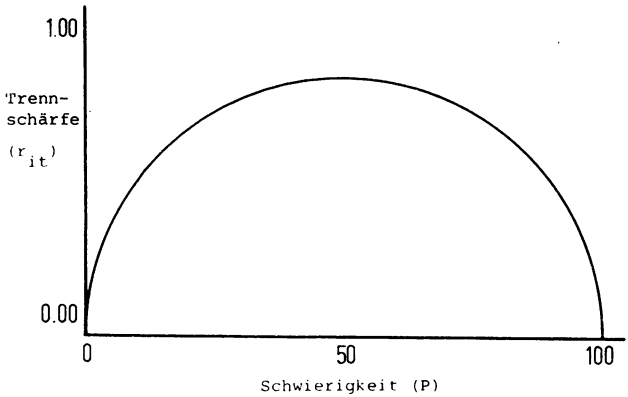
auf, ob eine Aufgabe in einem Test beibehalten werden soll oder nicht. Wenn eine Aufgabe einen relevanten Aspekt eines Lehrzieles erfaßt, dann hat sie ihre Berechtigung in einem lehrzielorientierten Test, gleichgültig ob sie nun leicht oder schwer ist. Aufgaben, die von allen oder von den meisten Schülern richtig beantwortet werden, entsprechen den Intentionen des Lehrers. Lehrzielrelevante Aufgaben mit geringen Lösungsprozentsätzen bedürfen einer besonderen Beachtung. Die Ursachen für die geringe Effektivität können verschiedener Art sein: Das Item weist formale oder inhaltliche Konstruktionsmängel auf; die geforderte Leistung wurde im Unterricht zu wenig berücksichtigt; die Bedeutung des Lehrziels war den Schülern nicht klar etc. Eine Eliminierung solcher Aufgaben wäre nicht zulässig, weil dadurch unter Umständen leicht Mängel des Lehr- und Lernprozesses bzw. der Testentwicklung verschleiert werden könnten.

3.2.2.1.2. Analyse der Aufgabentrennschärfe

Der Trennschärfeindex gibt an, wie gut ein Item zwischen Schülern mit hoher und niedriger Testleistung unterscheiden kann. Formal kann man den Trennschärfekennwert als Korrelation zwischen Itemantwort und Testrohwert definieren (r_{it}).

Eine hohe Trennschärfe liegt vor, wenn die Mehrzahl der Schüler mit niedrigem Gesamttestwert das Item falsch beantwortet, während die Mehrzahl der Schüler mit hohem Gesamttestwert die Aufgabe richtig löst. Von fehlender Trennschärfe spricht man, wenn sowohl schwache als auch gute Schüler die Aufgabe richtig bzw. falsch lösen. Schließlich gibt es noch die Möglichkeit einer negativen Trennschärfe, die dann auftritt, wenn gute Schüler bei einem Item mehr Falschlösungen aufweisen als schwache Schüler.

Abb. 3: Zusammenhang zwischen Schwierigkeit und Trennschärfe



In diesem Zusammenhang muß noch auf die Abhängigkeit zwischen Trennschärfe und Schwierigkeit einer Aufgabe hingewiesen werden. Zwischen beiden Kennwerten besteht eine ‚parabolische' Beziehung (LIENERT 1967).

Wie man aus der Abbildung erkennt, kann die Trennschärfe lediglich bei mittlerer Aufgabenschwierigkeit maximale Werte erreichen. Aufgaben, die von allen Probanden richtig bzw. falsch gelöst werden, besitzen keine Trennschärfe mehr.

Da bei kriterienbezogenen Tests häufig eine mehr oder weniger reduzierte Testwertevarianz zu beobachten ist, sind logischerweise auch nur niedrige Aufgabentrennschärfen zu erwarten. Dennoch lassen sich Trennschärfeindizes — wenn auch in anderer Form — für die Aufgaben eines kriterienbezogenen Tests bestimmen. Sie geben dann darüber Auskunft, wie gut ein Item zwischen den Schülern, die das Lehrziel erreicht haben und denen, die es nicht erreicht haben, diskriminieren kann.

3.2.2.1.3. Analyse der Aufgabengültigkeit

Die in der klassischen Testtheorie übliche Validierung von Testaufgaben an einem Außenkriterium ist bei lehrzielorientierten Tests von untergeordneter Bedeutung. Entscheidend ist hier in erster Linie die Inhaltsvalidität. Testaufgaben sind dann gültig, wenn sie eine repräsentative Verhaltensstichprobe des definierten Lehrzieles darstellen.

Trotz der Problematik des traditionellen Ansatzes werden für die Analyse lehrzielorientierter Tests in der Literatur erstaunlicherweise ausschließlich die Methoden der klassischen Testtheorie vorgeschlagen. Hier wird nicht konsequent verfahren. Man wird FRICKE zustimmen müssen, wenn er fordert: „Zu einem lehrzielorientierten Test gehört ... nicht nur eine lehrzielorientierte Testkonstruktion, sondern auch eine lehrzielorientierte Testanalyse" (FRICKE 1972 a). Die Schwierigkeiten sind sicher groß, da es noch keine gutfundierte Theorie kriterienbezogener Messungen gibt; dennoch existiert eine Reihe von interessanten Ansätzen. Einige sollen im folgenden dargestellt werden.

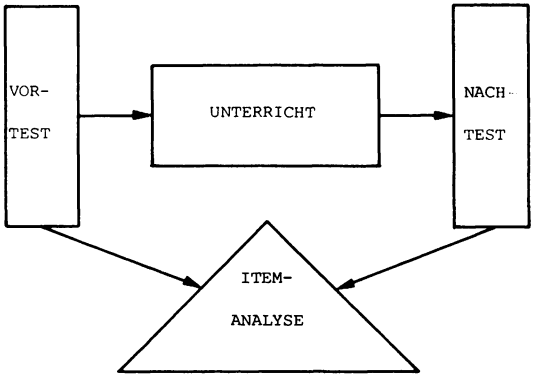
3.2.2.2. Verfahren zur Test- bzw. Itemanalyse kriterienbezogener Meßinstrumente

Die Notwendigkeit, für die Auswahl von kriterienbezogenen Testaufgaben spezielle Itemkennwerte zu ermitteln, wird allgemein anerkannt. Für die praktische Erprobung neuer Itemanalyse-Modelle haben sich vor allem COX & VARGAS, COX & GRAHAM, POPHAM und FRICKE eingesetzt.

3.2.2.2.1. Der Vortest-Nachtest-Differenzindex D_{pp} nach Cox & Vargas

Einer der ersten Ansätze zur Entwicklung eigenständiger Itemanalysetechniken bei kriterienbezogenen Tests, geht auf Cox & Vargas (1966) zurück. Ausgehend von der Erkenntnis, daß Aufgaben, die zwischen einer oberen und einer unteren Leistungsgruppe gut diskriminieren (klassische Trennschärfe), für kriterienbezogene Meßinstrumente kaum von Bedeutung sind, unternahmen sie den Versuch, das Konzept der Trennschärfe neu zu formulieren. Als mögliches Verfahren zur Bestimmung einer ‚lehrzielorientierten' Trennschärfe wird von den Autoren der Vergleich der Itemlösungen in einem Vortest-Nachtest-Design vorgeschlagen. Eine Aufgabe ist dann trennscharf, wenn sie zwischen den Vortest- und Nachtestgruppen gut unterscheiden kann.

Abb. 4: Itemanalyse nach einem Vortest- Nachtest-Design



Die Quantifizierung dieser Art von Trennschärfe erfolgt mit Hilfe des sogenannten Vortest-Nachtest-Differenzindizes D_{pp}. Er wird berechnet, indem man bei jedem Item vom Prozentsatz der Richtiglösungen in der Nachtestgruppe (P_{rn}) den Prozentsatz der Richtiglösungen in der Vortestgruppe (P_{rv}) abzieht.

$$D_{pp} = P_{rn} - P_{rv}$$

Beispiel: 25 Schüler bearbeiten einen Physiktest vor und nach der entsprechenden Unterrichtseinheit. Dabei wird die Aufgabe 3 im Vortest von 5 Schülern und im Nachtest von 20 Schülern richtig gelöst.

$$P_{rn} = 100 \cdot 20/25 = 80$$
$$P_{rv} = 100 \cdot 5/25 = 20$$
$$D_{pp} = 80 - 20 = 60$$

20 % der Schüler lösen also die Aufgabe 3 im Vortest richtig und 80 % waren im Nachtest erfolgreich. Der D_{pp}-Index ist demnach sowohl ein Maß

für die Effektivität des Unterrichts als auch ein Kennwert für die Trennschärfe der Aufgabe 3.

Um die Auswirkungen auf die Itemselektion zu demonstrieren, wurden von Cox und Vargas die D_{pp}-Werte verschiedener Schulleistungstests mit herkömmlichen Trennschärfeindizes (D) verglichen. Dabei zeigte sich, daß ein großer Teil der zu selegierenden Items bei beiden Analyseverfahren identisch war. Eine Reihe von Aufgaben, die gerade für die kriterienbezogene Messung eine besondere Bedeutung besaßen, hätten jedoch nach der traditionellen Itemanalyse eliminiert werden müssen.

3.2.2.2.2. Die Itemanalysen von Popham

Ähnlich wie Cox und Vargas sah auch Popham (1970) in den Veränderungen der Itemlösungen bei Vor- und Nachtest eine Möglichkeit der Trennschärfenbestimmung (Vier-Felder-Analyse). Des weiteren versuchte er, die Gültigkeit von kriterienbezogenen Testitems empirisch zu erfassen (Chi-Quadrat-Analyse; Ermittlung eines Prototyp-Items).

Vier-Felder-Analyse

Bei der Vier-Felder-Analyse werden die verschiedenen Itemlösungsmuster bei Vor- und Nachtest verglichen. Vier Möglichkeiten sind denkbar:

— Eine Aufgabe wird von den Schülern im Vortest falsch, im Nachtest richtig gelöst (Lösungsmuster 01);
— eine Aufgabe wird sowohl im Vortest als auch im Nachtest richtig beantwortet (Lösungsmuster 11);
— eine Aufgabe wird sowohl im Vortest als auch im Nachtest falsch gelöst (Lösungsmuster 00);
— eine Aufgabe wird im Vortest richtig, im Nachtest jedoch falsch beantwortet (Lösungsmuster 10).

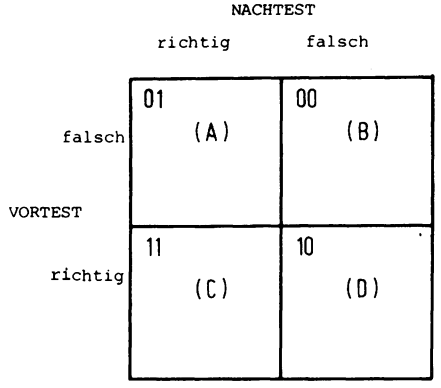


Abb. 5: Vortest-Nachtest-Lösungsmuster

Stellt man die Lösungsmuster in einer Vier-Felder-Tafel zusammen, so befinden sich in Feld A die positiven Veränderungen (01) und in Feld D die negativen Veränderungen (10). Von brauchbaren Items erwartet man nun, daß sie einen hohen Wert in der Kategorie 01 aufweisen und daß gleichzeitig die Zahl der unerwünschten 10-Fälle abnimmt, d. h. als Gütekriterium ist eine negative Korrelation der 01/10-Rangreihen wünschenswert.

Chi-Quadrat-Analyse

Ziel der Chi-Quadrat-Analyse ist die Überprüfung der Itemhomogenität. Von Testaufgaben, die dasselbe Lehrziel erfassen, erwartet man, daß sie vergleichbare Lösungsmuster haben, d. h. daß ihre Lösungshäufigkeiten annähernd gleich sind. Die Übereinstimmung der einzelnen Items kann durch einen Chi-Quadrat-Test festgestellt werden. Dazu vergleicht man die beobachteten und erwarteten Lösungshäufigkeiten miteinander.

Item-Nummer	01	00	11	10
1				
2				
3				
4				
5				
.				
.				
.				
k				

Abb. 6: Chiquadrat-Tabelle (beobachtete und erwartete Lösungsmuster)

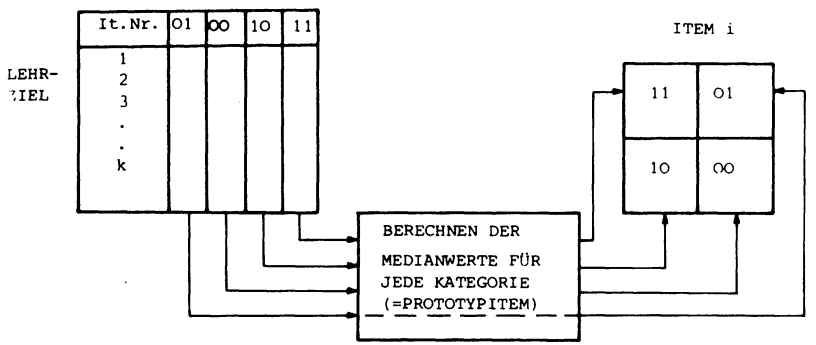
Ein signifikantes Chi-Quadrat weist auf Items hin, die von den anderen in die Analyse einbezogenen Testaufgaben stark abweichende Lösungsmuster zeigen. Diese Items erfassen unter Umständen nicht das gleiche Lehrziel. POPHAM ist der Ansicht, daß die Chi-Quadrat-Analyse als erste Grobschätzung der Aufgabenhomogenität brauchbar sei.

Prototyp-Items

Als vielversprechende Möglichkeit schlägt POPHAM schließlich noch ein Verfahren vor, bei dem es darum geht, ein für ein bestimmtes Lehrziel typisches Item, das sogenannte Prototyp-Item zu identifizieren. Das ist eine Aufgabe, die in besonderem Maße geeignet ist, das betreffende Lehrziel zu

erfassen. Die Frage ist nun, welches Item als Prototyp gelten kann. Nach POPHAM eignen sich für diesen Zweck am besten die Medianwerte der Lösungsmuster aller für dasselbe Lehrziel entwickelten Items.

Abb. 7: Bestimmung eines Prototyp-Items



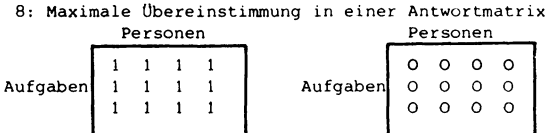
Vergleicht man jetzt die empirischen Vierfelderhäufigkeiten eines einzelnen Items mit dem Lösungsmuster des Prototyp-Items, d. h. mit den berechneten Medianwerten, so kann man mit Hilfe eines Chi-Quadrat-Tests die vom Prototyp abweichenden Items erkennen.

3.2.2.2.3. Der Ü-Koeffizient nach FRICKE

Der von FRICKE (1972 a) entwickelte Ü-Koeffizient dient der Bestimmung von Item- bzw. Testgütekriterien bei lehrzielorientierten Tests. Er hat eine dem Korrelationskoeffizienten ähnliche Funktion. Im Gegensatz zu den klassischen Verfahren der Kennwertquantifizierung ist er auch dann noch definiert, wenn keine oder eine sehr stark reduzierte Streuung der Testwerte vorliegt, d. h. wenn alle oder fast alle Schüler das Lehrziel erreicht bzw. nicht erreicht haben.

Der Übereinstimmungskoeffizient Ü gibt das Verhältnis zwischen der empirischen (beobachteten) und der maximal möglichen Übereinstimmung einer Antwortmatrix wieder. Maximale Übereinstimmung besteht bei homogenem Antwortvektor. Das ist beispielsweise der Fall, wenn ein Proband alle Testaufgaben richtig (1) oder falsch (0) beantwortet.

Abb. 8: Maximale Übereinstimmung in einer Antwortmatrix



$$\ddot{U} = \frac{\text{empirische Übereinstimmung } (D_{emp})}{\text{maximale Übereinstimmung } (D_{max})}$$

Die praktische Berechnung von Ü erfolgt über die Formel

$$\ddot{U} = 1 - \frac{4\,(k\Sigma x - \Sigma x^2)}{nk^2}$$

Dabei ist

n = die Anzahl der Testteilnehmer
k = die Anzahl der Items
x = der positive Wert (1) in der Matrix
(z. B. die richtige Aufgabenlösung)

Mit Hilfe einer nach Chi-Quadrat verteilten Prüfgröße können dann die berechneten Ü-Werte statistisch abgesichert werden.

$$\chi^2 = \frac{4n}{k\,(n-1)} \cdot (k\Sigma x - \Sigma x^2); \qquad df = n\,(k-1)$$

Der Ü-Koeffizient läßt sich also überall dort einsetzen, wo bei klassischen Verfahren eine Korrelationsberechnung nötig wäre; etwa bei der Berechnung der Trennschärfe oder bei der Bestimmung der Reliabilität bzw. Validität.

3.2.2.2.4. Das Rasch-Modell

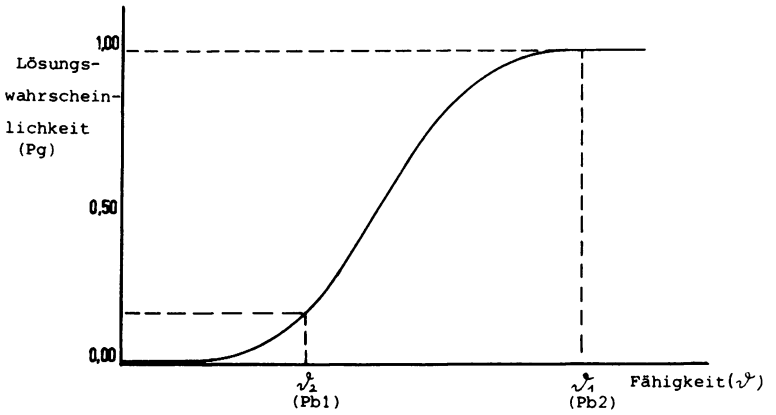
Die Möglichkeit, das logistische Testmodell von G. Rasch (1960) auf die Diagnostik von Schulleistungen zu übertragen, wurde in letzter Zeit von einigen Testtheoretikern und -praktikern untersucht (Wendeler 1968; Fricke 1972 b u. a.). Ziel des von Rasch entwickelten Ansatzes ist die Bestimmung der Aufgabenschwierigkeit unabhängig von den zufällig untersuchten Personen und die Abschätzung der Fähigkeit der Person unabhängig von der zufällig ausgewählten Itemstichprobe (spezifische Objektivität). Wright (1967) spricht in diesem Zusammenhang von einer ‚personfreien Testeichung' und einer ‚itemfreien Personmessung'. In einem komplizierten Algorithmus (maximum-likelihood-Verfahren) werden die Personen- und Aufgabenparameter geschätzt. Die Fähigkeit einer Person läßt sich dabei aus der Anzahl der von ihr richtig gelösten Aufgaben ermitteln (Summenwerte als erschöpfende Statistiken), während die Schwierigkeit einer Aufgabe über den Lösungsprozentsatz festgestellt wird. Die Grundgleichung des Rasch-Modells

$$P_{g+} = \frac{X_a}{X_a + D_i}$$

gibt die Wahrscheinlichkeit (P) einer richtigen Antwort auf Item g (g+) wieder. Wie man erkennt, ist diese Wahrscheinlichkeit lediglich von zwei Parametern abhängig, von X_a, der Fähigkeit der Person, gerade diese Art von Aufgaben lösen zu können und von D_i, der Aufgabenschwierigkeit.

Für den Zusammenhang zwischen der Lösungswahrscheinlichkeit einer Aufgabe und der zu messenden Fähigkeit (ϑ) wird eine ganz bestimmte Funktion, die kumulierte logistische Verteilungsfunktion, postuliert. Die Funktion hat folgendes Bild:

Abb. 9: Die Itemcharakteristikkurve (ICC)



Angenommen ein Proband (Pb 1) besäße die betreffende Fähigkeit in der Ausprägung ϑ_1. Entsprechend der abgebildeten Funktion (in der Regel ‚Itemcharakteristikkurve' genannt) ist die Wahrscheinlichkeit, daß er das Item richtig löst, ziemlich gering. Anders im Falle des Probanden 2. Die Fähigkeit ist bei ihm so stark ausgeprägt, daß die Wahrscheinlichkeit für eine korrekte Itemlösung annähernd 1.00 ist.

In einem nach RASCH skalierten Test müssen alle Items diese logistischen Itemcharakteristikkurven aufweisen. Aufgaben mit abweichenden Funktionen können mit Hilfe sogenannter Modelltests identifiziert werden. Da bei der Itemanalyse demnach nur Aufgaben mit identischen Charakteristikfunktionen selegiert werden, entsteht ein völlig homogener, d. h. eindimensionaler Test.

Die praktische Anwendung des RASCH-Modells auf lehrzielorientierte Tests ist problematisch, da zum einen eine große Anzahl von Probanden erforderlich ist, zum anderen der Konstruktions- und Rechenaufwand enorm ist und schließlich das Modell nicht mehr angewandt werden kann, wenn alle Probanden das Lehrziel erreicht (bzw. nicht erreicht) haben. Das

bedeutet, wenn überhaupt, dann ist das RASCH-Modell nur für normbezogene Messung sinnvoll.

Neben der Entwicklung spezieller Verfahren zur kriterienbezogenen Itemanalyse war es erforderlich, auch die entsprechenden Methoden der Testanalyse neu zu formulieren. Genau wie bei den herkömmlichen normbezogenen Tests wünscht man von einem kriterienbezogenen Test, daß er das, was er mißt, genau mißt (Reliabilität, Zuverlässigkeit, Meßgenauigkeit) und daß er das, was er messen soll, auch wirklich mißt (Validität, Gültigkeit). Einige Möglichkeiten der Testanalyse bei kriterienbezogenen Tests sollen kurz besprochen werden.

3.2.2.2.5. Reliabilitätsbestimmung nach CARVER

Der entscheidende Aspekt der Reliabilität ist nach CARVER (1970) die Wiederholbarkeit der Meßwerte. Diese Wiederholbarkeit ist aber nun keineswegs von der Testwertevarianz abhängig. Es ist durchaus möglich, daß die Ergebnisse von Paralleltests nahezu identisch sind, aber dennoch kaum Streuung aufweisen. In einem solchen Fall läge eine fast perfekte Wiederholbarkeit vor, die klassische Reliabilitätsbestimmung würde aber, da sie auf der Berechnung von Korrelationen beruht, völlig versagen. CARVER schlägt deshalb für kriterienbezogene Messungen folgende Prozedur vor: Man analysiert die Testwerte entweder von gleichwertigen Schülergruppen oder von entsprechenden Paralleltestformen. Die Reliabilität läßt sich dann dadurch abschätzen, daß man in den einzelnen Gruppen bzw. Testformen die Prozentsätze der Schüler, die das Kriterium erreicht haben, miteinander vergleicht. Je besser diese Vergleichswerte übereinstimmen, desto reliabler dürfte der Test sein.

3.2.2.2.6. Reliabilitätsbestimmung nach LIVINGSTON

In Anlehnung an die klassischen Verfahren der Testtheorie hat LIVINGSTON (1972) ein Reliabilitätsmodell entwickelt, das dadurch gekennzeichnet ist, daß nicht die Varianz um das arithmetische Mittel, sondern die Varianz um den Kriteriumswert Grundlage der Zuverlässigkeitsbestimmung darstellt. Die Berechnung eines Korrelationskoeffizienten erfolgt hierbei analog dem Produkt-Moment-Verfahren. Dieser Ansatz ist umstritten (HARRIS 1972; SHAVELSON, BLOCK, RAVITCH 1972; MERKENS 1972; HAMBLETON & NOVICK 1972). Nach HAMBLETON & NOVICK ist es kaum von Bedeutung zu wissen, wie weit ein Schüler von einem definierten Leistungsstandard entfernt ist; vielmehr ist es das Problem zu entscheiden, ob die wahre Schülerleistung oberhalb oder unterhalb dieses Kriteriums liegt.

MERKENS weist darauf hin, daß die Anwendung eines der Produkt-Moment-Korrelation ähnlichen Koeffizienten Normalverteilung der Testwerte um den Kriteriumswert voraussetze. Das bedeute aber nichts anderes, als daß in diesem Fall das Kriterium mit dem arithmetischen Mittel zusammenfalle. Dann könne man aber auch gleich die klassischen Verfahren anwenden. Vollends ungeklärt ist darüber hinaus noch die Verwendung des LIVINGSTON-schen Koeffizienten bei fehlender oder stark reduzierter Varianz, was dann auftritt, wenn alle oder die meisten Probanden den Kriteriumswert erreicht haben.

3.2.2.2.7. Reliabilitätsschätzung durch Skalogrammanalyse

Gesamttestwerte, die als Lösungssummen verschieden schwieriger Aufgaben gewonnen werden, können nur bedingt Informationen darüber vermitteln, welche Verhaltensweisen vom Probanden beherrscht oder nicht beherrscht werden. Um Auskunft darüber zu erhalten, wäre eine Überprüfung der Leistung bei jedem einzelnen Item notwendig. Eine Möglichkeit, dieses Problem zu lösen besteht in der Entwicklung sequentiell-skalierter Meßinstrumente. Bei einem so konstruierten Test ordnet man einem Lehrziel entsprechende Testitems nach aufsteigendem Schwierigkeitsgrad zu. Im Idealfall kann man dann aus der schwierigsten, vom Probanden noch gelösten Aufgabe den Grad der Lehrzielerreichung abschätzen, da der Gesamtwert gleichzeitig Auskunft über das Lösungsmuster des Probanden gibt. Der Zusammenhang der nach Schwierigkeit geordneten Rangreihe der Items und der Rangreihe der Probanden ist der Ausgangspunkt für die sogenannte Skalogrammanalyse nach GUTTMAN (1950). Die so gewonnenen Analysedaten ermöglichen neben einer Homogenitätsprüfung von Testaufgaben auch eine Abschätzung der Testreliabilität. Den praktischen Versuch, einen sequentiell-skalierten Rechentest zu entwickeln und mit den Möglichkeiten einer Skalogrammanalyse zu überprüfen, haben COX & GRAHAM (1966) unternommen. Es gelang ihnen für einen relativ einfachen und elementaren Bereich des Zahlenrechnens eine eindimensionale GUTTMAN-Skala zu konstruieren. Es ist jedoch fraglich, ob Ähnliches auch bei komplexeren Lehrzielen erreichbar ist. Von einigen Meßtheoretikern wird angenommen, daß GUTTMAN-Skalen lediglich für trivale Fälle entwickelt werden können.

3.2.2.2.8. Reliabilitätsschätzung nach CRONBACH

Betrachtet man kriterienbezogene Tests als Zufallsstichprobe aus einer ganzen Familie verwandter Tests (Testuniversum), dann kann man zur Reliabilitätsbestimmung CRONBACH's Generalisierungstheorie (CRONBACH et al.

1963) heranziehen. Eine praktische Anwendung dieser Theorie auf eine Gruppe von Arithmetik-Tests wird von HIVELY, PATTERSON und PAGE (1968) beschrieben.

3.2.2.2.9. Reliabilitätsbestimmung nach JACKSON

Von JACKSON (1970) stammt der Vorschlag, Parallelformen kriterienbezogener Tests unabhängig voneinander, jedoch nach identischen Konstruktionsvorschriften zu entwickeln und die Ergebnisse beider Formen miteinander zu vergleichen. Es wird ein sogenannter ‚index of agreement' (möglicherweise ein Kontingenzkoeffizient) zur Reliabilitätsberechnung empfohlen.

3.2.2.2.10. Reliabilitätsbestimmung nach FRICKE

Da der Ü-Koeffizient von FRICKE (1972) — wie bereits ausgeführt — eine dem traditionellen Korrelationskoeffizienten vergleichbare Funktion besitzt, ist er natürlich auch für die Bestimmung der Testgütekriterien, Reliabilität und Validität, geeignet. So lassen sich mit seiner Hilfe Retest-, Paralleltest- und auch Interitemkonsistenzmaße berechnen.

3.2.2.2.11. Klassische Reliabilitätsschätzung

Auch wenn man damit rechnen muß, daß bei kriterienbezogenen Tests eine mehr oder weniger starke Verringerung der Testwertevarianz auftritt, macht man in der Praxis die Erfahrung, daß im allgemeinen auch diese Meßwerte variieren. In diesen Fällen können dann ohne weiteres die gebräuchlichen Reliabilitätsmethoden (z. B. Bestimmung der internen Konsistenz) angewandt werden (JACKSON 1970).

3.2.2.2.12. Validitätsbestimmung

Das Hauptproblem bei der Entwicklung kriterienbezogener Meßverfahren ist die Überprüfung der Validität. Sie hängt in erster Linie von der Übereinstimmung der Testitems mit den definierten Lehrzielen ab. Die Messung muß Informationen über die Leistung des Schülers in bezug auf ein durch das Lehrziel festgelegtes Kriterium liefern. Die Effektivität kriterienbezogener Tests ist deshalb nur dann gewährleistet, wenn es gelingt, objektivierte Prozeduren, d. h. algorithmen-ähnliche Verfahren zu entwickeln, die eine standardisierte Itemkonstruktion ermöglichen. Darüber hinaus ist eine Weiterentwicklung empirischer Validierungsverfahren, wie z. B. der Konstruktvalidierung, erforderlich.

Ansätze zur standardisierten Itementwicklung finden sich bei OSBURN (1968), SHOEMAKER & OSBURN (1969), GUTTMAN & SCHLESINGER (1966), BORMUTH (1970).

OSBURN schlägt vor, sogenannte ‚Item-Muster' (item form) zu verwenden. Sie bestehen aus einem festen syntaktischen Gerüst und einem (bzw. mehreren) variablen Teil(en). Durch die Spezifizierung der die variablen Teile ersetzenden Elementenmenge wird eine Klasse von Items definiert. SHOEMAKER und OSBURN haben auf der Grundlage solcher Item-Muster Computer-Programme zur Erzeugung von Testaufgaben entwickelt. Die in das Programm eingegebenen Muster haben etwa die Form:

„Gegeben ist eine Normalverteilung mit dem Mittelwert von ..... und der Standardabweichung von ..... Wie groß ist die Wahrscheinlichkeit, daß ein zufällig aus der Verteilung ausgewählter Wert größer oder gleich ..... ist?" (zit. nach JACKSON 1970).

Ein Zufallsgenerator erzeugt hierzu im Rahmen eines für die Aufgabe realistischen Bereiches Werte für die Lücken im Item-Muster. Selbstverständlich sind auch Item-Muster mit verbalen Substitutionen möglich.

Ein ähnliches Verfahren — facet-design-method — stammt von GUTTMAN (GUTTMAN & SCHLESINGER 1966). Auch hier wird mit einem festen Itemgerüst und variablen Stellen gearbeitet. Die Art der Aufgabenerzeugung ermöglicht die systematische Entwicklung von Distraktoren bei Mehrfachwahlaufgaben. Dazu ein einfaches Beispiel:

„Ein Frischling ist ein (eine) $\underline{\quad x \quad}$ $\underline{\quad y \quad}$ "

Setzt man für die Variable x die Begriffe ‚groß', ‚klein', ‚jung', ‚alt' (x_1 = großer, kleiner, ...; x_2 = große, kleine, ...; x_3 = großes, kleines, ...) und für y die Begriffe ‚Hund' (y_1), ‚Esel' (y_1), ‚Katze' (y_2), ‚Kuh' (y_2), ‚Wildschwein' (y_3), ‚Pferd' (y_3), so lassen sich daraus eine ziemlich große Anzahl von Antwortalternativen zusammenstellen.

Schließlich soll noch auf den interessanten Ansatz von BORMUTH (1970) hingewiesen werden. Sein Ausgangspunkt ist der im Unterricht verwendete Lehrtext. Die Testitems werden mit Hilfe grammatischer und syntaktischer Regeln aus diesem Lehrtext abgeleitet. BORMUTH's Ziel ist die operational definierte, von der jeweiligen Sprachbegabung des Testautors unabhängige Aufgabe.

Alle die genannten Verfahren zur Standardisierung des Itemkonstruktionsprozesses sind problematisch. Sie sind in ihrer Anwendung meist sehr umständlich und sie versagen im allgemeinen bei komplexen Sachverhalten. Aus diesem Grund war man auch stets bemüht, zusätzlich Möglichkeiten zur empirischen Validierung kriterienbezogener Tests zu finden. Erste Schritte in dieser Richtung sind alle diejenigen Analyseverfahren, die die Homogenität

von Testitems, die ein und dasselbe Lehrziel repräsentieren, zu bestimmen versuchen (POPHAM 1970; COX & GRAHAM 1966; FRICKE 1972 a). Neuerdings werden von HERBIG (1973) zwei Verfahren zur empirischen Validierung lehrzielorientierter Tests vorgeschlagen. Die erste Methode basiert dabei auf dem Binomial-Modell (vgl. KLAUER 1972), die zweite stellt einen varianzanalytischen Ansatz dar.

3.2.3. Zusammenfassung

Es konnte gezeigt werden, daß das Konzept der kriterienbezogenen Messung voller Probleme steckt.

Neben terminologischen Unklarheiten wurden testtheoretische Schwierigkeiten diskutiert. Dabei ergaben sich folgende Schlußfolgerungen:

— Alle Schulleistungstests, seien es normbezogene oder kriterienbezogene, sind lehrzielbezogen.

— Ob eine Messung normbezogen oder kriterienbezogen klassifiziert werden kann, entscheidet die pädagogische Funktion. Dennoch gibt es entscheidende Unterschiede in der Konstruktion normbezogener und kriterienbezogener Meßverfahren.

— Die Tatsache, daß bei einem Test die Items auf ein Lehrziel bezogen sind, berechtigt noch nicht, von einem kriterienbezogenen Test zu sprechen. Zusätzliche Bedingung ist in jedem Falle die Standardisierung des Konstruktionsprozesses (d. h. Operationalisierung der Lernziele, explizite Kriterienbestimmung, Itemableitungsregeln etc.).

— Es ist notwendig, zwischen sinnvollen und weniger sinnvollen Kriterien zu unterscheiden.

— Wenn sich das Konzept der kriterienbezogenen Messung in der pädagogischen Praxis durchsetzen soll, dann müssen bestimmte Meßaspekte (Konzepte der Test- und Aufgabenanalyse) neu durchdacht werden. Es wurden einige Lösungsversuche beschrieben. Teils lehnten sie sich eng an klassische Verfahren normbezogener Messungen an, teils wurden neue Ansätze entwickelt. Insgesamt läßt sich aber sagen, daß es sich hierbei eher um erste Ansätze als um gut fundierte theoretische Modelle handelt.

— Kriterienbezogene Messungen haben genau wie normbezogene ihre eigene bedeutsame pädagogische Funktion. Keineswegs wird das eine Meßverfahren das andere verdrängen.

Die Diskussion um kriterienbezogene Messung wird in den kommenden Jahren sicherlich in verstärktem Maße einsetzen. Dabei wird sich zeigen, ob die pessimistische Beurteilung der Situation durch EBEL (1971) gerechtfertigt ist oder nicht. Seiner Meinung nach sind einwandfreie kriterienbezogene Messungen nur sehr schwer zu erhalten. Sie erfordern einen Grad an Spezifizierung der Lehrziele bzw. der entsprechenden Verhaltensaspekte,

welcher in der Praxis unrealistisch ist. Des weiteren sind kriterienbezogene Messungen nach EBEL nur sinnvoll in jenen meist elementaren Leistungsbereichen, in denen es auf einen hohen Grad von Fertigkeit in einer begrenzten Anzahl von Fähigkeiten ankommt. Im komplexen Bereich der Kenntnisse und des Verständnisses sei die Effektivität des kriterienbezogenen Ansatzes weniger wahrscheinlich. J. H. BLOCK (1971), der sich mit EBEL's Argumenten auseinandersetzte, kann diese Meinung jedenfalls nicht teilen. Für ihn und für viele andere hat die kriterienbezogene Messung eine ganz entscheidende Funktion im Lernprozeß. „... we cannot continue to use measurements which simply assess the outcomes of the teaching-learning process when measurements exist which can not only assess, but positively shape these outcomes for each learner. The potential of criterionreferenced measurements lies in their ability to promote the learning of all" (BLOCK 1971).

3.2.4. Literaturverzeichnis

Block, J. H.: Criterion-referenced Measurements: Potential. School Review, 1971, 289—298.

Bormuth, J. R.: On the theory of achievement test items. The University of Chicago Press, Chicago 1970.

Carver, R. P.: Special. problems in measuring change with psychometric devices. Evaluative Research: Strategies and Methods. Pittsburgh: American Institutes für Research, 1970.

Cox, R. u. *Graham, G. T.:* The development of a sequentially scaled achievement test. Journal of Educational Measurement, 1966, *3*, 147—150.

Cox, R. C. u. *Vargas, J. S.:* A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Ill., 1966.

Cronbach, L. J., Rajaratnam, N. u. *Gleser, G. C.:* Theory of generalizability: Aliberalization of reliability theory. British Journal of Statistical Psychology, 1963, *16*, 137—163.

Ebel, R. L.: Criterion-referenced Measurements: Limitations. School Review, 1971, 282—288.

Fricke, R.: Testgütekriterien bei lehrzielorientierten Tests (Ein Maß zur Bestimmung von Objektivität, Zuverlässigkeit, Gültigkeit und Trennschärfe bei lehrzielorientierten Tests). Zeitschrift für erziehungswissenschaftliche Forschung (ZeF), 1972 a, 150—175.

Fricke, R.: Über Meßmodelle in der Schulleistungsdiagnostik. Düsseldorf 1972 b.

Garvin, A. D.: The Applicability of Criterion-Referenced Measurement by Content Area and Level. In: *Popham, W. J.* (Ed.): Criterion-Referenced Measurement (An Introduction). Educational Technology Publications, Englewood Cliffs, N. J., 1971.

Glaser, R.: Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 519—521.

Glaser, R.: A Criterion-Referenced Test. In: *Popham, W. J.* (Ed.): Criterion-Referenced Measurement (An Introduction). Educational Technology Publications, Englewood Cliffs, N. J., 1971.

Guttman, L.: The basis for scalogram analysis. In: *Stouffer, S. A.*, et al.: Studies in social psychology in World War II, Vol. IV. Princeton, N. J., 1950.

Guttman, L. u. *Schlesinger, I. M.:* Development of diagnostic analytical and mechanical ability tests through facet design and analysis. The Israel Institute of Applied Social Research, Jerusalem 1966.

Hambleton, R. K. u. *Novick, M. R.:* Toward an integration of theory and method for criterion-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago 1972.

Harris, C. W.: An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 27—29.

Herbig, M.: Verfahren zur experimentellen Validierung lehrzielorientierter Tests Zeitschrift für erziehungswissenschaftliche Forschung (ZeF), 1973 (erscheint demnächst).

Hively, W., II et al.: A „univers defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 275—290.

Jackson, R.: Developing Criterion-Referenced Tests. TM Reports Number 1, ERIC Clearinghouse on Tests, Measurement & Evaluation, Princeton, N. J., 1970.

Klauer, K. J. et al.: Lehrzielorientierte Tests. Düsseldorf 1972.

Lienert, G. A.: Testaufbau und Testanalyse. Weinheim und Berlin 1967.

Livingston, S. A.: Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 13—25.

Merkens, H.: Zum Problem der Konstruktion von lehrzielorientierten Tests. Schule u. Psychol. 1972, 139—140.

Osburn, H. G.: Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 95—104.

Popham, W. J.: Indices of adequacy for criterion-referenced test items. A Symposium Presentation at a Joint Session of the National Council for Educational Measurement and the American Educational Research Association, Minnesota 1970.

Popham, W. J. (Ed.): Criterion-Referenced Measurement (An Introduction). Educational Technology Publications, Englewood Cliffs, N. J., 1971.

Popham, W. J. u. *Husek, T. R.:* Implications of criterion-referenced Measurement. Journal of Educational Measurement, 1969, 1—9.

Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks paedagogiske Institut, Copenhagen 1960.

Rasch, G.: An item analysis which takes individual differences into account. The British Journal of Mathematical and Statistical Psychology, 1966, 49—57.

Shavelson, R. J., Block, J. H. u. *Ravitch, M. M.:* Criterion-referenced testing: comments on reliability. Journal of Educational Measurement, 1972, 133—137.

Shoemaker, D. M. u. *Osburn, H. G.:* Computer-aided item sampling for achievement testing. Educational and Psychological Measurement, 1969, 165—172.

Wang, M. C.: Approaches to the validation of learning hierarchies. Western Regional Conference on Testing Problems (Proceedings) 1969, Princeton, N. J.: Educational Testing Service, 14—38.

Wendeler, J.: Eine Aufgabenanalyse anhand des Testmodells von Rasch. Archiv für die gesamte Psychologie, 1968, 218—230.

Wright, B. D.: Sample-free test calibration and person measurement. In: Invitational Conference on testing problems, Princeton 1967.

3.3. Zur Problematik der Klassifikation von Schultests

Bernhard Rosemann

Die verbreitete Unzufriedenheit mit den traditionellen Zensierungsmethoden hat die Pädagogen veranlaßt, nach Verfahren zu suchen, die geeignet sind, die Mängel des herkömmlichen Benotungssystems auszuschalten. Dieses Bemühen hat dem Test Eingang in unsere Schulen verschafft. LIENERT (1967) definiert *Test* als ein „wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung". Er unterteilt die Tests nach der Art des zu erfassenden Persönlichkeitsmerkmals in Intelligenztests, Leistungstests und Persönlichkeitstests. Obwohl alle drei Testarten in der Schule Verwendung finden können, wollen wir uns nur mit den *Leistungstests* befassen, da vor allem sie für das in diesem Buch behandelte Thema von Bedeutung sind. Ziel unserer Betrachtung soll sein, durch einen kritischen Überblick über die derzeitige Terminologie im Bereich der Schultests zu einer Präzisierung und Abgrenzung der in diesem Zusammenhang verwendeten Begriffe zu gelangen, wobei wir uns auf die gängigsten Konzepte beschränken werden.

3.3.1. Untersuchung der bisherigen Terminologie

3.3.1.1. Standardisierte und nichtstandardisierte Tests

LIENERT (1967) unterscheidet nach dem Allgemeinheitsgrad ihrer Anwendbarkeit zwischen standardisierten (geeichten) und nichtstandardisierten (informellen) Tests.

Er weist darauf hin, daß *standardisierte Tests* wissenschaftlich entwickelt werden, hinsichtlich der wichtigsten Gütekriterien untersucht und unter Standardbedingungen durchführbar und normiert sein müssen. Standardisierte Tests werden also von Experten nach den traditionellen Prinzipien der Testkonstruktion (z. B. LIENERT 1967, GULLIKSEN 1950) entwickelt, ein sowohl zeit- als auch kostenintensives Vorgehen. Beispiele für solche Tests finden sich etwa bei INGENKAMP (1962). GAUDE & TESCHNER (1971) verstehen unter einem *informellen Test* „ein von einem Lehrer oder einer Gruppe von Lehrern entwickeltes Prüfverfahren zur Erfassung bestimmter, möglichst exakt umschriebener Unterrichtsinhalte". Informelle Tests im bisher üblichen Sinne unterscheiden sich zunächst einmal nicht prinzipiell von den standardisierten Tests. An diese informellen Tests (auch

„classroom-tests", „teacher-made tests", „Standardarbeiten" usw. genannt) werden lediglich bescheidenere Anforderungen hinsichtlich der Konstruktion gestellt, da dem Lehrer weder die Zeit noch das notwendige Geld zur Konstruktion standardisierter Tests zur Verfügung stehen. Informelle Tests werden demzufolge nach vereinfachten Regeln der klassischen Testtheorie konstruiert, berücksichtigen dabei aber, wie z. B. BIGLMAIER (1971) vermerkt, weitgehend die Kriterien eines standardisierten Tests wie Objektivität, Gültigkeit (= Validität), Zuverlässigkeit (= Reliabilität), Normen für einen begrenzten Bereich und ökonomische Durchführung (vgl. Kap. 3.1. u. 3.2. in diesem Band). Gemeinsamkeiten und Unterschiede zwischen standardisierten und informellen Tests lassen sich vorläufig wie folgt darstellen:

	Standardisierte Tests	Informelle Tests
Objektivität der Durchführung, Auswertung und Interpretation	+	+
Reliabilität	in der Regel hoch zwischen .80 und .95 Retest bzw. Paralleltestreliabilität	in der Regel niedriger Reliabilität nach Konsistenzanalyse
Validität	+	+ -inhaltliche-
Anwendungsbereich	mehrere Klassenstufen einer oder mehrerer Schulen	meist für eine Klassen- oder Jahrgangsstufe
Testinhalt	relativ allgem. Unterr.ziele und -inhalte	auf spezielle Unterr.ziele und -inhalte bezogen
Normen	für versch. Klassen, Schulen und Schularten	für best. Klassenstufen einer Schule
Konstruktion	durch Experten	durch Lehrer

Aufgrund unserer obigen Ausführungen können wir feststellen: Die wesentlichen Unterschiede zwischen einem standardisierten Schulleistungstest und einem informellen Test liegen in den zu messenden Lernzielen bzw. Unterrichtsinhalten, der Population für die der Test anwendbar ist und in den Konstruktionsmerkmalen. Standardisierte und informelle Tests in diesem Sinne sind, um den Begriff vorwegzunehmen, „normbezogen".

Außer der obigen Einteilung finden sich weitere Unterscheidungen der in Schulen verwendeten Leistungstests. Diese Tatsache kommt zum Ausdruck durch Termini wie normbezogene, kriteriumsbezogene, lernzielorientierte oder lehrzielorientierte Tests, Kriteriumstests, mastery tests u. ä.

Bevor wir hierauf näher eingehen, sollen zunächst einige Betrachtungen über einen in diesem Zusammenhang wichtigen Begriff vorausgeschickt werden.

3.3.1.2. Der Begriff des Kriteriums

Der Begriff des Kriteriums hat seinen festen Platz im Bereich der klassischen Testkonstruktion, und zwar im Rahmen der Testvalidierung. Nach EBEL (1965) ist ein Kriterium ein Urteilsmaßstab (standard of judging). In der Testentwicklung meint man damit gewöhnlich ein Merkmal oder eine Merkmalskombination als Grundlage zur Beurteilung der Validität eines Tests oder eines anderen Meßverfahrens. Das Kriterium in diesem Sinne ist also ein Hilfsmittel, um über die Güte eines Tests befinden zu können, oder genauer gesagt, um überprüfen zu können, ob der Test das mißt, was er messen soll. (Auf eine eingehende Beschreibung der verschiedenen Kriteriumsarten können wir hier verzichten, sie findet sich ausführlich bei LIENERT 1967; vgl. ferner Kap. 3.1.4.2. sowie 3.2.1.4. oben).

Der Begriff Kriterium wird in Verbindung mit Schultests auch in anderer Weise verwendet. Dabei meint man dann mit Kriterium einen Leistungsstandard, eine absolut gesetzte Leistungsnorm. Hier wird also ein bestimmtes Leistungsniveau (Kriterium) festgesetzt, das ein Individuum erreichen kann oder nicht. Die Leistung einer Einzelperson wird mit diesem Kriterium verglichen, d. h. hier dient das gesetzte Kriterium zur Entscheidung darüber, wie „gut" eine Person ist.

Der Begriff „Kriterium" wird also im Kontext der Konstruktion von Schultests mindestens in zweierlei Weise verwendet:

Erstens bezeichnet das Wort Kriterium ein Merkmal oder eine Merkmalskombination, mit deren Hilfe man entscheiden kann, wie valide ein Test ist.

Zweitens meint Kriterium einen Leistungsstandard, anhand dessen man entscheiden kann, ob ein Lernziel erreicht wurde oder nicht, ob also eine bestimmte Fertigkeit bei einer Person vorliegt oder nicht. Allerdings besteht noch keine Übereinstimmung darüber, ob das Kriterium zu einer Alternativ-

entscheidung führen — Lernziel erreicht oder nicht erreicht —, oder ob die Leistung des Individuums auf einem Kontinuum lokalisiert und damit der Grad, in dem das Lernziel erreicht wurde, erfaßt werden sollte (s. GLASER 1963; BÜSCHER, Kap. 3.2. in diesem Buch).

Wir müssen diese mehrfache Verwendung des Begriffes Kriterium bei der Diskussion der oben angeführten Konzepte im Auge behalten.

3.3.1.3. Normbezogene versus lernzielorientierte Tests

Häufig wird zwischen normbezogenen und lernzielorientierten Tests unterschieden. Wie auch BÜSCHER (s. Kap. 3.2.) darauf hinweist, ist diese Unterscheidung nicht zwingend. Auch ein normbezogener Test, sei es nun ein standardisierter oder ein informeller Test, bezieht sich, wenn er sorgfältig konstruiert wurde, auf Lernziele; anderenfalls wäre sein Einsatz schlechterdings sinnlos. Auch INGENKAMP (1971) äußert: „Die sogenannten ‚normbezogenen Schulleistungstests' sind ja nicht vorwiegend auf Normen bezogen, sondern weit mehr lehrplanorientiert."

Wie wir in unserem zweiten Beitrag (s. Kap. 4.2. in diesem Buch) ausführlicher dargestellt haben, werden die Items auch normbezogener (informeller) Tests aus den vorher operational definierten Lernzielen abgeleitet — s. Spezifikationstabelle. Das gleiche geschieht bei den sogenannten „lernzielorientierten" Tests. Unterschiede können bestehen im Allgemeinheitsgrad der gemessenen Lernziele, sie bestehen in der weiteren Verarbeitung der erstellten Items. Bei normbezogenen Tests werden die Leistungen des einzelnen Schülers auf den Items mit den Leistungen der anderen Schüler seiner Bezugsgruppe verglichen, bei lernzielorientierten Tests mit einem gruppenunabhängigen, absolut gesetzten Leistungsstandard. Unter bestimmten Voraussetzungen sind die Items aus normbezogenen und lernzielorientierten Tests austauschbar (s. auch BALDWIN 1971). Die Unterscheidung zwischen normbezogenen und lernzielorientierten Tests ist also keine eindeutige, Lernzielbezogenheit ist für beide Testarten anzusetzen.

3.3.1.4. Normbezogene versus kriteriumsbezogene Tests

Normbezogene Tests können, wie erwähnt, voll standardisierte Tests (hier gelten die Normen für einen weiten Bereich) oder informelle Tests (hier gelten die Normen nur für eine Klasse oder Jahrgangsstufe einer Schule) sein. Es ist dabei allerdings zu fragen, ob der Begriff „Norm" bei den informellen Tests nicht zu anspruchsvoll ist. Wie auch immer, ausschlaggebendes Merkmal der normbezogenen Tests ist, daß die Leistung eines Individuums zu den Leistungen anderer in bezug gesetzt wird, d. h. die Leistung wird immer relativ zu den Leistungen anderer bewertet.

„Ein kriterienbezogener Test ist ein Leistungstest, der entwickelt wurde, um das Vorhandensein oder Fehlen eines spezifischen Kriteriumsverhaltens zu messen, wie es im Lernziel beschrieben wurde" (WANG 1969; zit. n. BÜSCHER, Kap. 3.2.). Die Leistung eines Schülers wird also mit einem gesetzten Leistungsstandard verglichen.

Der Begriff „kriteriumsbezogener" Test kann aber nun zu Mißverständnissen führen. Verwenden wir den Begriff Kriterium im ersten Sinne, also als einen Maßstab für die Validität eines Tests, dann sind selbstverständlich auch normbezogene Tests kriteriumsbezogen. Im Falle des Schulleistungstests sind „die Testaufgaben selbst das bestmögliche Kriterium für das zu untersuchende Persönlichkeitsmerkmal" — inhaltliche Validität (LIENERT 1967). Sollen die Meßwerte als Prediktorvariablen verwendet werden, so tritt allerdings noch ein Außenkriterium hinzu. Benutzen wir den Begriff Kriterium im Sinne eines Leistungsstandards, den eine Person erreichen muß, dann hindert uns nichts daran, normbezogene Tests „kriteriumsbezogen" zu verwenden. Es ist ohne weiteres möglich, einen bestimmten Leistungsstandard als Kriterium anzusetzen, dessen Erreichen für die getestete Person von einer definierten Bedeutung ist. Wir können z. B. festlegen, daß eine Person einen Wert x in einem normbezogenen Test erreichen muß, um in den nächsthöheren Unterrichtskurs aufsteigen zu können.

Fassen wir die Diskussion noch einmal zusammen:

a) Ein Kriterium kann sein
   (1) ein Merkmal oder eine Merkmalskombination zur Überprüfung der Validität eines Tests,
   (2) ein Leistungsstandard, eine absolut gesetzte Leistungsnorm, mit dem eine individuelle Leistung verglichen wird.
b) Die Unterscheidung „normbezogener Test" versus „lernzielorientierter Test" ist nicht eindeutig, auch normbezogene Tests sind lernzielorientiert.
c) Die Unterscheidung „normbezogener Test" versus „kriteriumsbezogener Test" ist problematisch, denn auch normbezogene Tests sind kriteriumsbezogen, wenn Kriterium im Sinne a1) verwendet wird.
d) Diese Unterscheidung ist auch dann nicht zwingend, wenn Kriterium im Sinne a2) aufgefaßt wird, denn auch normbezogene Tests können zu einer kriteriumsbezogenen Messung herangezogen werden.

Wo liegen dann die zweifellos vorhandenen Unterschiede zwischen den „normbezogenen" und den „nicht-normbezogenen" Tests? Sie liegen einmal, wie auch BÜSCHER darstellt, in der Vorgehensweise bei der Testkonstruktion, wobei man grob so differenzieren kann: Bei *normbezogenen* (standardisierten oder informellen) Tests werden die Regeln der klassischen Testkonstruktion mehr oder minder vollständig angewendet, insbesondere bezüglich der Aufgabenanalyse, der Reliabilitäts- und Validitätskontrolle. Bei *nicht-normbezogenen* Tests haben diese Regeln — nicht aber die ent-

sprechenden Gütekriterien! — nur sehr eingeschränkte Bedeutung. Zum anderen ist bei normbezogenen Tests die Bezugsgruppe des Individuums Bewertungsmaßstab für seine Leistung, bei nicht-normbezogenen ein absolut gesetzter Leistungsstandard, eine „absolute Leistungsnorm".

3.3.2. Gedanken zu einer pädagogisch begründeten Klassifikation der Schultests

Die bisher übliche *terminologische* Trennung zwischen normbezogenen und nicht-normbezogenen Tests gründet sich hauptsächlich auf testtheoretischen Überlegungen. Sie macht dem Benutzer solcher Verfahren aber nicht ihre vorhandene unterschiedliche pädagogische Bedeutung hinreichend klar. Obwohl auch BÜSCHER auf die unterschiedlichen pädagogischen Intentionen der diskutierten Verfahren hinweist, bleibt er bei der bisherigen Terminologie stehen. Ziel unserer folgenden Überlegungen ist eine Gruppierung der in der Schule verwendeten Tests nach ihrem pädagogischen Stellenwert und nicht nach testtheoretischen Überlegungen. Wir werden dabei die Darstellungen von BLOOM et al. (1971) berücksichtigen.

3.3.2.1. Die Unterscheidung zwischen Leistungsfeststellung und Leistungsbewertung

Grundlegend für den folgenden Gedankengang ist unsere Unterscheidung zwischen „Leistungsfeststellung" und „Leistungsbewertung". Im ersteren Falle verschafft man sich lediglich Information darüber, was die Schüler im Verlaufe des Unterrichtsgeschehens gelernt bzw. nicht gelernt haben. Diese Informationen per se kann der Lehrer in vielfältiger Weise verwenden. Im zweiten Falle geht man einen Schritt weiter, man will die festgestellten Leistungen der Schüler bewerten, wobei verschiedene Bezugspunkte für die Bewertung in Betracht kommen können. Welche Bedeutung hat diese Unterscheidung für die Testanwendung bzw. für eine Klassifikation der Tests in der Schule?

Beginnen wir mit der *Leistungsbewertung*. Bei der Leistungsbewertung ist es das Hauptziel des Lehrers, für den Schüler einen Meßwert zu erhalten bzw. eine Note festzulegen, die dessen Leistungsstand charakterisiert. Nach Abschluß einer größeren Unterrichtseinheit wird also beim Schüler ein Endverhalten evoziert, anhand dessen der Lehrer die Effizienz des Lehr- und Lernprozesses beurteilen kann. Lehr- und Lernprozeß einerseits und Bewertungsprozeß andererseits sind damit voneinander getrennt. Der Schüler lernt über eine Zeitspanne einen bestimmten Stoff, nach Abschluß eines kleineren oder größeren Stoffgebietes wird das, was der Schüler gelernt hat, gemessen und bewertet.

Der moderne Lehrer wird für diesen Bewertungsprozeß einen Test heranziehen. Er könnte einen standardisierten Test benutzen. In der Regel wird er aber darauf verzichten, weil die damit abgeprüften Lernziele zu allgemein und seinem speziellen Unterricht zu wenig angemessen sind. Deshalb wird er einen informellen normbezogenen Test, den er selbst oder ein anderer Lehrer für die entsprechende Klassenstufe und das zur Frage stehende Unterrichtsgebiet dieser Schule konstruiert hat, verwenden. Wie wir oben hervorgehoben haben, ist auch ein solcher Test lernzielorientiert (s. Spezifikationstabelle). Allerdings liegt das Schwergewicht eines solchen Tests in der Festlegung der relativen Position des Schülers innerhalb seiner Bezugsgruppe. Bei Verwendung multipler Scores (bezüglich der Verhaltens-Inhalts-Kombinationen der Spezifikationstabelle) können jedoch auch hiermit Stärken oder Schwächen des Schülers hinsichtlich bestimmter Lernziele oder Lernzielbereiche sichtbar gemacht werden. Grundsätzlich ist aber der Bezugspunkt für die Bewertung der Schülerleistung im Falle des informellen normbezogenen wie auch des standardisierten Tests die Bezugsgruppe dieses Schülers.

Der Lehrer kann aber zur Gewinnung einer Note für die Leistung eines Schülers auch einen „kriteriumsbezogenen" Test einsetzen, der bekanntlich den Vorzug hat, die Leistung des Schülers nicht an der Leistung seiner Mitschüler zu messen, sondern direkt an einem für ein bestimmtes Unterrichtsgebiet gesetzten Leistungsstandard. Es wird also jeweils geprüft, ob ein Schüler das Lernziel erreicht oder nicht erreicht hat. Wie Weis (1971) formuliert, „kann man zum Zweck des Überganges von erreichten Lernzielen zu Noten zwei Modelle anwenden: ein sequentielles oder ein additives Modell. Das sequentielle Modell ist anwendbar in hierarchisch gegliederten Gegenstandsbereichen, in denen einzelne Lernziele klar aufeinander aufbauen, so daß ein höheres Ziel nicht erreicht werden kann, wenn nicht alle vorausgegangenen Teillernziele erreicht worden sind. In diesem Fall bestimmt sich die Note aus der Höhe des erreichten Niveaus. Wendet man dagegen das additive Modell an, dann ergibt sich die Note aus dem Prozentsatz der erreichten Lernziele, wobei bestimmte Ziele, etwa als Kernstoff, obligatorisch gemacht werden können" (s. ferner auch Wendeler 1969).

Um also zu einer Bewertung der Leistung seiner Schüler zu gelangen, kann der Lehrer anwenden: standardisierte Tests, informelle normbezogene Tests, kriteriumsorientierte Tests mit Benotung. Im Hinblick auf das Lernziel stellt diese Aufzählung eine Rangreihe zunehmender Spezifität und Unterrichtsbezogenheit der abgeprüften Lernziele dar.

Welche Bedeutung hat nun die sogenannte *Leistungsfeststellung?* Sie unterscheidet sich von der Leistungsbewertung entscheidend durch die damit verfolgte pädagogische Intention. Leistungsfeststellung soll *nicht* zu einer Bewertung der Schülerleistung im Sinne etwa einer Note führen, sondern sie dient allein der Steuerung eines noch in Gang befindlichen Lernprozesses

und ist damit Teil dieses Lernprozesses. Leistungsfeststellung als integraler Bestandteil des Lehr- und Lernprozesses ermöglicht durch einen ständigen feedback-Prozeß zwischen Lehrern und Schülern ein optimales Erreichen der gesetzten Lernziele. Die Leistungsfeststellung geschieht nicht erst nach Abschluß eines größeren Stoffgebietes, sondern erfordert die Aufgliederung des Gesamtstoffes in sachlogisch begründbare Teilabschnitte oder Lerneinheiten. Die Bewältigung der in diesen Lerneinheiten gestellten Lernziele ist Gegenstand der Leistungsfeststellung. Noch während des Lernprozesses wird ermittelt, welche Schritte auf dem Weg zum Lernziel der Schüler nicht bewältigt hat, welche Hilfen ihm gegeben werden müssen, ob und wie Lehrmethoden geändert werden sollen. Die einzig relevante Aussage in diesem Zusammenhang ist die, ob ein Schüler einen Lernschritt, ein Lernelement bewältigt hat oder nicht. Wann dies der Fall ist, wird anhand eines vorher festzulegenden Kriteriums entschieden. Die Ableitung des Kriteriums erfolgt sinnvoll aus der jeweils vorgefundenen Struktur des Lernstoffes; allerdings ist das Problem der optimalen Setzung des Kriteriums unseres Wissens noch nicht hinreichend gelöst. Leistungsfeststellung erfolgt in erster Linie mit Hilfe „kriteriumsbezogener" Tests *ohne* anschließende Benotung des Schülers, denn ein die Leistung eines Schülers zusammenfassend beschreibender Meßwert (Note) hat in diesem Zusammenhang kaum einen Nutzen.

Prinzipiell ist es aber auch möglich, etwa informelle normbezogene Tests zur Leistungsfeststellung heranzuziehen, wenn diese sich auf genügend kleine Lerneinheiten beziehen, wobei die Auswertung allerdings multiple Scores unter Berücksichtigung der Spezifikationstabelle beinhalten müßte.

3.3.2.2. Lernsteuerungstests und Lernkontrolltests

Die auf der Grundlage der Unterscheidung von Leistungsfeststellung und Leistungsbewertung vorzunehmende Klassifikation von Tests orientiert sich an der Stellung der Tests im pädagogischen Gesamtgeschehen. Tests sind zum einen Hilfsmittel zur Steuerung des ablaufenden Lerngeschehens, zum anderen Instrumente zur Bewertung der Ergebnisse eines abgeschlossenen Lernprozesses. Wir wollen daher, um diese zweifache Funktion von Tests im Bereich der Schule deutlich zu machen, folgende Terminologie vorschlagen:

Verfahren, die zur *Leistungsfeststellung* mit dem Ziel der Beeinflussung des Lernprozesses dienen, nennen wir *Lernsteuerungstests*. Dazu sind in der bisherigen Benennung zu rechnen: (informelle) kriteriumsbezogene Tests ohne nachfolgende Benotung, seltener normbezogene Tests für kleinere Lerneinheiten mit multiplen Scores.

Verfahren, die zur *Leistungsbeurteilung* (Leistungsbewertung) dienen, nennen wir *Lernkontrolltests*. Dazu rechnen: standardisierte Tests, infor-

melle normbezogene Tests, (informelle) kriteriumsbezogene Tests mit nachfolgender Benotung.

Lernsteuerungstests finden ihren Sinn, durch einen permanenten Informationsaustausch zwischen Lernenden und Lehrenden zu einer optimalen Unterrichtsgestaltung und damit zu einem möglichst weitgehenden Erreichen der gesetzten Lernziele beizutragen.

Lernkontrolltests dienen zu einer relativen oder absoluten Bewertung der Leistung eines Schülers und haben ihren entscheidenden Vorteil in der Objektivierung des Bewertungsvorganges in z. T. engem Bezug zu den angestrebten Lernzielen. Ob eine Leistungsbewertung in dieser Form überhaupt noch erforderlich ist, ist ein Problem, das an dieser Stelle nicht diskutiert werden kann.

Im Hinblick auf die Testanwendung in den Schulen scheint uns die vorgeschlagene Terminologie insofern von Vorteil zu sein, als sie es auch dem in der Testtheorie noch nicht ausgebildeten Lehrer möglich macht, auf Anhieb zu erkennen, für welche pädagogischen Zwecke ein bestimmter Test einsetzbar ist. In der von uns gewählten Terminologie werden zugleich zwei entscheidende Aufgaben des Pädagogen angesprochen, die Lenkung und Steuerung des Lernprozesses und die Bewertung der Ergebnisse dieses Vorganges.

3.3.3. Literaturverzeichnis

Baldwin, T. S.: Evaluation of learning in industrial education. In: *Bloom* et al. 1971, 854—905.

Biglmaier, F.: Leistungsmessung durch informelle Lehrertests. In: *Lichtenstein-Rother, Ilse:* Schulleistung u. Leistungsschule. Bad Heilbrunn/Obb. 1971.

Bloom, B. S.; Hastings, J. Th. & *Madaus, G. F.:* Handbook on formative and summative evaluation of student learning. New York 1971.

Ebel, R. L.: Measuring educational achievement. Englewood Cliffs, Prentice-Hall, 1965.

Gaude, P. & *Teschner, W. P.:* Objektivierte Leistungsmessung in der Schule. Frankfurt am Main 1971.

Glaser, R.: Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, *18*, 519—521.

Gulliksen, H.: Theory of mental tests. New York 1950.

Ingenkamp, K.: Die deutschen Schulleistungstests. Weinheim 1962.

Ingenkamp, K.: Tests in der Schulpraxis. Weinheim 1971.

Lienert, G.: Testaufbau und Testanalyse. Weinheim 1967 (2. Aufl.).

Weis, V.: Zensierungsmodelle und ihre pädagogische Konsequenzen. Die Deutsche Schule, 1971, 542—553.

Wendeler, J.: Standardarbeiten. Weinheim 1969.

4. Objektive Verfahren der Leistungsbeurteilung in der Schule

Einleitender Kommentar

Im folgenden werden nun die wichtigsten Methoden pädagogischer Leistungsbeurteilung dargestellt. Diese lassen sich in zwei große Gruppen zusammenfassen: in ‚objektive' und in ‚subjektive' Verfahren. Zu den *objektiven* Verfahren rechnen wir praktisch alle *Test*verfahren, also die formellen (standardisierten) und informellen (kriterienbezogenen) Tests, die nach der ROSEMANNschen Terminologie wiederum in Lernsteuerungs- und Lernkontrolltests unterschieden werden können. Unter dem Gesichtspunkt der Leistungs*beurteilung* i. e. S. interessieren hier vor allem die Lernkontrolltests. Zu den *subjektiven* Verfahren schulischer Leistungsbeurteilung zählen alle Formen von Lehrerurteilen, z. B. Verhaltensbeobachtung, Aufsatzbeurteilung sowie jedwede Art von Notengebung traditioneller Prägung. Diese Verfahren werden im letzten (fünften) Hauptkapitel dieses Buches behandelt. Zunächst wenden wir uns den objektiven Verfahren zu.

Wenn auch in der standardisierten Schulleistungsmessung bislang zu wenig beachtet, so stellt die Forderung nach exakter Lernzieldefinition keineswegs ein Spezifikum sog. kriterienbezogener Leistungsmessung dar. Es ist freilich das unbestreitbare Verdienst der neuen Testentwicklung, die Aufmerksamkeit gerade auf diesen Punkt der Unterrichtsplanung gelenkt zu haben. Sowohl der Unterrichtserfolg als auch entsprechende Effizienzkontrollen (z. B. durch sog. lehr- oder lernzielorientierte Tests; indirekt freilich auch durch die sog. standardisierten Schulleistungstests) hängt ja in hohem Maße von exakten Lehr-/Lernzieldefinitionen ab. Logischerweise steht deshalb der Beitrag von R. HORN zum Thema „Leistungsmessung und Lernzieldefinition" am Anfang dieses Hauptkapitels. Die *Operationalisierung von Lernzielen* sowie die damit in Zusammenhang stehende Formulierung von Prüfungsaufgaben sind Hauptgegenstand seiner Ausführungen. Eigene Präzisierungsvorschläge zum BLOOMschen Taxonomie-Modell (vgl. Lernzielanalyseschema auf S. 175 unten) sowie eine Reihe in der Praxis erprobter Unterrichtsbeispiele werden m. E. die Arbeit des Lehrers in wesentlichen Punkten der Unterrichtsplanung und -kontrolle erleichtern.

Der folgende Beitrag von ROSEMANN befaßt sich mit der Erstellung sog. informeller Schulleistungstests. Diese werden hier vorab unter dem Aspekt der Lernleistungs*kontrolle* abgehandelt (zur Problematik der Konstruktion von Lern*steuerungs*tests vgl. Kap. 3.2. oben). Erklärte Absicht des Autors ist es, „den Lehrer, der nicht gedenkt, sich eingehender mit Theorie und Praxis der Testkonstruktion zu befassen, ... in die Lage (zu) versetzen, den einen oder anderen Test selbst zu entwickeln bzw. bestehende Tests kritisch zu beurteilen". Darüber hinaus werden Probleme in bezug auf die Anwendung

informeller Tests sowie Methoden der Testauswertung besprochen. Abgesehen von ihren meßtheoretischen Qualitäten (Objektivität, Reliabilität und Validität) sind *informelle Tests* vielleicht wie kein anderes Meßinstrument geeignet, dem Lehrer nicht nur die schwierige Aufgabe der Leistungsbewertung zu erleichtern, sondern auch die Notengebung auf eine gerechtere (weil lernzielabhängigere und somit praxisnähere, d. h. der konkreten Unterrichtssituation angemessenere) Urteilsbasis zu stellen. Der Einsatz informeller Tests als pädagogisches Hilfsmittel zur Beurteilung von Schülerleistungen sollte deshalb weit mehr als bisher an die Stelle der üblichen Klassenarbeiten treten.

Der nächste Beitrag von R. Horn gibt einen kurzen Überblick über die wichtigsten zur Zeit in der BRD lieferbaren *standardisierten* Schulleistungstests. Neben Lesetests, Rechtschreibtests und Rechentests sowie allgemeinen, d. h. fächerübergreifenden (sog. Omnibusverfahren) Schulleistungstests werden auch Fremdsprachentests und Tests für spezielle Unterrichtsfächer (Erdkunde, Geschichte, Naturlehre) beschrieben. Die übersichtlich gegliederte Darstellung mit zahlreichen Testaufgabenbeispielen vermittelt auch dem testungeübten Leser eine gute Orientierung über Aufbau und Anwendungsmöglichkeiten standardisierter Schulleistungstests, die im Rahmen schulischer Leistungsbeurteilung auch in Zukunft unentbehrlich sein werden. Ein nach Testkategorien und Klassenstufen geordnetes Übersichtsschema sowie das Verzeichnis einschlägiger Testverlage am Ende des Beitrags informieren schnell und nahezu lückenlos über das derzeitige Angebot von Schultests.

4.1. Leistungsmessung und Lernzieldefinition

Ralf Horn

4.1.1. Voraussetzungen der Leistungsmessung

Bei Untersuchungen der schulischen Leistungsbeurteilung standen lange Zeit die formalen Gütekriterien wie Objektivität, Reliabilität und Validität im Vordergrund (INGENKAMP 1971). Da ein großer Teil der Untersuchungen von Psychologen gemacht wurde, die die Begriffe der Testkonstruktion auf die schulischen Beurteilungsmöglichkeiten übertrugen, ist es nicht überraschend, daß der Zusammenhang zwischen Unterricht und Leistungsfeststellung wenig beachtet wurde.

Es kommt hinzu, daß die Regeln der Testkonstruktion auf die schulische Leistungsmessung angewendet werden können, ohne daß man sich um die einzelnen Abschnitte des Lehrplans zu kümmern braucht. Es würde also ausreichen, zu einem bestimmten Unterrichtsabschnitt, der inhaltlich festgelegt ist (etwa Bruchrechnen), einfach Aufgaben zu konstruieren. Die Aufgaben werden dann einer Analyse unterzogen und die geeigneten dann zu einer endgültigen Testform zusammengestellt, die dann noch geeicht werden muß.

Das beschriebene Verfahren stellt eine bewußte Übertreibung dar, zeigt aber, daß die Bindung von standardisierten Verfahren an den Unterricht, an dem die Schüler teilgenommen haben, gering ist. Das ist auch deswegen notwendig, weil die standardisierten Tests nicht nur in einem regional begrenzten Bereich, sondern an der Gesamtpopulation von Schülern einer bestimmten Klasse und Klassenstufe eingesetzt werden sollen. Damit dies trotz unterschiedlicher Curricula und Schulbüchern möglich ist, muß die Bindung an den Unterricht notwendigerweise locker sein.

Es liegt auf der Hand, daß derartige Verfahren nicht in optimaler Weise auf den Unterricht zugeschnitten sind. Es ist offensichtlich notwendig, neben dem Stoff des Unterrichts noch eine weitere Festlegung zu treffen, die das Ziel des Unterrichts in irgendeiner Form widerspiegelt.

Die Notwendigkeit der exakten Planung von Unterrichtseinheiten wurde zuerst von den Vertretern des programmierten Unterrichts erkannt. Damit eine optimale Sequenz in einem linearen Programm (etwa nach SKINNER) festgelegt werden kann, müssen für jeden Teilabschnitt Aussagen darüber vorliegen, was vom Schüler erwartet wird. Diese Ziele werden als „operationalisierte" Ziele bezeichnet. Darüber, ob man von operationalisierten „Lernzielen" (MAGER 1965) oder operationalisierten „Lehrzielen" (KLAUER et al. 1972) sprechen soll, besteht zur Zeit keine Übereinstimmung. Die Verwendung des einen oder anderen Begriffs hängt von der Betonung der ver-

schiedenen Aspekte ab. Wenn man von geplantem Unterricht spricht, sollte man den Begriff „Lehrziele" verwenden. In den Lehr- oder Lernzielen für eine bestimmte Unterrichtseinheit kommt das verwendete Curriculum recht deutlich zum Ausdruck und sie liefern daher eine Grundlage für eine Leistungsmessung, die sich am konkreten Unterricht orientiert. Die Vorteile der Festlegung von operationalisierten Lernzielen lassen sich an einem Beispiel am besten aufzeigen. Wenn in einem Lehrplan lediglich eine Festlegung von zu unterrichtenden Stoffen festgelegt ist, kann der Unterricht von zwei verschiedenen Lehrern an unterschiedlichen Schulen stark voneinander abweichen. So kann das Drama „Wilhelm Tell" von Schiller einmal als typischer Vertreter der Gattung „Drama" behandelt werden und zum anderen als Beschreibung der sozialen Verhältnisse in der Schweiz des frühen 15. Jahrhunderts. Die Leistungen beider Schülergruppen sind trotz des gleichen Unterrichtsthemas nicht vergleichbar.

Eine angemessene Messung der Schülerleistung ist nur dann möglich, wenn die Ziele des Unterrichts in operationalisierter Form vorliegen.

Bei den meisten Einsatzmöglichkeiten schulischer Leistungsmessung ist eine Diagnose allein, die sich bei den standardisierten Verfahren außerdem auf den Vergleich mit dem Durchschnitt der vergleichbaren Population beschränkt, nicht sinnvoll. Für den Lehrer reicht es nicht aus, zu wissen, wie gut oder wie schlecht ein Schüler ist, er benötigt auch Hinweise dafür, wie er den Schüler, der relativ schlecht abgeschnitten hat, fördern kann. Diese „therapeutischen Maßnahmen" sind jedoch nur dann möglich, wenn eine Leistungsfeststellung anhand von ausreichend definierten Lernzielen vorgenommen wird.

Eine lernzielbezogene Leistungsmessung ist daher aus folgenden Gründen wünschenswert:

1. Mit Lernzielen ist das Curriculum, das dem Unterricht zugrunde liegt, genau fixiert, so daß das gemessen wird, was auch unterrichtet wurde.
2. Die Messung der Schülerleistung an operational definierten Lernzielen liefert Ansatzpunkte zu gezielten Förderungsmaßnahmen bei einzelnen Schülern.

Diese Form der Testverfahren, die für den Unterricht besonders gut geeignet ist, läßt sich jedoch nur dann konstruieren, wenn einige Voraussetzungen gegeben sind. Dazu gehört in jedem Fall, daß der Unterricht durch operationalisierte Lernziele gesteuert wird.

4.1.2. Operationalisierte Lernziele

Der Begriff der Operationalisierung von Lernzielen spielt in der Didaktik (MÖLLER 1969) aber auch in der Pädagogischen Psychologie eine zunehmend wichtigere Rolle. Zunächst bedeutet Operationalisierung, daß die Verhal-

tensweisen, die der Schüler am Ende einer Unterrichtseinheit zeigen soll, festgelegt werden. Diese Verhaltensweisen müssen direkt beobachtbar sein. Daher findet man bei Lernzielen häufig Formulierungen wie „Der Schüler soll . . . nennen (aufschreiben, lösen usw.) können".

Ein Lernziel im Naturkundeunterricht könnte etwa das folgende sein: Der Schüler soll die verschiedenen Teile der Pflanze (Stengel, Blatt, Blüte) richtig benennen können. Derartige Ziele sind sehr einfach und es ist einzusehen, daß diese Lernziele (Feinziele im Sinn von MÖLLER) nicht in ein Curriculum aufgenommen werden können, da damit nur eine Aufblähung der Zahl der zu erreichenden Ziele stattfinden würde.

Wenn man das beispielhaft gegebene Ziel analysiert, so stellt man fest, daß damit lediglich die Vermittlung bestimmter Wissenselemente abgedeckt wird, d. h. Ziele dieser Art können nur der niedrigsten Stufe der Taxonomie von BLOOM (1972) zugeordnet werden. Da jedoch gerade das Wissen über bestimmte Sachgebiete einem schnellen Wandel unterliegt, kann es nicht Aufgabe der Schule sein, lediglich Wissen zu vermitteln. Das wird in einigen Lehrplänen zur Zeit dadurch angedeutet, daß verlangt wird, der Schüler sollte etwas „wirklich verstehen". Das Verständnis wird dabei aber nicht so konkret gefaßt, daß es sich anhand dieser Formulierung überprüfen ließe. Um jedoch den Erfolg eines Unterrichts erfassen zu können, sollten auch komplexere geistige Leistungen so operationalisiert werden, daß es möglich ist, einen entsprechenden Test zu entwickeln.

Bei komplexeren geistigen (kognitiven) Prozessen wird das Niveau der einzusetzenden Fähigkeit weitgehend durch die Art des verwendeten Problems festgelegt. Daß diese Festlegung relativ leicht möglich ist, beweist GUILFORD's (1971) Erfahrung bei der Konstruktion von Tests für bestimmte geistige Leistungen aus dem Strukturmodell, die vorher noch nicht bekannt, d. h. operationalisiert waren.

Daraus läßt sich folgern, daß bei komplexeren Lernzielen jeweils ein Problem als „Musteraufgabe" für die Festlegung des Niveaus des Lernziels mitgegeben sein muß. Die Lernziele lassen sich daher in ein kognitives „Grundmuster" und in einen entsprechenden fachspezifischen Inhalt aufgliedern (vgl. HORN 1972). Bei der Untersuchung von HORN konnte belegt werden, daß es möglich ist, komplexere kognitive Lernziele auf diese Weise festzulegen.

Bis jetzt wurde so argumentiert, als stünden sich nur Lernziele, die Wissen vermitteln und solche, die sich auf höhere kognitive Prozesse beziehen, gegenüber. GAGNÉ (1969) unterstützt diesen Gesichtspunkt, in dem er als höchste Stufe seiner Lernhierarchie „Problemlösen" verwendet. Detaillierte Untersuchungen, die sich speziell mit dem Problem der Lernziele im Bereich der Schule beschäftigen, kommen zu wesentlich detaillierteren Aufgliederungen (vgl. BLOOM 1972, WOOD & SKURNIK 1970 u. a.).

BLOOM unterscheidet in der Taxonomie sehr verschiedene Komplexitätsstufen, die, um einen Überblick zu geben, hier aufgeführt sind:

1. Wissen
2. Verstehen
3. Anwendung
4. Analyse
5. Synthese
6. Evaluation

Dabei werden die Kategorien 2—6 unter „Intellektuelle Fähigkeiten und Fertigkeiten" zusammengefaßt und gegen das mit zahlreichen Unterteilungen versehene Niveau „Wissen" abgesetzt.

WOOD und SKURNIK kommen bei einer ausgedehnteren Analyse von Lernzielen im Fach Mathematik als Vorstufe zum Aufbau einer Aufgabensammlung (Itembank) zu folgenden Kategorien:

1. Wissen (Knowledge)
2. Fertigkeiten (Skills)
3. Verstehen (Comprehension)
4. Anwendung (Application)
5. Erfindungsgabe (Inventiveness)

Um die verschiedenen Komplexitätsniveaus bei den Lernzielen unterscheiden zu können, benötigt man möglichst eindeutige Definitionen. Da die Taxonomie von BLOOM bereits in zahlreichen Untersuchungen bestätigt werden konnte, was auch für den hierarchischen Aufbau gilt (KROPP & STOKER 1966), werden wir uns bei der Aufstellung der Definitionen darauf beschränken.

Die Kategorie „Wissen" kann, wie einige Überlegungen zeigen, ausgeklammert werden. Die komplexeren Formen des kognitiven Verhaltens lassen sich etwa als Strategien des Problemlösens auffassen. Diese Strategien sind an sich frei von Bindungen an irgendwelche fachspezifischen Inhalte. Diese unabhängige Struktur wird in der Folge als „Grundmuster" bezeichnet. Um zu erreichen, daß die Schüler bestimmte Problemlösungsstrategien beherrschen, ist es offensichtlich nicht notwendig, sich an einen ganz bestimmten Stoff zu halten. Das ist auch einleuchtend, da die meisten der komplexeren Ziele in den verschiedenen Fächern auftreten.

Die von BLOOM (1972) angegebenen Klassifikationen der geistigen Fähigkeiten und Fertigkeiten sind für die praktische Anwendung leider etwas zu unscharf. Für die Planung von Unterricht oder die Aufstellung von Lernzielen wird die folgende Präzisierung empfohlen:

Verständn. = Übertragung von einer Kommunikationsform in eine andere
Beispiel: Übertragen von Formeln in verbale Aussagen, Nacherzählungen usw.

An-
wendung = Gelernte Regeln werden auf Situationen übertragen, die *neu* für den Schüler sind
Beispiel: Voraussagen von Veränderungen, die erfolgen, wenn eine spezifierte Maßnahme getroffen wird

Analyse = Beurteilung von Situationen anhand gegebener Richtlinien
Beispiel: Beurteilung der Angemessenheit von Schlußfolgerungen (logische Stimmigkeit)

Synthese = Herstellen von mehreren Lösungen, die zumindest für das betreffende Individuum neu sind
Beispiel: Auffinden von Verfahren zur Überprüfung von Annahmen oder Behauptungen

Evalution = Beurteilung von Situationen an internalisierten Wertsystemen

Nach der Untersuchung von KROPP & STOKER ist es fraglich, ob die Evaluation an der Spitze der Hierarchie bei der Taxonomie richtig placiert ist. Sinngemäß könnte sie auch bei der Analyse eingeordnet werden. Da die Forderung der Beurteilung von Situationen nach internalisierten Regeln einen längeren Lernprozeß impliziert, kann man bei allen Fragen der Planung begrenzter Unterrichtseinheiten die Kategorie zunächst vernachlässigen.

Mit dieser Operationalisierung ist es möglich, Lernziele zu formulieren, die den verschiedenen Stufen der Taxonomie entsprechen. Für den Unterricht in einem bestimmten Fach ist es dann noch notwendig, einen entsprechenden Inhalt zu finden, an dem die Schüler eine bestimmte Lernerfahrung machen sollen.

An einem Beispiel läßt sich das am besten zeigen:

Wenn etwa im Fach Biologie die Ernährung der Pflanze behandelt werden soll, lassen sich folgende Lernziele u. a. verfolgen:

1. Wissen

„Der Schüler soll anhand einer Zeichnung die Leitungsbahnen einer Pflanze ankreuzen können" (HORN 1972, S. 34).

2. Verstehen

„Der Schüler soll die gegebenen Informationen richtig interpretieren können. Beispiel: Pflanzen gedeihen auf verschiedenen Böden unterschiedlich gut. Schwere Böden speichern mehr Wasser als leichte Böden. Wie wirkt sich das auf das Wachstum bestimmter Pflanzen aus, die viel bzw. wenig Wasser vertragen?" (HORN 1972, S. 35).

3. Anwendung

„Der Schüler soll in der Lage sein, den Effekt von Veränderungen verschiedener Faktoren vorauszusagen. Beispiel: Es gibt chemische Stoffe, die auf die Pflanzen so wirken, daß alle Blätter abfallen (Entlaubung). Welche Folgen hat die Anwendung dieser Stoffe für die Pflanze?" (HORN 1972, S. 36).

4. Analyse

„Der Schüler soll die Richtigkeit von Schlußfolgerungen an gegebenen Annahmen oder Informationen überprüfen können. Beispiel: Alle Pflanzen verdunsten Wasser. Wenn eine Pflanze wenig Wasser verdunsten kann ...
a) wächst sie langsamer
b) erzeugt sie mehr Nährstoffe
c) bildet sich die Wurzel besser aus
d) werden die Blätter größer (nach HORN 1972, S. 36).

5. Synthese

„Der Schüler soll in der Lage sein, Wege zu finden, die es ermöglichen, Hypothesen (Annahmen, Behauptungen) zu überprüfen. Beispiel: Die Pflanzen produzieren in der Nacht in den Blättern bestimmte Substanzen, die nicht weitergeleitet werden. Wie läßt sich feststellen, ob diese Behauptung richtig oder falsch ist?"
(HORN 1972, Anhang)

Lernziele, die der Evaluation entsprechen, lassen sich nach den zur Zeit vorliegenden Erfahrungen aus zwei Gründen schlecht operationalisieren. Erstens nimmt das Erreichen dieses Zieles so lange Zeit in Anspruch, daß es bei der Planung von Unterrichtseinheiten nicht berücksichtigt werden kann, und zweitens ist dieses Ziel im Gegensatz zu den anderen kognitiven Zielen schwer überprüfbar.

Wie bereits angedeutet, hat die Operationalisierung von Lernzielen einige Auswirkungen auf die Gestaltung des Unterrichts. Nicht alle Formen des Unterrichts (Partnerarbeit, Frontalunterricht, Stillarbeit) sind allen Komplexitätsstufen der Lernziele angemessen. Bei höheren Zielen muß der Aktivität der Schüler zum Ausprobieren und Durchdenken von Lösungen mehr Zeit eingeräumt werden als etwa bei Wissenszielen.

Darüber hinaus haben operationalisierte Lernziele, die den Schülern bekanntgegeben werden, den positiven Effekt, daß sie den Unterricht erleichtern und dazu führen, daß sie schneller erreicht werden. Die optimistische Behauptung von MAGER (1965) „Wenn Sie jedem Lernenden eine Ausfertigung Ihrer Lernzielbeschreibung geben, werden Sie selbst nicht mehr viel zu tun haben" konnte bei einem Experiment von HASTINGS (1972) bestätigt werden. Allerdings wurden bei der Untersuchung von HASTINGS Studenten als Versuchspersonen verwendet. Es liegt jedoch nahe zu vermuten, daß zumindest bei älteren Schülern ähnliche Vorteile zu verzeichnen wären.

Nach den Vorstellungen verschiedener Gruppen, die an der Erstellung von Curricula arbeiten (Freiburger Arbeitsgruppe für Lehrplanforschung FAL, Institut für Pädagogik der Naturwissenschaften in Kiel), wird das Curriculum nicht aus einer Sammlung von unterrichtsrelevanten Stoffen, sondern aus Lernzielen bestehen. Der erste Lehrplan, der aufgrund dieser Überlegungen konstruiert wurde, war der Lehrplan der Primarschule des Kantons Freiburg. (Beschreibungen dieses Projekts sind in verschiedenen Berichten

der FAL zu finden.) Für den Unterrichtenden besteht dann nicht die Notwendigkeit, Lernziele selbst zu formulieren, aber er muß aus dem Lernzielangebot einen Teil auswählen.

Bei den Lernzielen müßte ein entsprechender Hinweis darüber angebracht werden, welchem Komplexitätsniveau das Ziel angehört, oder der Lehrer müßte die Ziele analysieren, bevor er daraus eine Unterrichtseinheit macht. Da die Zahl der Lernziele für ein Schuljahr und ein Fach bereits ziemlich groß ist (vgl. den in Zusammenarbeit mit der FAL entwickelten Lehrplan der Primarschule des Kantons Freiburg), können nicht alle Lernziele so formuliert sein, daß sie unmittelbar für den Unterricht zu verwenden wären.

Lernzielanalyseschema

1. Kann zwischen Inhalt und Grundmuster beim Lernziel unterschieden werden?
   - nein → Wissen
   - ja ↓
2. Sind alle Informationen für die Lösung der Aufgabe vorhanden?
   - ja → Verständnis
   - nein ↓
3. Müssen vom Schüler Informationen zur Lösung des Problems beigesteuert werden?
   - ja → Anwendung
   - nein ↓
4. Wird vom Schüler eine Untersuchung vom Sachverhalt nach vorgegebenen Kriterien erwartet?
   - ja → Analyse
   - nein ↓
5. Ist es möglich, mehr als eine Lösung für das Problem zu finden?
   - ja → Synthese

(Horn 1972, S. 40)

Es ist daher notwendig, die Ziele des Lehrplans zunächst zu analysieren. Zweckmäßigerweise verwendet man dazu das von HORN (1972) entwikkelte Schema. Dieses Analyseschema besteht aus einem Blockdiagramm, bei dem eine Reihe von aufeinanderfolgenden Fragen eine Lokalisation des Zieles erlaubt.

Nach der Analyse des Lernzieles kann mit der Planung des Unterrichtsablaufes begonnen werden. Parallel zur Unterrichtsplanung sollten dann bereits die Aufgaben bzw. Probleme festgelegt werden, die zur Kontrolle des Unterrichtserfolges verwendet werden sollen. Damit gibt es zwischen dem Inhalt und der Intention des Unterrichts und der Kontrolle des Erfolges keine Verschiebungen, die sich negativ auf die Schülerleistungen auswirken.

4.1.3. Erstellung von Prüfungsaufgaben

Nachdem die Lernziele für eine Unterrichtseinheit festgelegt und entsprechend operationalisiert sind, kann die konkrete Planung des Unterrichts und die gleichzeitige Konstruktion der Prüfungsaufgaben beginnen. Damit zwischen der Intention des Unterrichts und der Prüfung keine Lücke entsteht, empfiehlt es sich, bereits bei der Planung des Unterrichts die zur Überprüfung verwendeten Aufgaben zu entwickeln. Allerdings kann das nicht bedeuten, daß der Unterricht nur auf etwas ausgerichtet wird, was überprüfbar ist. Eine derartige Möglichkeit der Einschränkung wäre leicht gegeben, wenn nur Wissensziele im Unterricht berücksichtigt werden.

Die erste Frage, die bei der Erstellung von Prüfungsaufgaben zu beantworten ist, bezieht sich auf das numerische Verhältnis von Aufgaben zu einem Lernziel. Auf der einen Seite hängt die Zahl der zu verwendenden Aufgaben davon ab, wie viele Unterteilungen gemacht werden sollen. Mit einer einzigen Aufgabe kann man im günstigsten Fall eine Trennung in zwei Gruppen vornehmen. Wenn Sie etwa die Körperkraft einer Gruppe von Menschen untersuchen und Sie nur eine Aufgabe (ein Gewicht) verwenden, dann kann die Gruppe in diejenigen eingeteilt werden, die dieses Gewicht heben können und diejenigen, die dazu nicht in der Lage sind.

Die Verhältnisse liegen bei Aufgaben, die zu Lernzielen gestellt werden, zweifellos etwas schwieriger. Obwohl auch hier nur eine Trennung in zwei Schülergruppen angestrebt wird, diejenigen, die das Lernziel erreicht und diejenigen, die es nicht erreicht haben, muß mehr als eine Aufgabe gestellt werden. Diese größere Zahl von notwendigen Aufgaben geht zum einen darauf zurück, daß es eine so einfache Beziehung wie zwischen dem Hochheben eines Gewichts und der Körperkraft bei Lernzielen nicht gibt. Da auch bei lernzielbezogenen Aufgaben die Grundregeln der Aufgabenkonstruktion und die Möglichkeit objektiver Auswertung berücksichtigt werden

muß, können die Aufgaben auch durch „Zufall" gelöst werden. Darüber hinaus ist Wissen, das in Form von Mehrfachwahlaufgaben geprüft wird, keine „Alles-oder-Nichts-Funktion".

Es gibt also zwischen völligem Wissen und dem Nichtwissen zahlreiche Stufen von Halbwissen, die durchaus zu Lösungen von Aufgaben führen können.

Während auch bei Wissenslernzielen nach Möglichkeit mehrere Aufgaben gestellt werden müssen, sind die Verhältnisse bei höheren kognitiven Lernzielen ähnlich, wenn auch aus anderen Gründen.

Nach der Taxonomie von BLOOM, die inzwischen mehrfach bezüglich ihres hierarchischen Charakters bestätigt werden konnte, schließt jedes übergeordnete Ziel alle anderen, die ihm untergeordnet sind, ein. Bei der Beantwortung von Aufgaben, die sich auf ein bestimmtes Problem beziehen, wird jedoch auf die für das Individuum ökonomischste Weise der Problemlösung zurückgegriffen. Wenn etwa bei einer Aufgabe die Anwendung einer Regel gefordert wird, der Schüler aber die Lösung des Problems bereits kennt, dann wird er nicht nochmals die Lösung des Problems durchführen, sondern lediglich die Lösung aus seinem Gedächtnis abrufen. Da die Lernerfahrung und die Lerngeschichte des Schülers weit über das hinausgehen, was in der Schule unterrichtet wird, ist vor allem bei praxisbezogenen Aufgaben damit zu rechnen, daß ein Teil von ihnen deswegen gelöst wird, weil die Schüler gerade dieses Problem schon kennengelernt hatten.

Daraus ergibt sich die Schlußfolgerung, daß bei Lernzielen, die komplexere Probleme stellen, wesentlich mehr Aufgaben erforderlich werden als bei einfacheren Zielen. Diese Schwierigkeit kann auch nicht dadurch umgangen werden, daß man auf die Form der Mehrfachantwortauswahlaufgaben verzichtet und zu einer weniger objektiven Form der Aufgabenstellung übergeht. In der Untersuchung von HORN (1972) konnte gezeigt werden, daß bei offener Aufgabenstellung (keine vorgegebenen Antworten) von den Schülern die Fragen nur dann beantwortet werden, wenn sie irgendwelche Vorinformationen über die eigentliche Problemlösung hatten. Das bedeutet letzten Endes nichts anderes, als daß offene Aufgaben zur Überprüfung von komplexeren Lernzielen recht wenig geeignet sind.

Da bei lernzielbezogenen Testverfahren die durchschnittliche Aufgabenschwierigkeit recht niedrig liegt (hohe Zahlenwerte), ist es schwierig, eine genaue Zahl von Aufgaben anzugeben, die zu einem bestimmten Lernziel konstruiert werden sollen. Da es außerdem keine zwingende Vorschrift für die Ableitung derartiger Aufgaben aus den Unterrichtstexten gibt und es diese auch vorläufig nicht geben kann, hängt die Zahl der konstruierten Aufgaben im wesentlichen von dem Einfallsreichtum dessen ab, der diese Aufgaben konstruiert. Man sollte jedoch darauf achten, daß mindestens 5 Aufgaben jedem Lernziel zugeordnet werden.

Auch dann ist wegen der Möglichkeit, daß die Lösungen des Schülers zufällig zustandekommen, nur dann eine Entscheidung über den Lernzustand des Schülers möglich, wenn er entweder alle Aufgaben beantwortet oder alle Aufgaben falsch löst. Das ergibt sich aus den Überlegungen, die das Vertrauensintervall von bestimmten Lösungshäufigkeiten berücksichtigt.

Bei der eigentlichen Aufgabenkonstruktion ist es am sinnvollsten, mit den komplexen Lernzielen zu beginnen und dort entsprechende Probleme zu suchen. Aus komplexeren Problemen kann man durch Einschränkungen oder durch Hinzufügen von Informationen relativ schnell einfachere Probleme machen, während es umgekehrt nicht möglich ist, aus einfach strukturierten Aufgaben kompliziertere zu machen. Bei der Konstruktion der Aufgaben sollte man die Entscheidung darüber, welcher Komplexitätsstufe die Aufgabe zuzuordnen ist, anhand der vorliegenden Planungsunterlagen des Unterrichts und dem Analyseschema für Lernziele treffen.

Ein konkretes Beispiel läßt sich infolgedessen auch nur dann beschreiben, wenn man sowohl die Lernziele des Unterrichts und die inhaltliche Festlegung als auch die Prüfungsaufgaben kennt.

Sehr gut geeignet, um das Verfahren der Aufgabenkonstruktion zu zeigen, ist ein Lernziel, das der Komplexitätsstufe „Anwendung" zuzuordnen ist. Auch spielt dabei eine Rolle, daß die Aufgabe der Schule wohl am besten bei Lernzielen deutlich wird, die eine Anwendung der in der Schule vermittelten Informationen oder Regeln fordern. Das Niveau der Anwendung läßt sich beschreiben als „Finden und Anwenden von Regeln auf Situationen, die dem Schüler *neu* sind".

Bei derartigen Lernzielen kommt es vor allem darauf an, daß der Schüler mit Problemen konfrontiert wird, die nicht Gegenstand des Unterrichts waren. Da Lernen nicht nur in der Schule stattfindet, sollten die Probleme aus Bereichen stammen, die auch nicht zu der Erfahrung der Schüler gehören.

Wenn in einer 5. Klasse etwa das Thema „Ernährung der Pflanzen" behandelt wird, und es Ziel des Unterrichts ist, die Schüler so weit zu bringen, daß sie die gelernten Regeln auf neue Situationen übertragen können, kann folgendes Lernziel formuliert werden:

Der Schüler soll voraussagen können, welche Wirkungen bei einer Pflanze auftreten, wenn Eingriffe bei der Ernährung der Pflanzen vorgenommen werden.

Ein derartiges Lernziel hat, wie sich auf den ersten Blick zeigt, eine relativ große Bedeutung, die auch für Schüler zu erkennen ist. Bei diesen komplexen Lernzielen ist zu überlegen, ob sie nicht den Schülern vor der betreffenden Unterrichtseinheit mitgeteilt werden sollen.

Bei der Planung des Unterrichts können jetzt einige Situationen zur Demonstration der zu übermittelnden Informationen konstruiert werden.

Voraussetzung zur Erreichung des Lernziels ist, daß die Schüler zunächst erfahren, wie die Ernährung bei den Pflanzen vor sich geht. Mit einiger Wahrscheinlichkeit ist es besser, bei derartigen komplexen Zielen die Eigenaktivität der Schüler zu fördern, damit sie in ausführlicher Weise mit Strategien des Problemlösens vertraut gemacht werden und nicht später vor neuen Fragestellungen einfach aufgeben. Außerdem wird durch eine geeignete Fragestellung auch die Motivation der Schüler verstärkt (extrinsische Motivation).

Bei der angegebenen Unterrichtseinheit könnte man mit dem folgenden Versuch beginnen:

In zwei mit Leitungswasser gefüllte Gläser (Meßzylinder) stellt man je 2—3 bewurzelte Sprosse der Ampelpflanze. Um die Verdunstung an der Wasseroberfläche zu vermeiden, überschichtet man bei einem Glas die Wasseroberfläche mit etwas Paraffinöl. Über die beiden Gläser stülpt man je eine Glas- oder Kunststoffglocke und stellt das Ganze etwa zwei Stunden in die Sonne. Danach ist auf der Innenseite der Behälter ein Beschlag mit feinen Wassertröpfchen festzustellen (nach HORN 1972, S. 30 f.).

An diesem einfachen Versuch lassen sich einige der Vorgänge bei der Ernährung der Pflanzen zeigen, die sich später auch überprüfen lassen. Aus dem Versuch lassen sich einige Prinzipien ableiten, die am Ende des Unterrichts überprüft werden können, etwa mit folgender Aufgabe:

„Bäume verdunsten mehr Wasser als kleinere Pflanzen. Wenn man jedoch gleich große Blätter verschiedener Pflanzen vergleicht, sind Unterschiede in der Verdunstung von Wasser festzustellen. Was könnte für diesen Unterschied verantwortlich sein?
A) Die Länge der Leitungsbahnen im Blatt
B) Die Oberfläche des Blattes
C) Die Beschaffenheit des Blattrandes (gesägt oder glatt)
D) Die Lichtdurchlässigkeit (HORN 1972, Anhang).

Wenn der Unterrichtsablauf und die eigentlichen Prüfungsaufgaben parallel entwickelt werden, wie es bei diesem Beispiel demonstriert wurde, kann relativ leicht die Effektivität des Unterrichts überprüft werden.

Für die Konstruktion der Prüfungsaufgaben ist jedoch die Frage zu stellen, wie man Probleme für die Aufgaben findet.

Es ist einsichtig, daß eine ganze Reihe von praxisbezogenen Problemen für eine Aufgabenstellung wenig geeignet sind, da in den meisten Fällen nicht nur ein einziger Faktor oder eine Regel eine Rolle spielt, sondern mehrere. In diesen Fällen sollte man die Konstruktion von fiktiven Situationen erwägen, wie es auch von BLOOM (1972) vorgeschlagen wird.

Dabei werden von BLOOM drei Grundtypen der Problemsituation unterschieden:

A) Darstellung einer fiktiven Situation,
B) Verwendung von Material, das dem Schüler wahrscheinlich nicht vorher bekannt ist. Solche Situationen bestehen häufig in vereinfachten Versionen komplexen Materials, das normalerweise viel später im Kurs dargeboten würde (. . .),
C) die Aufnahme eines neuen Gesichtspunktes bei Situationen, die der getesteten Gruppe bekannt sind. In hochtechnisierten Bereichen können „bekannte" Probleme für Nichtfachleute ziemlich unbekannt sein (BLOOM 1972, S. 139).

In der Taxonomie von BLOOM finden sich zahlreiche Aufgabenbeispiele für die verschiedenen Grundtypen der Problemsituation. Die drei Grundtypen sollen hier lediglich kurz anhand von einigen Problemstellungen charakterisiert werden:

A) fiktive Situation:
Eine physikalische Aufgabe, bei der ein Aufzug mit konstanter Beschleunigung „g" abwärts fährt, stelle eine fiktive Situation dar. Man kann daran allerdings sehen, daß sich derartige Situationen gut eignen, um komplexere Phänomene auf einfache zu reduzieren, ohne daß die Situation allzu unrealistisch wird.

B) Zu Typ B läßt sich nur dann ein Beispiel finden, wenn der Unterricht vollständig geplant wurde.
C) Aufgaben dieses Typs lassen sich relativ leicht finden. Sie sind auch für die Überprüfung von anderen Lernzielen sehr gut geeignet. Ein Beispiel liefert folgendes Problem:

„Zu welcher Tageszeit holt sich jemand, der ein Sonnenbad nehmen möchte, am wahrscheinlichsten einen schweren Sonnenbrand? Er wird sich am wahrscheinlichsten den schweren Sonnenbrand mittags holen (11.00 bis 13.00 Uhr), weil:
( ) wir über Mittag der Sonne etwas näher sind als am Morgen oder am Nachmittag.
( ) die Mittagssonne einen stärkeren „Brand" erzeugt als die Morgen- oder Nachmittagssonne.
( ) wenn die Sonnenstrahlen direkt (senkrecht) auf eine Oberfläche fallen, diese Oberfläche mehr Energie aufnimmt, als wenn die Strahlen schräg auf sie treffen.
( ) wenn die Sonne senkrecht über uns steht, ihre Strahlen durch weniger absorbierende Atmosphäre dringen, als wenn sie niedriger am Himmel steht.
( ) die Luft mittags gewöhnlich wärmer ist als zu anderen Tageszeiten.
( ) für den Sonnenbrand das ultraviolette Licht verantwortlich ist."
(BLOOM 1972, S. 146).

4.1.4. Zusammenfassung

Die Messung schulischer Leistungen kann in zuverlässiger und valider Weise nur dann vorgenommen werden, wenn die Lernziele vorliegen. Dabei sollten die Ziele des Unterrichts, um Schwierigkeiten bei der Interpretation auszu-

schalten, möglichst weitgehend operationalisiert sein. Da auch Lernziele komplexerer Art, etwa Problemlösungsverhalten, ebenfalls operationalisiert werden können, sollten diese Ziele auch im Unterricht entsprechend berücksichtigt werden.

Die Probleme bei der Aufgabenkonstruktion beginnen bei den Aufgaben, die komplexere Ziele überprüfen. Diese Schwierigkeit kann jedoch teilweise überwunden werden, wenn die Konstruktion der Aufgaben parallel zur Planung des Unterrichts vorgenommen wird. Ein Beispiel aus einer Unterrichtseinheit des Fachs Naturkunde wurde besprochen, um dieses Vorgehen zu illustrieren.

4.1.5. Literaturverzeichnis

Bloom, B. S. et. al.: Taxonomie der Lernziele im kognitiven Bereich. Weinheim 1972.

Gagné, R. M.: Die Bedingungen des menschlichen Lernens. Berlin, Darmstadt 1969.

Guilford, J. P.: The Analysis of Intelligence, New York 1971.

Hastings, G. R.: Independent Learning Based on Behavioral Objectives. In: Journal of Educational Research, May/June 1972.

Horn, R.: Lernziele und Schülerleistung, Weinheim 1972.

Ingenkamp, K. (Hrsg.): Die Problematik der Zensurengebung. Weinheim 1971.

Klauer, K. et. al.: Lehrzielorientierte Tests, Düsseldorf 1972.

Kropp, R. P. and *Stoker, H. W.:* The Construction and Validation of Tests of the Cognitive Processes as Decribed in the Taconomy of Educational Objectives, Florida State University 1966.

Mager, R. F.: Lernziele und programmierter Unterricht. Weinheim 1965.

Möller, C.: Technik der Lernplanung, Weinheim 1969.

Wendeler, J.: Standardarbeiten, Weinheim 1969.

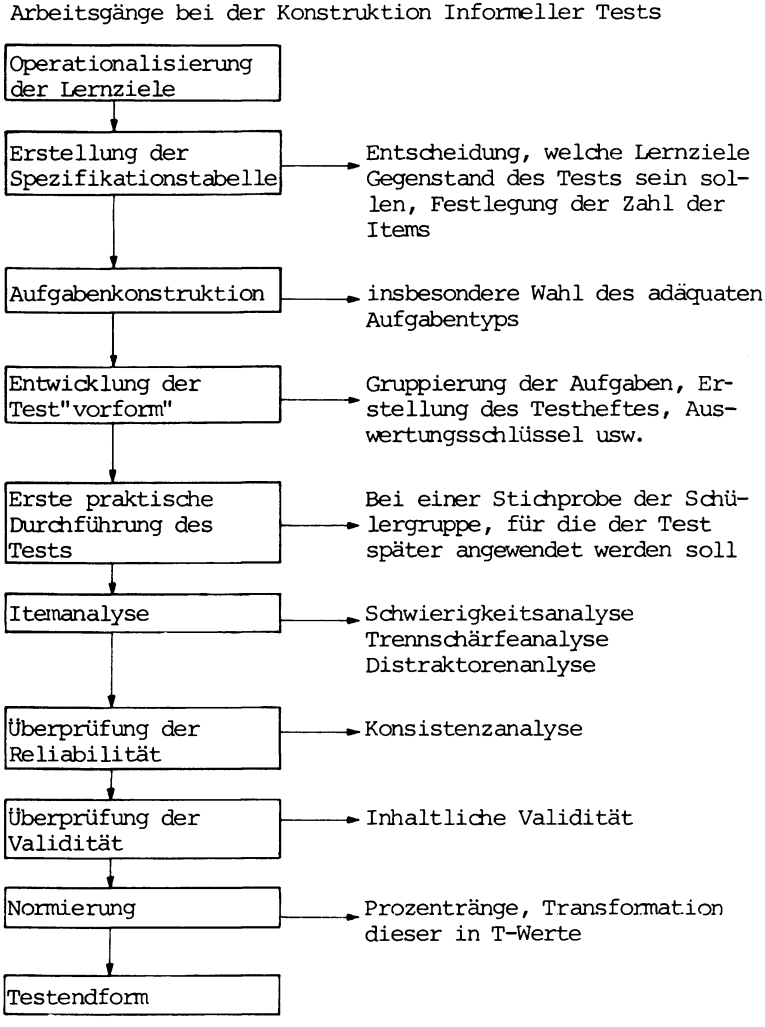
Wood, R. and *Skurnik, L. S.:* Item Banking. National Foundation for Educational Research in England and Wales, 1969.

4.2. Konstruktion und Einsatz von Informellen Tests zur Leistungsbeurteilung (Lernkontrolltests)

Bernhard Rosemann

4.2.1. Einleitung

Im folgenden Beitrag sollen die wichtigsten Schritte bei der Konstruktion informeller Tests dargestellt werden. Entsprechend unserer terminologischen Abgrenzung (s. ROSEMANN, Kap. 3.3., bes. S. 163 ff. in diesem Band) handelt



es sich dabei um *Lernkontrolltests*, also Tests, die zur Bewertung einer gegebenen Schülerleistung eingesetzt werden. Die Probleme der Konstruktion von *Lernsteuerungstests* werden insbesondere von BÜSCHER (Kap. 3.2.) behandelt. Den Lehrer, der nicht die Gelegenheit hatte, sich eingehend mit Theorie und Praxis der Testkonstruktion zu befassen, sollten unsere Ausführungen in die Lage versetzen, den einen oder anderen Test selbst zu entwickeln bzw. bestehende Tests kritisch zu beurteilen. Wer sich eingehender diesem Gegenstand widmen will, dem bleibt es nicht erspart, sich mit einer Reihe von Standardwerken vertraut zu machen (s. Literaturverzeichnis).

Wir stellen unseren Ausführungen ein Ablaufdiagramm (s. S. 182) voran, das in chematischer Form die notwendigen Arbeitsgänge der Testkonstruktion veranschaulicht.

4.2.2. Die Formulierung der Items

4.2.2.1. Die Spezifikationstabelle

Bevor man den ersten Schritt zur Konstruktion eines Tests tut, nämlich die Erstellung eines Aufgabenpools, müssen dem Überlegungen vorausgehen, von denen die Qualität des Tests ganz entscheidend abhängt.

Es muß überprüft werden, ob der Unterrichtsstoff, über den der Test konstruiert werden soll, in der Form von operational definierten Lernzielen vorliegt (s. HORN, Kap. 4.1. in diesem Band). Ist dies nicht der Fall, dann ist es zunächst erforderlich, den Stoff in operational definierte Lernziele umzuschreiben. Allein schon diese Arbeit, die z. T. sehr aufwendig sein kann, zwingt den Lehrer zu einer intensiven Auseinandersetzung mit der Frage „Was will ich den Schüler eigentlich lehren?" oder anders „Welche Verhaltensänderungen will ich beim Schüler bewirken?". Dabei dürfte manchem Lehrer klar werden, daß er so recht eigentlich bisher gar nicht gewußt hat, was er denn seinen Schülern im einzelnen beizubringen beabsichtigt, obwohl er, hätte man ihn nur danach gefragt, sicher das Gegenteil dessen behauptet hätte.

Der Testautor muß sich dann Gedanken darüber machen, welche Lernziele er in welchem Umfang in dem Test abprüfen will. Will er (im Sinne der BLOOMschen Taxonomie) nur Wissen abfragen, oder will er feststellen, ob der Schüler den Unterrichtsstoff verstanden hat, ob er ihn anwenden kann, oder will er mehreres gleichzeitig? Diese Überlegungen muß der Lehrer anstellen, will er nicht in die Gefahr geraten, das Testergebnis falsch zu interpretieren. Prüft ein Test beispielsweise nur Wissenstatbestände, die der Schüler womöglich mechanisch auswendig gelernt hat, und der Lehrer interpretiert ein gutes Abschneiden in diesem Test in der Weise, daß er annimmt, der Schüler habe diesen Stoff auch verstanden, dann begeht er eine Selbsttäuschung, die zwar sein Gewissen beruhigen mag, aber den eigent-

lichen Sinn des Testeinsatzes verfehlt er damit. Ein nützliches Hilfsmittel, sich über die Frage klar zu werden „Welche Lernziele soll mein Test erfassen?", ist die sog. *Spezifikationstabelle* (s. dazu auch GRONLUND 1968, BLOOM et al. 1971).

Wie die Abbildungen 1 und 2 zeigen, handelt es sich bei der Spezifikationstabelle um eine zweidimensionale Matrix. Diese Matrix enthält in der Horizontalen die Verhaltenskomponenten der Lernziele (1 bis k), in der Vertikalen die Inhaltsbereiche (1 bis j). Es ergeben sich damit k x j Zellen, die jeweils durch ein bestimmtes Verhalten und einen spezifischen Inhalt determiniert sind.

Abb. 1:

Allgemeine Spezifikationstabelle für den Bereich der kognitiven Fähigkeiten und Fertigkeiten (s. BLOOM et al. 1956)

		Verhalten					
	1	2	3	4	5	6	
Inhalt	Wissen	Verständnis	Anwendung	Analyse	Synthese	Evaluation	Total
1. ...							
2. ...							
3. ...							
4. ...							
5. ...							
6. ...							
: ...							
: ...							
j							
Total							

Der Testautor hat jetzt zu entscheiden, ob er alle möglichen Kombinationen der Spezifikationstabelle in seinen Test aufnehmen oder ob er sich auf bestimmte beschränken will. Ist sein Ziel etwa eine reine Wissensprüfung, dann würden alle vertikalen Zellen außer der Spalte 1 leer bleiben.

An diesem Punkte läßt sich der Unterschied zwischen Informellen Tests zur Leistungsbeurteilung vs. Tests zur Leistungsfeststellung erneut deutlich machen. Für die Leistungs*beurteilung* wird der Testautor eine Stichprobe

Abb. 2:

Spezifikationstabelle für einen Informellen Test im Geschichtsunterricht

	Verhalten			
	1	2	3	
Inhalt	Wissen	Verständnis	Anwendung	Total
1. Grundherr u. Bauer	4	4	4	12
2. Mittelalterl. Stadt, Bauten	2	2	3	7
3. Bürger, Bauern, Mönche im Osten	5	7	6	18
4. Die Hanse in Europa	6	4	3	13
Total	17	17	16	50

Nach TBR 9, Inst. f. Film u. Bild, München 1969.

aus den Zellen entnehmen und von dieser auf die Leistung des Schülers hinsichtlich des Gesamtumfanges der Tabelle schließen. Er wird z. B. aus 200 Zellen nur 30 auswählen und diese als eine Schätzung der Leistung des Schülers hinsichtlich der 200 Zellen benutzen.

Bei der Leistungs*feststellung* zum Zwecke der Lernsteuerung wird er nur einen engen Ausschnitt aus der Tabelle wählen, etwa eine Zeile oder eine Spalte, und diesen dafür sehr detailliert und erschöpfend erfassen, um notwendigenfalls gezielt den Lernprozeß beeinflussen zu können. Hat sich der Testautor nun entschlossen, welche Zellen, d. h. welche Verhaltens-Inhalts-Kombinationen er in seinen Test aufnehmen will, dann muß er ferner darüber entscheiden, mit welchem Gewicht er sie jeweils im Test vertreten haben will. Konkret heißt das, er muß entscheiden, wieviel Items seines Tests eine bestimmte Verhaltens-Inhalts-Kombination repräsentieren sollen; je wichtiger sie ihm erscheint, desto mehr Items wird er dafür in den Test aufnehmen.

Der Lehrer trägt also in jede der Zellen, die er zu erfassen beabsichtigt, die Zahl der geplanten Items ein. Die *Spalte* „Total" der Spezifikationstabelle liefert dann die Information darüber, welches Gewicht verschiedene Verhaltenskomponenten in dem zu konstruierenden Test haben werden,

während die *Zeile* „Total" das Gewicht der Inhaltsbereiche deutlich werden läßt.

Nach welchen Gesichtspunkten hat man sich nun bei der Verteilung der Gewichte zu richten? Wie GRONLUND (1968) vorschlägt, sollte man dabei drei Aspekte berücksichtigen:

a) die Unterrichtszeit, die man für einen bestimmten Gegenstand verwendet hat;
b) die Bedeutung, die der Lehrer dem Gegenstand beimißt;
c) die Schwerpunkte, die man im Unterricht gesetzt hat.

Wurde die Spezifikationstabelle nach diesen Kriterien erstellt, dann sollte jedes Lernziel seiner Bedeutung im Unterricht entsprechend repräsentiert sein. Bei der Aufgabenkonstruktion empfiehlt es sich dann allerdings, mehr Aufgaben zu formulieren als in der Spezifikationstabelle vorgesehen sind, da mit großer Wahrscheinlichkeit sich einige Aufgaben bei der Itemanalyse als unbrauchbar erweisen werden.

4.2.2.2. Itemtypen und ihre Konstruktion

Bei der Beschreibung der Itemkonstruktion wollen wir so vorgehen, daß wir, wie es in der Regel gehandhabt wird, die Items nach der Art der geforderten Antworten klassifizieren, ihren Verwendungszweck aufzeigen, Konstruktionshinweise geben sowie die Vor- und Nachteile der verschiedenen Itemtypen deutlich machen werden. Unsere Erörterungen werden sich dabei auf Papier- und Bleistift-Tests beschränken.

Zunächst erscheint es sinnvoll, einen Überblick über die hier zu behandelnden Aufgabenarten zu geben.

Aufgaben

mit

gebundenen Antworten	nicht gebundenen Antworten
Auswahlantworten	Ergänzungsantworten
Alternativantworten Mehrfachwahlantworten	Kurzantworten
Ordnungsantworten	Kurzaufsatz
Zuordnungsantworten Umordnungsantworten	

Interpretationsübungen

sowohl gebundene als auch nicht-gebundene Antworten möglich

Im folgenden werden wir nun im einzelnen auf die verschiedenen Aufgabenarten eingehen und vor allem solche ausführlicher besprechen, die zum einen bereits breite Verwendung gefunden haben, sich zum anderen aber auch besonders zur Konstruktion Informeller Tests anbieten.

4.2.2.2.1. Aufgaben mit gebundenen Antworten

Die Aufgaben mit Auswahlantworten. Die Aufgaben mit Auswahlantworten lassen sich einteilen in zwei Untergruppen:

— die Alternativ-Antworten (alternative responses)
— die Mehrfachwahlantwort (multiple-choice-form).

Die Alternativ-Antwort. *a) Beschreibung.* Einige der Alternativ-Antwort-Items bestehen aus einer Feststellung (Aussagesatz), die der Schüler als richtig oder falsch zu kennzeichnen hat. Da diese Form sehr häufig verwendet wird, werden diese Aufgaben auch oft als „Richtig-Falsch-Items" (True-False-Items) bezeichnet. Andere Formen verlangen vom Schüler, mit „ja" oder „nein" zu antworten oder fordern ihn auf, zuzustimmen oder abzulehnen.

Beispiele:

Lies jede der folgenden Feststellungen sorgfältig durch. Wenn eine Feststellung richtig ist, kreuze "R" an, wenn sie falsch ist, kreuze "F" an!

	R	F
Der Stoff, der die Blätter grün färbt, wird als Chlorophyll bezeichnet.	☒	☐
Eines der Hauptexportgüter Venezuelas ist Öl.	☒	☐

Lies jede der folgenden Fragen. Wenn die Antwort "ja" ist, kreuze "ja" an, wenn sie "nein" ist, kreuze "nein" an!

	Ja	Nein
Ist 51 % von 38 mehr als 19?	☒	☐
Wenn 60 % einer Zahl 9 ist, ist die Zahl kleiner als 9?	☐	☒

Lies jede der folgenden Feststellungen. Wenn eine Feststellung richtig ist, kreuze "R" an, wenn sie falsch ist, kreuze "F" an, wenn sie eine Meinung darstellt, kreuze "M" an.

	R	F	M
Die Erde ist ein Planet	☒	☐	☐
Die Erde dreht sich um den Mond.	☐	☒	☐
Auf dem Mars gibt es keine Pflanzen oder Tiere.	☐	☐	☒

b) Zweck. Nach GRONLUND (1968) kann mit Hilfe dieser Aufgaben vor allem überprüft werden:

— Die Fähigkeit des Schülers, Aussagen über Fakten als richtig oder falsch zu beurteilen.

— Die Kenntnis von Begriffsdefinitionen, Regeln u. dergl. m.

— Die Fähigkeit des Schülers, Fakten von bloßen Meinungen unterscheiden zu können.

— Das Vermögen, einfache logische Schlüsse zu ziehen.

c) Konstruktionshinweise. GRONLUND (1968) nennt einige Regeln zur Ausarbeitung von Aufgaben mit Alternativantworten.

(1) Die Feststellungen sollen eindeutig richtig oder falsch sein.

Schlechtes Beispiel:

	R	F
Die Mondkrater werden durch Meteore verursacht	☐	☐

Der gute Schüler wird durch solche Aufgaben verwirrt, denn er weiß, daß die Krater auf dem Mond zum Teil auch durch vulkanische Aktivität entstanden sind.

Besser formuliert lautet dieses Item:

	R	F
Eine Ursache der Mondkrater war der Einfluß von Meteoren.	☐	☐

(2) Vermeide zu allgemeine Aussagen.

Dazu gehört, auch in der Formulierung darauf zu achten, daß man nicht ungewollt Hinweise auf die zutreffende Antwort gibt. So erscheinen Zusätze wie „gewöhnlich", „im allgemeinen", „oft", „manchmal" häufig in richtigen Feststellungen, während Worte wie „immer", „nie", „alle", „keiner" oder „nur" eher bei falschen Feststellungen zu finden sind.

(3) Vermeide zu triviale Feststellungen.

Solche Aussagen haben meist wenig Bedeutung für den Gesamtunterrichtsstoff.

(4) Vermeide negative Aussagen oder doppelte Verneinungen.

(5) Vermeide lange, inhaltsreiche Sätze.
(6) Vermeide es, zwei gedankliche Inhalte in einer Feststellung zu vereinen.
(7) Falsche und richtige Items sollten ungefähr gleich lang und gleich häufig sein.

d) Vorteile. Ein Vorzug der Aufgaben mit Alternativ-Antwort ist, daß sie sich sehr leicht herstellen lassen. Sie sind ferner einfach und objektiv auswertbar, und der Test wird in der Regel nur kurze Zeit beanspruchen. Allerdings sind diese sehr rasch erstellten Items dafür oft auch relativ simpel und liefern dem Testenden kein wirkliches Bild vom Leistungsstand des Schülers. Alternativ-Antworten zu konstruieren, die tatsächlich relevante Lernziele abprüfen, erfordert bereits ein hohes Maß an Können.

Eine ausführliche Beschreibung der Konstruktion von Alternativ-Antworten findet sich im übrigen bei EBEL (1965). Uns scheint dieser Autor allerdings eine etwas zu optimistische Haltung diesem Aufgabentyp gegenüber einzunehmen.

e) Nachteile. Ein wesentlicher Nachteil der Aufgaben mit Alternativ-Antworten liegt in dem sehr begrenzten Bereich der Lernziele, die mit ihnen überprüft werden können. In der Hauptsache verwendet man sie zur Überprüfung von Faktenwissen. Eine weitere Schwierigkeit besteht darin, Feststellungen zu formulieren, die absolut richtig oder absolut falsch sind.

Der wohl schwerwiegendste Einwand gegen die Alternativ-Antworten ist die hohe Ratewahrscheinlichkeit für richtige Lösungen, nämlich 50 %. Diese überaus hohe Ratewahrscheinlichkeit erhöht sich noch, wenn in den Items bereits irgendwelche Hinweise auf die zutreffende Lösung enthalten sind (s. o.). NUNNALLY (1964, 1972) weist darauf hin, daß der Testautor sich nicht in dem Glauben wiegen sollte, diese auf die Ratemöglichkeit zurückzuführenden Meßfehler durch eine der zahlreichen Korrekturformeln ausschalten zu können. Die einzige Möglichkeit, einen derartigen Meßfehler zu verringern, sieht er in einer Verlängerung des Tests über 60 Aufgaben, was wiederum andere Probleme mit sich bringt (z. B. Testzeit, Motivation, Ermüdung usw.). Eine andere Fehlerquelle liegt in der Wirksamkeit von sog. Reaktionstendenzen (response sets) des Schülers. Manche Schüler neigen dazu, solche Aufgaben, die sie nicht beantworten können, eher mit „Richtig" anzukreuzen, während andere eher ihr Kreuz bei „Falsch" machen. Solche Tendenzen lassen sich auch durch Instruktionen kaum ausschalten.

Diese und einige andere Nachteile, die Aufgaben mit Alternativ-Antworten mit sich bringen, lassen die Schlußfolgerung berechtigt erscheinen, Alternativ-Antworten nur sparsam und nur dann zu verwenden, wenn andere Aufgabenformen nicht eingesetzt werden können. Daran sollte auch die positive Haltung EBEL's diesem Aufgabentyp gegenüber nichts ändern.

Aufgaben mit Mehrfachwahlantworten (multiple-choice form). *a) Beschreibung.* Die Mehrfachwahlantwort ist die derzeit verbreitetste Form der Aufgabenbeantwortung. Ein Mehrfachwahl-Item setzt sich zusammen aus einem Problem und einer Reihe von vorgegebenen Antworten bzw. Lösungen. Das Problem, auch *„Stamm"* des Items genannt, hat die Form einer Frage oder einer unvollständigen Feststellung. Die vorgegebenen Antworten bzw. Lösungen werden als *„Alternativen"* bezeichnet, wobei die richtige Alternative einfach *„Antwort"*, die übrigen (falschen) Alternativen *„Distraktoren"* genannt werden. Wie wir sehen werden, stellen letztere ein spezielles Problem bei der Konstruktion dar.

Beispiele:

Wie nennt man die Zahl 48 in der Gleichung 6 x 8 = 48 ?

a) Dividend ☐
b) Multiplikand ☐
c) Produkt ☐
d) Quotient ☐

Südafrika ist führend in der Welt in der Gewinnung von

a) Bauxit ☐
b) Diamanten ☐
c) Eisenerz ☐
d) Kohle ☐

Wenn oben von einer „richtigen" Alternative (Antwort) unter den vorgegebenen die Rede war, so muß dabei noch differenziert werden zwischen *„richtiger Antwort"* und *„bester Antwort"*. Kann ein Mehrfachwahl-Item so konstruiert werden, daß außer einer einzigen Alternative alle anderen eindeutig falsch sind, so spricht man vom Typ „Richtige Antwort" (correct answer type). Ist es aber nicht möglich, aufgrund der im Stamm des Items gegebenen speziellen Fragestellung eine einzige richtige Alternative zu konstruieren, so wird man mehrere Alternativen vorgeben, die zwar alle richtig sind, aus denen aber der Schüler dann die beste auszuwählen hat (Typ „Beste Antwort" — best answer type). Einleuchtenderweise stellt die Konstruktion eines Mehrfachwahl-Items vom Typ „Beste Antwort" erhöhte Anforderungen an den Testautor.

Beispiel:

Welchem der folgenden Faktoren wird die meiste Aufmerksamkeit geschenkt, wenn eine Stadt zur Hauptstadt eines Landes bestimmt wird?

a) Lage ☐
b) Klima ☐
c) Autobahnen ☐
d) Bevölkerung ☐

b) Zweck. Zu seiner weiten Verbreitung mag entscheidend dazu beigetragen haben, daß das Mehrfachwahl-Item zur Überprüfung einer Vielzahl von Lernzielen geeignet ist (s. GRONLUND 1968; NUNNALLY 1964, 1972; EBEL 1965; GAUDE & TESCHNER 1970 u. a.). Nach GRONLUND können mit Mehrfachwahlaufgaben überprüft werden:

— Das Wissen des Schülers, also die Kenntnis von Fachausdrücken und spezifischen Fakten (etwa Fragen von der Form „Wo", „Was", „Wann", „Wer"), ferner die Kenntnis von Prinzipien, Regeln, Methoden usw.

— Das Verständnis des Schülers, und zwar die Fähigkeit, Kenntnisse und Regeln anzuwenden, Ursache-Wirkungs-Zusammenhänge zu interpretieren, bestimmte Vorgehensweisen und Methoden begründen zu können usw.

Mit dieser Aufzählung ist der Verwendungsbereich der Mehrfachwahlaufgabe nur annähernd umrissen. Wir werden weiter unten noch auf andere Möglichkeiten zur Verwendung von Mehrfachwahlaufgaben eingehen.

c) Konstruktionshinweise. Die optimale Nutzung der mit dem Mehrfachwahl-Item gegebenen Möglichkeiten hängt ab von der klaren Formulierung des Problems (Stamms), der Bildung plausibler, im strengen Sinne gleichwahrscheinlicher Alternativen und der Vermeidung von Hinweisen auf die zutreffende Antwort. Im einzelnen sollte man in Anlehnung an GRONLUND und NUNNALLY folgende Regeln beachten:

(1) Der *Stamm* des Items soll ein umschriebenes, bedeutsames Problem darbieten und alle notwendigen, aber keine irrelevanten Informationen enthalten. Dabei sollte eine negative Formulierung vermieden werden, soweit sie nicht sachlogisch erforderlich ist.

Beispiel:

```
Hans, ein stattlicher Junge mit guten Schulleistungen, mit
einem IQ von 105. Seine Intelligenz ist einzustufen als

a) hervorragend      ☐
b) genial            ☐
c) durchschnittlich  ☐
d) unterdurchschnittlich ☐

Besser:

Hans hat einen IQ von 105. Seine Intelligenz ist einzu-
stufen als
....
```

(2) Die vorgegebenen *Alternativen* sollen grammatikalisch mit dem Stamm übereinstimmen, damit nicht allein schon durch den Satzbau Hinweise auf die richtige Lösung gegeben werden. Dabei sollte nur eine Alternative die richtige oder beste Antwort enthalten.

Beispiel (nach INGENKAMP 1962):

Im Gegensatz zu WILSONS 14 Punkten wurden durch den Versailler Vertrag

a) das deutsche Heer abgerüstet ☐
b) das deutsche Kolonialreich zum größten Teil dem Völkerbund unterstellt ☐
c) deutsche Ostgebiete ohne Abstimmung abgetreten ☐
d) die Auslieferung der deutschen Kriegsflotte verlangt ☐

Hier kann die richtige Antwort (c) allein durch die grammatische Struktur von Stamm und richtiger Antwort erraten werden.

(3) Besondere Beachtung ist den *Distraktoren* (den falschen Alternativen) zu schenken.

Sie müssen alle die gleiche Plausibilität besitzen wie die richtige Antwort, denn ihr eigentlicher Sinn besteht ja darin, die Ratewahrscheinlichkeit zu vermindern. Sie sollen dem Schüler, der die Aufgabe aufgrund ungenügender Kenntnisse nicht beantworten kann, idealerweise alle als mögliche Lösungen erscheinen.

Das Auffinden solcher gleichwahrscheinlicher Distraktoren stellt aber den Testautor häufig vor erhebliche Schwierigkeiten. Vermag er diese Schwierigkeiten nicht zu überwinden und haben die Distraktoren eine unterschiedliche Wahrscheinlichkeit, die richtige Lösung zu sein, so wird das Ziel, ein Mehrfachwahl-Item zu konstruieren, verfehlt, und die Vorteile dieser Itemart werden verschenkt.

Beispiel:

Wer entdeckte den Nordpol?

a) Christoph Columbus ☐
b) Ferdinand Magellan ☐
c) Robert Peary ☐
d) Marco Polo ☐

Hier genügt es dem Schüler zu wissen, daß die in den Alternativen a), b), und d) genannten Männer lange Zeit vor der Nordpolentdeckung lebten, er findet die richtige Antwort auf dem Wege der Elimination.

(4) Des weiteren ist zu beachten, daß die richtige Antwort nicht immer an der gleichen Position der Alternativenliste steht, also nicht immer an erster, zweiter usw. Stelle. Die Positionen der richtigen Antworten sollten nach dem Zufall bestimmt werden.

(5) Auch das äußere Bild der richtigen Antwort sollte nicht konsistent unterschiedlich von dem der Distraktoren sein, da der Schüler damit unbeabsichtigte Hinweise auf die zutreffende Lösung erhält.

Beispiel:

Der Siedepunkt des Wassers beträgt

a) 80° C ☐
b) 95° C ☐
c) 100° C in einem offenen Behälter in Meereshöhe ☐
d) 70° C ☐

Besser:
Der Siedepunkt des Wassers in einem offenen Behälter in Meereshöhe beträgt:

a) 80° C ☐
b) 95° C ☐
c) 100° C ☐
d) 70° C ☐

(6) Alternativen von der Form „Keines von allen", „Alle sind richtig" zur Erhöhung der Schwierigkeit des Items sollten sparsam eingesetzt werden.

(7) Jedes Item sollte von jedem anderen Item unabhängig sein. Die Beantwortung einer Aufgabe sollte also einerseits nicht die Beantwortung einer später folgenden erleichtern, andererseits auch nicht von der richtigen Beantwortung einer vorhergehenden Aufgabe abhängig sein.

d) Vorteile. Wie oben erwähnt, erlaubt das Mehrfachwahl-Item einen weit größeren Bereich von Lernergebnissen abzuprüfen als die Alternativ-Antwort, wenn es sorgfältig konstruiert wurde. Die Ratewahrscheinlichkeit für die richtige Lösung wird entsprechend der Zahl der Alternativen verringert, Reaktionstendenzen der Schüler werden weitgehend ausgeschaltet.

Des weiteren besitzen diese Items eine hohe Auswertungsobjektivität, wobei die Auswertung mit Hilfe der EDV sehr ökonomisch gestaltet werden kann.

e) Nachteile. Die größten Schwierigkeiten bietet die Mehrfachwahlaufgabe bei der Konstruktion von Distraktoren. Wird dieses Problem nicht gelöst, dann werden aus Mehrfachwahl-Items sehr oft Alternativ-Antwort-Items (vgl. Abschn. 4.2.4.2. und 4.2.4.3.).

Die Nachteile, die sich aufgrund ungenügender Formulierung ergeben, können mit etwas Geschick vermieden werden.

Zur Prüfung produktiver, schöpferischer (kreativer) Leistungen sind allerdings auch die Mehrfachwahlaufgaben nur bedingt geeignet.

Trotzdem meint NUNNALLY (1964): „It is strongly recommended that the multiple-choice item be employed for most objective tests."

Aufgaben mit Ordnungsantworten (matching exercises). Die zu den gebundenen Antworten gehörenden Ordnungsantworten lassen sich aufgliedern in:
— Zuordnungsantworten
— Umordnungsantworten

Zuordnungsantworten. *a) Beschreibung.* Aufgaben mit Zuordnungsantworten bestehen aus zwei Reihen von Wörtern, Zahlen, Symbolen, Feststellungen u. ä. Die Symbole, denen ein anderes zugeordnet werden soll, heißen *Prämissen*, diejenigen, aus denen die Wahl getroffen wird, *Antworten*. Dabei müssen nicht alle Symbole aus der Reihe der Antworten verwendet werden — sog. „unvollständige Zuordnung".

Beispiel:

Spalte A	Spalte B
----- entdeckte den Pazifischen Ozean	a) Balboa
----- umsegelte als erster die Erde	b) Cortez
----- erforschte als erster Mexico	c) Hudson
----- eroberte Peru	d) Magellan
	e) Marquette
	f) Pizarro
	g) Ponce de Leon

b) Zweck. Aufgaben mit Zuordnungsantworten dienen vor allem der Messung von Faktenwissen, insbesondere solchen Wissens, wo es auf die Verknüpfung (Assoziation) zweier Tatbestände ankommt. Beispiele dafür liefert GRONLUND (1968):

Menschen (Männer)	---	Leistungen
Daten	---	Geschichtliche Ereignisse
Begriffe	---	Definitionen
Regeln	---	Beispiele
Symbole	---	Konzepte
Autoren	---	Buchtitel
Fremdwörter	---	Deutsche Äquivalente
Maschinen	---	Gebrauchsmöglichkeiten
Tiere, Pflanzen	---	Klassifikation
Objekte	---	Namen der Objekte
u. a.		

c) Konstruktionshinweise. (1) Die Liste der Prämissen muß inhaltlich homogen sein, und für jede Prämisse sollten mehrere plausible Antworten vorgegeben werden. Die Forderung nach Homogenität der Prämissen ist

bei weitem die wichtigste. Sie bedeutet, daß die Prämissen aus einem eng umgrenzten Wissensgebiet stammen und nicht aus mehreren (im Beispiel: Entdeckungen des 16. Jahrhunderts).
(2) Die Listen der Prämissen und Antworten sollten verschieden lang sein — die eine möglichst um 50 % länger als die andere —, um so die Ratemöglichkeit zu verringern. Welche Liste dabei länger ist, ist von untergeordneter Bedeutung, jedoch sollten beide Listen nicht zu lang sein.
In der Instruktion ist darauf hinzuweisen, daß die Antworten einmal, mehrmals oder manche überhaupt nicht benutzt werden können.
(3) Die Liste der Antworten sollte nach einem erkennbaren Prinzip aufgebaut sein, nach dem Alphabet, der Zeit, einer Zahlensequenz. Dies erleichtert dem Schüler die Übersicht, gibt ihm andererseits aber keine Hinweise auf die zutreffende Kombination.
(4) Den Aufgaben soll eine ausführliche Instruktion vorangestellt werden, die die Prozedur der Zuordnung erklärt. So wird eine Verwirrung des Schülers verhindert und außerdem Testzeit gespart.

Beispiel:

Auf die Linie links von jeder Feststellung in Spalte A ist der Buchstabe des Wortes aus Spalte B zu schreiben, das Deiner Meinung nach richtig ist.
Jedes Wort (Name usw.) der Spalte B kann einmal, mehr als einmal oder gar nicht verwendet werden.

d) Vorteile. Die Aufgaben mit Zuordnungsantworten erlauben die schnelle Prüfung eines umfangreichen Bereiches von aufeinanderbezogenen Tatsachenmaterials. Sie lassen sich außerdem für bestimmte Stoffgebiete (z. B. Geschichtsunterricht, Sprachunterricht) relativ leicht konstruieren, und sie sind objektiv auswertbar.
e) Nachteile. Diese Itemart ist weitgehend begrenzt auf die Überprüfung von Faktenwissen.
Ein schwerwiegendes Problem ist die Wahrung der Homogenität der Prämissen, ein anderes (ähnlich wie bei den Mehrfachwahlaufgaben) die Bildung einer ausreichenden Zahl von plausiblen Antworten. Jede richtige Antwort für eine Prämisse muß ja noch als plausible Antwort für die anderen Prämissen gelten können.

Aufgaben mit Umordnungsantworten. *a) Beschreibung.* Eine weniger verwendete Variante der Ordnungsantwort ist die Unordnungsantwort. Wir wollen deshalb nur kurz auf sie eingehen.
Dem Schüler werden hier eine Reihe von Wörtern, Begriffen, Zahlen, Symbolen o. ä. vorgegeben. Er muß sie dann nach einem bestimmten Prinzip in eine sinnvolle Reihe umordnen, z. B. einen Satz bilden, Länder nach der Größe ordnen usw.

Beispiel:

Ordne die gegebenen Länder Europas nach ihrer Flächengröße (1 = größtes Land, 6 = kleinstes Land)

a) Deutschland (BRD)	1._____
b) England	2._____
c) Frankreich	3._____
d) Norwegen	4._____
e) Italien	5._____
f) Spanien	6._____

b) Zweck. Am ehesten angemessen ist dieser Itemtyp für die Prüfung von Faktenwissen.
c) Konstruktionshinweise. Die Liste der Begriffe, Symbole usw. sollte nicht zu lang sein, der Schüler verliert sonst die Übersicht.
d) Vorteile und *e) Nachteile.* Bei geschicktem Einsatz lassen sich auch kompliziertere Denkvollzüge erfassen. Jedoch schlagen bei dieser Itemart der große Platzbedarf, vor allem aber die Schwierigkeit einer angemessenen Bewertung negativ zu Buche. So erscheint es sinnvoll, diese Aufgabenart nur in besonderen Fällen einzusetzen.

4.2.2.2.2. Aufgaben mit nicht-gebundenen Antworten

Charakteristisches Merkmal dieser Aufgabenart ist, daß dem Schüler nicht zwei oder mehrere Antworten zur Auswahl bzw. Zu- oder Umordnung vorgegeben werden, sondern daß er selbst eine Antwort auf eine Frage formulieren oder einen Satz bzw. einen Teil eines Satzes ergänzen muß. Die Problematik dieser Items liegt weniger in ihrer Konstruktion als in der erreichbaren Objektivität der Bewertung der Antworten. Die folgende Aufzählung dieser Aufgaben stellt gleichzeitig eine Rangreihe abnehmender Objektivität dar.
Es lassen sich folgende Aufgabenarten mit nicht gebundenen oder besser, nicht vorformulierten, Antworten unterscheiden:
Aufgaben mit
— Ergänzungs- bzw. Kurzantworten (fill in bzw. short answer response),
— Kurzaufsatzantworten (essay tests).

Aufgaben mit Ergänzungs- bzw. Kurzantworten. *a) Beschreibung.* Ergänzungsantworten und Kurzantworten unterscheiden sich voneinander lediglich durch die Art der Problemstellung in der Aufgabe. Im ersteren Fall muß eine unvollständige Feststellung (mit Lücken) durch ein Wort, eine Zahl, ein Symbol o. ä. ergänzt werden. Im zweiten Fall hat der Schüler eine direkt gestellte Frage mit einem Wort, einer Zahl, einem Namen, einer kurzen Aussage, einem Symbol o. ä. zu beantworten.

Beispiele:

Ergänzungsantwort

Kolumbus entdeckte Amerika im Jahre - - - - -

Wenn $\frac{x}{b} = \frac{3}{b-1}$, dann ist x = - - - - -

Kurzantwort

Welches ist der Fachausdruck für Kurzsichtigkeit? _____

Wer umsegelte als erster die Welt? _____

b) Zweck. Im wesentlichen lassen sich mit diesem Aufgabentyp Wissenstatbestände überprüfen, wie die Kenntnis von Begriffen, spezifischen Fakten, Regeln, Methoden und einfache Interpretationen von Daten.
Komplexe Leistungen können allerdings erfaßt werden, wenn die Kurzantwort-Aufgabe zur Interpretation von Diagrammen, Karten, Kurven und Bildmaterial verwendet wird (s. a. Abschn. 4.2.2.2.3.). Insbesondere bei mathematischen und naturwissenschaftlichen Aufgaben, wo die Lösung durch Zahlen oder Symbole ausgedrückt werden kann und damit objektiv bewertbar ist, können mit diesem Aufgabentyp durchaus auch komplexe Lern- und Denkvorgänge untersucht werden. GRONLUND (1968) ist der Meinung, daß gerade bei mathematischen und naturwissenschaftlichen Problemen die Kurzantwort-Aufgabe vor anderen zu bevorzugen ist.
c) Konstruktionshinweise. (1) Das Kurzantwort-Item sollte so formuliert werden, daß es nur eine kurze Antwort erfordert und nur eine einzige Antwort die richtige ist. Verständlicherweise ist die letzte Forderung diejenige, die bei der Konstruktion die größten Schwierigkeiten bereitet.
Schlechtes Beispiel:
Sauerstoff ist wichtig für _____.
Mögliche richtige Antworten sind hier „Verbrennung", „Atmung", „Taucher" u. a.
Besser wäre:
Welches der atmosphärischen Gase ist wichtig für die Verbrennung? _____

(2) Es sollten bei der Ergänzungsantwort nur ein oder zwei Lücken gelassen werden, wobei nur wichtige Begriffe ausgelassen werden sollten.
Schlecht:
Im Jahre 1492 _____ Kolumbus Amerika.
Besser:
Kolumbus entdeckte Amerika im Jahre _____
Ebenfalls verbesserungsbedürftig ist folgendes Item:
Die meisten Autounfälle werden verursacht durch _____, _____ und _____.

(3) Eine direkte Frage ist einer unvollständigen Feststellung gegenüber vorzuziehen. Die Frageform entspricht eher den bisherigen Schulerfahrungen des geprüften Schülers, die Situation ist eindeutiger strukturiert.
(4) Wird das Kurzantwort-Item verwendet, dann sollten die Lücken für die einzusetzenden Antworten gleich lang sein und möglichst am rechten Rand der Seite des Testheftes aufgeführt werden; s. die Beispiele unter (a).
d) Vorteile. Die Möglichkeit, die richtige Lösung allein aufgrund von Raten zu finden, ist hier weitgehend reduziert. Darüberhinaus sind solche Aufgaben relativ leicht zu konstruieren.
e) Nachteile. Mit diesem Aufgabentyp kann im wesentlichen nur Faktenwissen abgeprüft werden (Ausnahmen s. o.).
Bedeutsame Schwierigkeiten ergeben sich bei der Aufgabenauswertung. Wenn das Problem nicht ganz exakt formuliert wurde, können verschiedene Antworten gegeben werden, die zwar nicht diejenigen sind, die der Testautor erwartete, die andererseits aber auch nicht falsch sind. Diese Einschränkung der Auswertungsobjektivität entfällt allerdings bei Aufgaben mathematischer oder naturwissenschaftlicher Art, wo die Antwort aus einer Zahl oder einem Symbol besteht.

Aufgaben mit Kurzaufsatzantwort (Essay-Test). Wie GAUDE & TESCHNER (1970) formulieren, versteht man darunter solche Aufgaben, bei denen der Schüler auf eine Frage, Anweisung, Skizze, Tabelle usw. mit einer verbalen oder nicht-verbalen Antwort (Zeichnung, graphische Darstellung) reagieren muß. Der Schüler ist also weitgehend frei in der Formulierung seiner Antwort.
GRONLUND (1968) unterscheidet noch Aufgaben mit eingeschränkter Antwort (restricted response questions) und freier Antwort (extended response questions). Die entscheidende Schwierigkeit dieser Aufgaben, die hauptsächlich zur Erfassung komplexerer Leistungen einsetzbar sind, liegt in der mangelnden Objektivität ihrer Bewertung. Da solche Aufgaben nur in Ausnahmefällen in objektive Tests aufgenommen werden sollten, wollen wir hier auf eine ausführliche Diskussion verzichten. Der interessierte Leser mag sich bei GRONLUND 1968, GAUDE & TESCHNER 1970, EBEL 1965, NUNNALLY 1972 u. a. informieren.

4.2.2.2.3. Aufgaben zur Messung komplexer Leistungen

Wie bei den verschiedenen Aufgabentypen bereits vermerkt wurde, eignen sich einige von ihnen auch zur Erfassung komplexer Leistungen. So konnten Kurzantwort-Items für Problemlösungsaufgaben aus Mathematik und Naturwissenschaften, Alternativ-Antwort-Aufgaben zur Erfassung von Ursache-Wirkungs-Zusammenhängen eingesetzt und Mehrfachwahlaufgaben

zur Überprüfung des Verständnisses von Unterrichtsstoff benutzt werden. Dabei mußte man sich jedoch jeweils auf Einzelitems stützen.

Als weitere Möglichkeit, komplexe Leistungen zu messen, haben wir die Essay-Tests erwähnt, jedoch zugleich auf die Bewertungsprobleme hingewiesen, die den Einsatz solcher Tests drastisch einschränken.

Eine andere Form von Aufgaben, die sich zur Erfassung komplexer Leistungen u. E. recht gut eignen, stellen die auf R. W. TYLER (1946) zurückgehenden sog. „interpretive exercises" (Interpretationsübungen) dar. Variationen von ihnen werden auch als „classification exercise", „keytype items" oder „master-list items" bezeichnet.

Die Interpretationsübungen (interpretive exercises). *a) Beschreibung.* Bei diesen Aufgaben wird dem Schüler ein bestimmter Datensatz in Form von geschriebenem Material, Tabellen, Karten, Zeichnungen, Bildern o. ä. vorgegeben.

Der Schüler hat nun eine Reihe von Aufgaben zu bearbeiten, die ihm zu dem vorgelegten Material gestellt werden (er muß gewisse Schlußfolgerungen ziehen, Beziehungen zwischen den Daten herausarbeiten, aufgrund dieser Daten gemachte Annahmen auf ihre Gültigkeit überprüfen usw.).

Die Aufgaben haben in der Regel die Form von Mehrfachwahlaufgaben oder Aufgaben mit Alternativantworten, sie sind also objektiv auswertbar. Der dem Schüler vorgelegte Datensatz kann hinsichtlich seines Informationsgehaltes variiert werden, je nachdem, welches Lernziel abgeprüft werden soll.

Beispiel (nach GRONLUND 1968):

(Die folgende Aufgabe soll insbesondere prüfen, daß der Schüler in der Lage ist, zu erkennen, ob Schlußfolgerungen und Verallgemeinerungen, die aufgrund eines gegebenen Datenmaterials getroffen wurden, gerechtfertigt sind oder nicht.)

Tab. 1: Sterblichkeit durch Autounfälle unter der weißen Bevölkerung in den USA von 1957 bis 1958

Altersgruppen	Sterblichkeitsziffer pro 100 000	
	Männer	Frauen
Alle Altersstufen zusammen	32.9	11.1
1-4	10.5	8.0
5-14	10.4	5.4
15-19	54.2	16.4
20-24	76.3	12.7
25-44	35.6	9.1
45-64	33.1	12.9
65 und älter	58.4	22.5

Instruktion:

Die folgenden Feststellungen beziehen sich auf die Zahlen in der oben abgebildeten Tabelle. Lies jede Feststellung sorgfältig durch und kreuze "R" an, wenn die Feststellung durch die Zahlen in der Tabelle gestützt wird;
kreuze "F" an, wenn die Feststellung durch die Zahlen in der Tabelle widerlegt wird;
kreuze "W" an, wenn die Feststellung durch die Zahlen in der Tabelle weder gestützt noch widerlegt wird.

	R	F	W
1. Es sterben mehr Männer durch Autounfälle als Frauen.	☒	☐	☐
2. Autounfälle sind eine Haupttodesursache bei jungen Männern im Alter von 20 bis 24 Jahren.	☐	☐	☒
3. Männer über 65 Jahre fahren nicht sicherer als junge Burschen zwischen 15 und 19 Jahren.	☐	☐	☒
4. Die meisten der durch Autounfälle getöteten Personen sind 65 Jahre und älter.	☐	☒	☐
5. Wenn man alle Altersgruppen zusammenfaßt, können nur ungefähr 11 Prozent der Todesfälle bei Frauen auf Autounfälle zurückgeführt werden.	☐	☒	☐

b) Zweck. Wie eingangs erwähnt, sollen diese Interpretationsübungen vor allem zur Überprüfung komplexer Leistungen wie Verständnis, schlußfolgerndes Denken, kritisches Denken, Kreativität, Problemlösen usw. herangezogen werden. Weitere Beispiele für den Gebrauch der Interpretationsübungen finden sich bei GRONLUND (1968). Er führt vor allem folgende Lernziele an, die mit Hilfe der Interpretationsübungen gemessen werden können:

— die Fähigkeit, Regeln anzuwenden
— die Fähigkeit, Beziehungen zu interpretieren
— die Fähigkeit, Schlußfolgerungen als solche erkennen und selbst welche ziehen können
— die Fähigkeit, die Relevanz von Informationen abzuschätzen
— die Fähigkeit, haltbare Hypothesen zu entwickeln
— die Fähigkeit, gültige Schlüsse zu ziehen und den Schlüssen zugrunde liegenden Annahmen zu erkennen
u. a.

c) Konstruktionshinweise. Die Konstruktion von Interpretationsübungen umfaßt zwei Hauptabschnitte:

— die Auswahl des Informationsmaterials,
— die Konstruktion von Aufgaben zu diesem Material.
(1) Das Informationsmaterial sollte sich inhaltlich niveaumäßig auf der Ebene des Unterrichtsstoffes bewegen, wobei das Material jedoch für die Schüler neu sein sollte, um zu verhindern, daß lediglich auswendig Gelerntes reproduziert wird. Um den Einfluß der Lesefertigkeit der Schüler gering zu halten, sollten die Informationen nicht zu lang, präzise und klar sein.
(2) Bei der Konstruktion der Aufgaben ist darauf zu achten, daß ihre Beantwortung nicht bereits im Material enthalten ist, sondern daß sie eine Interpretation der vorgegebenen Daten erfordert. Dabei sollte die Zahl der zu beantwortenden Aufgaben in einem vernünftigen Verhältnis zur Menge der Informationen stehen. Im übrigen sind natürlich all die Hinweise zu beachten, die für die Konstruktion objektiver Testitems gegeben wurden.
d) Vorteile. Mit Hilfe der Interpretationsübungen können, wie bereits festgestellt, komplexere Lernprozesse einer Prüfung unterzogen werden. Hierbei handelt es sich vorwiegend um solche Lernziele, die für den Schüler sowohl in der Schule als auch im täglichen Leben immer mehr an Bedeutung gewinnen, z. B. das Lesen- und Interpretierenkönnen von Tabellen, Problemlösen und so fort. Es ist möglich, diese Lernziele unabhängig davon zu prüfen, ob der Schüler bestimmte konkrete Einzelfakten präsent hat, da ihm diese ja mit dem Informationsmaterial geliefert werden. Man ist also in der Lage, etwa schlußfolgerndes Denken per se auch dann zu prüfen, wenn dem Schüler die Detailinformationen zu Beginn des Tests fehlen. Bei einer entsprechenden einzelnen Mehrfachwahlaufgabe beispielsweise würde er aufgrund dieses fehlenden Faktenwissens zwangsläufig versagen müssen und damit einen falschen Eindruck hinsichtlich seiner Fähigkeiten hinterlassen.
e) Nachteile. Hauptschwierigkeiten liegen bei der Konstruktion von Interpretationsübungen in der Zusammenstellung des Informationsmaterials und in der Formulierung der zu beantwortenden Aufgaben. Beides erfordert relativ viel Zeit und Geschick. Ein weiteres Problem ist das der Lesegeschwindigkeit der einzelnen Schüler. Um nicht die Lesefertigkeit unbeabsichtigt zum eigentlichen Prüfungsgegenstand zu machen und den Schüler mit schwachen Leseleistungen zu benachteiligen, sollten die Informationen relativ einfach und kurz gehalten werden.

4.2.3. Entwicklung der Test„vorform"

4.2.3.1. Gruppierung der Items und Erstellung des Testheftes

Nachdem der Testkonstrukteur die in der Spezifikationstabelle vorgesehene Zahl von Aufgaben konstruiert hat, müssen die Items in übersichtlicher Form so zusammengestellt werden, daß sie den Schülern zur Beantwor-

tung vorgelegt werden können. Dazu eignet sich am besten das sog. Testheft.

Ein *Testheft* besteht quasi aus zwei Teilen. Der erste Teil (meist die Vorderseite des Heftes) enthält die *Bezeichnung des Tests, Angaben* zur *Person* des Schülers (Name, Vorname, Klasse, Schule usw.) und die *Instruktion* mit Lösungsbeispielen. Im zweiten Teil folgen dann die vom Schüler zu lösenden Aufgaben.

Besondere Bedeutung kommt dabei der *Testinstruktion für den Schüler* zu. Zunächst sollte in wenigen Sätzen das *Ziel dieses Tests* erläutert werden, um dem Schüler den Zusammenhang des Tests mit dem bisherigen Unterrichtsverlauf deutlich zu machen. Des weiteren muß dem Schüler erklärt werden, wie er bei der Lösung der gestellten Aufgaben vorzugehen hat, d. h. es muß ihm die *Art der Beantwortung der Aufgaben* hinreichend deutlich gemacht werden. Hierbei ist zunächst verbal zu schildern, was der Schüler zu tun hat, also z. B. die Kästchen neben den richtigen Alternativen anzukreuzen, Umordnungen vorzunehmen usw. Darauf folgen zwei oder drei *Beispiele* mit Lösungen für *jede* im Test verwendete *Aufgabenart.*

Beispiele für Instruktionen:

Mehrfachwahlantwort
Lies jede der folgenden Aufgaben gründlich durch. Zu jeder Aufgabe sind 4 Antworten gegeben. Nur eine davon ist richtig. Die richtige Antwort soll herausgefunden und angekreuzt werden.
... Lösungsbeispiele ...

Alternativantwort (R-F-Typ)
Lies jede der folgenden Aufgaben sorgfältig durch. Wenn die Feststellung richtig ist, kreuze das „R" an. Wenn die Feststellung falsch ist, kreuze das „F" an.
... Lösungsbeispiele ...

In der Instruktion ist dem Schüler auch mitzuteilen, welche *Hilfsmittel* (Lineal, Konzeptpapier, Lexika usw.) er bei dem Test verwenden darf. Wichtig ist ferner die Angabe der *Bearbeitungszeit,* also der Zeit, die dem Schüler zur Lösung der Aufgaben zur Verfügung steht (in der Regel eine Schulstunde). Wieviel Zeit die Schüler tatsächlich für den Test benötigen, kann vor der ersten Testdurchführung nur grob geschätzt werden. Die erforderliche Bearbeitungszeit ist u. a. abhängig vom Alter der Schüler, der Zahl der Aufgaben und der erforderlichen Schreibarbeit (etwa bei Kurzantwort- oder Ergänzungsantwort-Items). NUNNALLY (1962) hält für einen Test mit 40 Mehrfachwahlaufgaben mit fünf Alternativen bei älteren Schülern eine Testzeit von 50 Minuten für angemessen. Als allgemeine Regel gilt, daß in der vorgegebenen Zeit *90 % der Schüler* die Aufgaben ohne zu großen Zeitdruck sollten lösen können. Die Zahl der Aufgaben pro Test sollte nicht mehr als 50—60 Items betragen.

Im zweiten Teil des Testheftes folgen dann die zu lösenden *Aufgaben*. Bei der Zusammenstellung der Aufgaben sind einige wichtige Punkte zu beachten.

In einem Test sollten möglichst nicht mehr als zwei verschiedene Aufgabenarten verwendet werden. Werden mehrere Aufgabentypen benutzt, dann hat das einmal zur Folge, daß die Instruktion unnötig lang wird, da ja für jeden Aufgabentyp Lösungsbeispiele gebracht werden müssen. Zum anderen erfordert ein häufiger Wechsel des Aufgabentyps ein ständiges Umdenken und Sich-Umstellen-Müssen des Schülers. Das kostet Zeit, verwirrt den Schüler und kann sich damit nachteilig auf seine Testleistung auswirken.

Bei der Reihung der Aufgaben ist es günstig, an den Anfang des Tests einige leichtere Aufgaben zu stellen („Eisbrecher"). Der Schüler wird dadurch nicht von vornherein entmutigt, am Test weiterzuarbeiten. Im allgemeinen gilt für die dann folgenden Aufgaben die Regel, daß sie entsprechend ihrem Schwierigkeitsgrad ansteigend angeordnet werden, also von leichten hin zu schweren Aufgaben am Ende des Tests. (Vor Durchführung der Itemanalyse [s. Abschn. 4.2.4.] kann der Schwierigkeitsgrad natürlich vom Lehrer nur geschätzt werden.) Würde man die leichten und schweren Aufgaben im Test z. B. nach dem Zufall anordnen, dann könnte der Fall auftreten, daß sich ein Schüler zu lange bei einer vorne stehenden schwierigen Aufgabe aufhält und gar nicht erst zu jenen weiter hinten stehenden leichteren Aufgaben, die er lösen könnte, vordringt. Werden in einem Test mehrere Stoffgebiete gleichzeitig abgeprüft, dann sollte für jedes Stoffgebiet die Aufgabenreihung getrennt in der beschriebenen Weise vorgenommen werden. Allerdings muß dann die Testzeitbegrenzung für jedes Stoffgebiet gesondert angegeben werden, also etwa für das Stoffgebiet A 10 Minuten, für B 20 Minuten usw. Der Schüler muß in der Instruktion hierauf gesondert hingewiesen werden. Im übrigen sind die Aufgaben im Testheft möglichst übersichtlich und in einheitlicher Schreibweise darzustellen, z. B.:

Aufgaben-nummer	Aufgabe	Alternativen	Raum zum Ankreuzen
1	.....	.....	☐
		.....	☐
		.....	☐
		.....	☐
		.....	☐
2	.....	.....	☐
		.	
		.	
.		. .	
.			
.			

Ökonomisch ist es, die Schüler die richtigen Antworten nicht in das Testheft selbst eintragen zu lassen, sondern dafür einen gesonderten Antwortbogen (s. Abb. 3) zu benutzen, auf dem die Schüler die richtigen Antworten ankreuzen. Das hat den Vorteil, daß die Testhefte mehrfach verwendbar sind und die Auswertung der Testergebnisse erleichtert wird (z. B. mittels einer Lösungsschablone oder über eine elektronische Datenverarbeitungsanlage).

Abb. 3: Beispiel für einen Antwortbogen bei Mehrfachwahlaufgaben

-Antwortbogen-

Name Vorname

Klasse Datum Fach

Lies die Anweisungen im Testheft sorgfältig durch und beachte sie genau. Kreuze für jede Aufgabe den Buchstaben im Antwortbogen an, der die richtige Lösung bezeichnet. Achte darauf, daß die Aufgabennummer im Testheft mit der Aufgabennummer im Antwortbogen übereinstimmt.
Beispiel:
Bei einer Aufgabe ist „B" die richtige Antwort.
In diesem Falle wird wie folgt auf dem Antwortbogen angekreuzt:

Aufgabennummer	
...	A X C D

Aufgabennummer	Richtige Antwort	Aufgabennummer	Richtige Antwort
1.	A B C D	21.	A B C D
2.	A B C D	22.	A B C D
3.	A B C D	23.	A B C D
4.	A B C D	24.	A B C D
.	. . . .	.	. . . .
.	. . . .	.	. . . .

a) Bewertung. Es ist sinnvoll, für jede richtig gelöste Aufgabe einen Punkt (1), für jede falsche Lösung null Punkte (0) zu geben und auf eine gewichtete Bewertung zu verzichten. Die Höchstpunktzahl ist dann immer gleich der Zahl der Aufgaben im Test. Bei Ergänzungsantworten kann für jede richtig ergänzte Lücke ein Punkt gegeben werden. Das gleiche gilt für jede richtige Zuordnung bei Ordnungsantworten.

b) Auswertung. Zur leichteren und schnelleren Auswertung empfiehlt es sich, eine Auswertungsschablone herzustellen. Dazu schneidet man aus einem Blatt Kartonpapier, das die Größe des Antwortbogens hat und auf dem die Buchstaben der richtigen Lösungen an die gleiche Stelle eingetragen wurden, an der sie im Antwortbogen stehen, die Kästchen mit den richtigen Lösungen heraus. Durch Auflegen dieser Schablone auf den Antwortbogen läßt sich sehr schnell die Zahl der richtigen Lösungen feststellen. Jede falsche Antwort wird (durch das Loch in der Schablone) mit einem Farbstift gekennzeichnet. Der *Gesamttestwert* des Schülers ergibt sich, indem man von der Gesamtzahl der (bearbeiteten) Items die Zahl der falschen (farbig markierten) abzieht.

Auf die Korrektur des Testwertes mit Hilfe von Korrekturformeln bei Aufgaben mit Ratemöglichkeiten sollte verzichtet werden. Eine ausführliche Diskussion dieses Problems findet sich bei NUNNALLY (1972).

4.2.4. Die Itemanalyse

Nachdem die Aufgaben zusammengestellt, ihre Bewertung festgelegt und die Instruktion formuliert wurden, kann der Test zur ersten praktischen Erprobung eingesetzt werden. Man läßt den Test von einer Stichprobe des Schülerkreises ausfüllen, für den man späterhin diesen Test anzuwenden gedenkt. Damit hat man bereits genügend Daten gewonnen, um den Test einer ersten empirischen Analyse unterziehen zu können.

Diese erste Analyse, die *Aufgaben-* oder *Itemanalyse,* hat dreierlei Ziele:

— festzustellen, wie schwierig die einzelnen Aufgaben für die Schüler sind (Schwierigkeitsgrad),
— festzustellen, ob sich die Aufgaben eignen, zwischen guten und schlechten Schülern zu unterscheiden, also zwischen solchen, die den Unterrichtsstoff weitgehend gelernt haben und solchen, die das nicht oder nur in geringerem Ausmaße getan haben (Trennschärfe),
— festzustellen, ob bei Mehrfachwahlaufgaben die Distraktoren plausible Alternativen zu richtigen bzw. besten Antworten darstellen (Distraktorenanalyse).

Man bezeichnet diese Phase der Testerprobung auch mit „testing the test".

4.2.4.1. Berechnung des Schwierigkeitsgrades einer Aufgabe

Der Schwierigkeitsgrad (Schwierigkeitsindex) P eines Items ist definiert als der Prozentsatz einer bestimmten Stichprobe von Schülern, die die Aufgabe löst.

Also:

$$P = \frac{N_R}{N} \cdot 100$$

N_R = Anzahl der Schüler, die die Aufgabe richtig gelöst haben;
N = Gesamtzahl der Schüler.

Beispiele:
Von 120 Schülern haben 102 Schüler eine Testaufgabe richtig gelöst. Der Schwierigkeitsgrad für diese Aufgabe errechnet sich:

$$P = \frac{102}{120} \cdot 100 = 85$$

Eine andere Testaufgabe wurde in einer Stichprobe von 100 Schülern von 50 Schülern richtig gelöst.

$$P = \frac{50}{100} \cdot 100 = 50$$

Der Schwierigkeitsindex wird also um so größer, je mehr Schüler die Aufgabe gelöst haben. Eine Aufgabe mit dem Schwierigkeitsgrad 0 ist also sehr schwer — sie wurde von keinem Schüler gelöst—, eine Aufgabe mit dem Index 100 ist sehr leicht — sie wurde von allen Schülern richtig beantwortet.

Eine Aufgabe sollte nur dann im Test weiter verwendet werden, wenn ihr *Schwierigkeitsindex zwischen 20 und 80* liegt. Leichtere Items ($P>80$) sollten ausgeschieden oder nur zu Beginn des Tests als „Eisbrecher" eingesetzt werden.

Es ist wichtig zu beachten, daß ein Schwierigkeitsindex immer nur für die untersuchte Gruppe Gültigkeit hat. Es gibt also keine Schwierigkeit des Items per se. Man kann z. B. nicht sagen, die Schwierigkeit eines Items ist $P = 68$, sondern nur, die Schwierigkeit dieses Items für die Gruppe X ist $P = 68$. Soll eine Aufgabe in einen Test für eine andere Schülerpopulation aufgenommen werden, dann ist ihr Schwierigkeitsgrad erneut zu bestimmen.

4.2.4.2. Berechnung der Trennschärfe einer Aufgabe

Die Trennschärfe eines Items bedeutet den Grad, mit dem die Aufgabe zwischen guten (leistungsstarken) und schlechten (leistungsschwachen) Schülern trennen (unterscheiden) kann. Eine Aufgabe ohne jede Trennschärfe würde

von guten und schlechten Schülern in gleicher Weise beantwortet werden und könnte dem Lehrer damit keine Hilfe für seine Leistungsbeurteilung bieten.

Man wird einwenden, daß die Trennschärfe einer Aufgabe ja wohl von ihrer Schwierigkeit abhinge, eine schwere Aufgabe beispielsweise könnte eben nicht von den schlechten Schülern gelöst werden und sei deshalb zwangsläufig trennschärfer als eine sehr leichte Aufgabe, die eben von allen Schülern richtig beantwortet wird. Tatsächlich besteht ein Zusammenhang zwischen Schwierigkeit und Trennschärfe, jedoch ist dies nicht ein linearer, sondern ein paraboloider. Wir werden weiter unten näher darauf eingehen (s. Abschn. 4.2.5.).

Für die Berechnung der Trennschärfe gibt es eine Reihe verschiedener Möglichkeiten. Wir wollen hier nur eine sehr einfache, für die Konstruktion von informellen Tests gut verwendbare Methode darstellen.

Die Antwortbogen der getesteten Schüler werden so zu einem Stapel geordnet, daß oben der Antwortbogen des Schülers mit dem höchsten Gesamttestwert liegt, darunter der mit dem zweithöchsten Gesamttestwert, während der Bogen mit dem niedrigsten Gesamttestwert ganz unten liegt. Die Antwortbogen sind also in eine Rangreihe gebracht worden: oben der höchste, unten der niedrigste Rang.

Nun werden diesem Stapel die oberen 25 % (Obergruppe) und die unteren 25 % (Untergruppe) der Antwortbogen entnommen. Bei 100 Schülern werden also 25 Bogen von oben her und 25 Bogen von unten her entnommen. Jetzt wird für jedes Item die Zahl der richtigen Antworten in der Ober- bzw. Untergruppe ausgezählt (s. Tab. 2).

Beispiel:

Tab. 2: Berechnung der Trennschärfe

Item Nr.	Zahl der richtigen Antworten in der		Trennschärfeindex
	Obergruppe (R_O)	Untergruppe (R_U)	T i
1	18	10	.32
2	5	15	-.40
3	17	2	.60
.	.	.	.
.	.	.	.
.	.	.	.

Die ausgezählten Häufigkeiten werden in folgende Formel eingesetzt:

$$Ti = \frac{R_O - R_U}{\frac{N}{4}}$$

Es bedeuten:
Ti = Trennschärfeindex,
R_O = Zahl der richtigen Antworten in der Obergruppe,
R_U = Zahl der richtigen Antworten in der Untergruppe,
N = Gesamtzahl der getesteten Schüler (Gesamtstichprobe).

Im Beispiel ergibt sich für Item 1:

$$Ti = \frac{18-10}{25} = \frac{8}{25} = .32$$

Für Item 2 und 3 wird errechnet:

Item 2: $Ti = \frac{5-15}{25} = \frac{-10}{25} = -.40$

Item 3: $Ti = \frac{17-2}{25} = \frac{15}{25} = .60$

Wie die Beispiele zeigen, wird die Trennschärfe um so größer, je unterschiedlicher die Zahl der richtigen Antworten in Ober- und Untergruppe ist (vgl. Item 1 und 3). In der Regel werden die Trennschärfeindizes positiv sein. Ein negativer Trennschärfeindex (vgl. Item 2) kommt zustande, wenn mehr Schüler der Untergruppe als Schüler der Obergruppe die Aufgabe richtig beantworten. Solche Aufgaben sind oft unklar formuliert, extrem schwer oder leicht oder mit schlechten Alternativen versehen. Solche Aufgaben sind entweder neu zu konstruieren oder (besser) wegzulassen. EBEL (1965) klassifiziert die Items hinsichtlich ihrer *Trennschärfe* wie folgt:

.40 und größer	sehr gute Items
.30 bis .39	gute Items, aber möglicherweise zu verbessern
.20 bis .29	Items, die einer Verbesserung bedürfen
.19 und kleiner	schlechte Items, wegzulassen oder durch Revision zu verbessern.

Items mit sehr niedrigen Trennschärfeindizes ($Ti \leqq .20$) leisten also kaum einen Beitrag zur Differenzierung von guten und schlechten Schülern.

4.2.4.3. Distraktorenanalyse

Zur Analyse der Distraktoren sind nicht unbedingt Berechnungen notwendig, obwohl man analog zum Trennschärfeindex vorgehen könnte.

Von einem guten Distraktor muß man erwarten, daß er eher von den schlechten Schülern für die richtige Antwort gehalten wird als von den guten.

Beispiel:
Item x

Alternativen	*A*	*B*	*C*	*D**
Obergruppe (N=12)	5	5	0	2
Untergruppe (N=12)	3	4	0	5

* A ist die Antwort,
B, C und D sind Distraktoren.

Distraktor B wird eher von den besseren Schülern als richtige Antwort angekreuzt, widerspricht also dem Ziel eines Distraktors. Distraktor C ist völlig unzureichend, er wird überhaupt nicht gewählt, die Vierfach-Wahlaufgabe reduziert sich damit auf eine mit drei Wahlmöglichkeiten. Distraktor D erfüllt die Ansprüche an einen guten Distraktor durchaus, er „verleitet" eher die schlechten Schüler, ihn für die richtige Antwort zu halten als die guten.

Das Ergebnis dieser Analyse bedeutet also, daß diese Aufgabe vor Aufnahme in den endgültigen Test einer erneuten Bearbeitung unterzogen werden muß.

4.2.5. Itemselektion und -revision

Auf die Itemanalyse folgt nun die Auswahl der für den endgültigen Test zu verwendenden Items. Die mittels der Itemanalyse erhaltenen Daten erlauben eine Abschätzung der Brauchbarkeit der ursprünglich formulierten Aufgaben. Items mit zu hohem oder zu niedrigem Schwierigkeitsgrad bzw. Trennschärfeindex können, wie aus den obigen Betrachtungen abzuleiten ist, für die Testendform nicht verwendet werden. Ebenso müssen solche Aufgaben entfallen, bei denen die Distraktorenanalyse aufgezeigt hat, daß einige der Alternativen keine plausible Lösungsmöglichkeit darstellen. Wie wir oben angeführt haben (s. Abschn. 4.2.4.1.), sollten nur solche Aufgaben im Test belassen werden, deren Schwierigkeitsindex zwischen 20 und 80 liegt (s. auch NUNNALLY 1972, S. 186 ff.). Dabei ist allerdings darauf zu achten, daß vor allem auch genügend Aufgaben mittlerer Schwierigkeit ($P \approx 50$) im Test vertreten sind. Bei der Auswahl der Aufgaben genügt es jedoch nicht, als Auswahlkriterium allein den Schwierigkeitsgrad heranzuziehen. Es kann vorkommen, daß eine Aufgabe, deren Schwierigkeit im angegebenen Bereich liegt, eine zu geringe Trennschärfe aufweist und deshalb unbrauchbar ist. Gleiches gilt umgekehrt. Eine Aufgabe mit hinreichender Trennschärfe kann zu schwierig oder zu leicht sein.

Wie oben angedeutet, besteht ein Zusammenhang zwischen Schwierigkeitsgrad und Trennschärfe. Dieser Zusammenhang ist jedoch nicht linear, eine Aufgabe ist also nicht um so trennschärfer, je schwieriger sie ist. Vielmehr besteht zwischen Schwierigkeit und Trennschärfe eine paraboloide Abhängigkeit (s. Abb. 4); vgl. ferner S. 143 in diesem Buch.

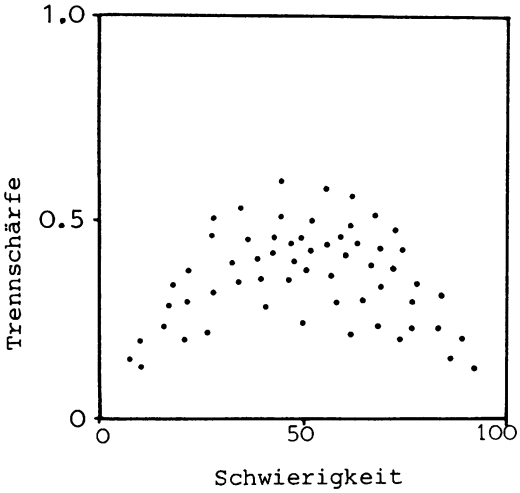


Abb. 4: Beziehung zwischen Schwierigkeit und Trennschärfe

nach LIENERT

Wie der Abbildung zu entnehmen ist, haben sehr schwierige und sehr leichte Aufgaben eine geringe Trennschärfe. Aufgaben mit mittlerer Schwierigkeit besitzen im allgemeinen die höchste Trennschärfe.

Die simultane Auswahl der Aufgaben nach Schwierigkeit und Trennschärfe „endet", wie LIENERT (1969) bemerkt, „nicht selten mit vielerlei Kompromissen!". Man mag sich diese Auswahlprozedur erleichtern, wenn man die Kennwerte der Aufgaben (Schwierigkeits- und Trennschärfeindex) in folgendes Diagramm einträgt (s. Abb. 5).

Bei der Auswahl der Aufgaben wird man zunächst so vorgehen, daß man vorab alle Aufgaben mit negativer Trennschärfe ausscheidet. Sollen die im Test verbleibenden Aufgaben einen Schwierigkeitsgrad zwischen 20 und 80 haben, dann errichtet man in diesen beiden Punkten das Lot und läßt alle Aufgaben wegfallen, die links bzw. rechts dieser beiden Geraden liegen. Will man weiterhin nur Aufgaben mit einer Trennschärfe von .30 und größer berücksichtigen, dann werden alle Aufgaben ausgeschieden, die unterhalb der Waagerechten liegen, die durch den Punkt .30 der Trennschärfeskala geht.

Allerdings wird man sich auch bei dieser Vorgehensweise jedes Schematismus' enthalten. Man sollte vielmehr darauf achten, daß in genügender Zahl leichte und schwere Aufgaben sowie eine angemessene Zahl mittelschwerer Aufgaben für den Test zurückbehalten werden, selbst wenn dies auf Kosten einer hinreichenden Trennschärfe geht. Die Aufgaben sollten sich hinsichtlich ihrer Schwierigkeit so verteilen, daß das Maximum bei den Aufgaben mit $P \approx 50$ liegt.

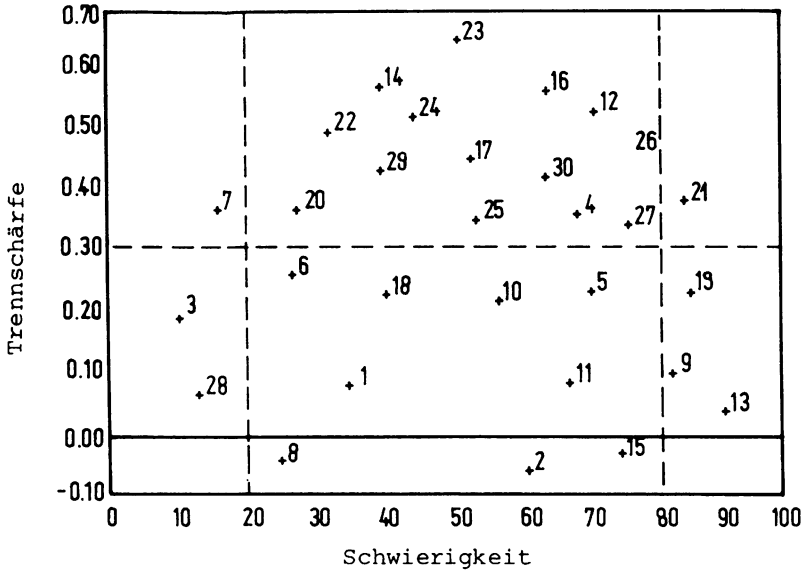


Abb. 5: Hilfsdiagramm zur Aufgabenauswahl (in Anlehnung an LIENERT 1969) Beispiel: Das Item mit der Nummer 3 hat einen Schwierigkeitsindex 10 und einen Trennschärfeindex .20)

Geht man zu starr und schematisch vor, dann kann der Fall eintreten, daß letztlich nicht genügend Aufgaben für den Test übrig bleiben. Dies soll natürlich kein Plädoyer dafür sein, „schlechte" Aufgaben in den Test aufzunehmen, sondern lediglich die Notwendigkeit einer flexiblen Haltung bei der Auswahlprozedur unterstreichen.

Trotzdem kann es vorkommen, daß die Zahl der in der Spezifikationstabelle eingetragenen Aufgaben größer ist als die Zahl der brauchbaren Items. D. h., einige der Verhaltens-Inhalts-Kombinationen, die man beabsichtigte zu prüfen, würden in der Testendform unzureichend oder gänzlich unberücksichtigt bleiben. In diesem Falle ist der Testautor gezwungen, die unbrauchbaren Aufgaben zu überarbeiten, um auf die angestrebte Zahl der Testaufgaben zu kommen. Dies ist besonders dann notwendig, wenn die betreffenden Aufgaben ein dem Lehrer wesentlich erscheinendes Lernziel prüfen sollen und er auf die Überprüfung nicht verzichten will. Er kann dabei so vorgehen, daß er die gesamte Aufgabe neu konstruiert (unter Berücksichtigung der gegebenen Konstruktionshinweise). Meist ist aber eine vollständige Neuformulierung der Aufgaben nicht unbedingt erforderlich. Be-

sonders bei Aufgaben mit Auswahlantworten genügt es häufig, die in der Distraktorenanalyse gegebenen Hinweise zu beachten und schlechte Distraktoren durch bessere zu ersetzen. Auf die Revision der Aufgaben muß natürlich erneut die Prozedur der Aufgabenanalyse folgen. Der Testautor darf sich also nicht damit zufrieden geben, z. B. einen neuen Distraktor einzuführen und dann diese Aufgabe ohne wiederholte Itemanalyse in den Test zu übernehmen. Er würde damit den Zweck und Sinn der Aufgabenanalyse überhaupt verfehlen. Wie wir oben angeführt haben, erfolgt die Auswahl der Items für die Testendform aufgrund der Aufgabenkennwerte Schwierigkeitsindex und Trennschärfeindex bzw. der Kombination beider. Dieses Vorgehen entspricht der sog. „rationalen Selektion" im Sinne LIENERTS.

4.2.6. Die Reliabilität des Tests

Unter der Reliabilität (Zuverlässigkeit) eines Tests versteht man den Grad der Genauigkeit, mit der ein Test das mißt, was er mißt. Von einem reliablen Test wird also verlangt, daß er bei wiederholten Messungen ähnliche Resultate liefert, identische Resultate sind nicht zu erwarten (s. dazu Kap. 3.1.3. in diesem Band).

Die Frage nach der Brauchbarkeit eines Tests wird durch die Überprüfung seiner Reliabilität nur zum Teil beantwortet. Ein Test kann hoch reliabel sein, aber dennoch unbrauchbar, wenn er nicht valide ist, d. h. wenn er nicht das Merkmal mißt, das er messen soll.

Die Reliabilität eines Tests wird numerisch mit einem Reliabilitätskoeffizienten beschrieben, dessen Bedeutung aber davon abhängt, wie er zustande gekommen ist, d. h. davon, welche Art der Reliabilitätsbestimmung man verwendet hat. Es soll in diesem Zusammenhang nur ein kurzer Überblick über die verschiedenen *Methoden der Reliabilitätsbestimmung* gegeben werden (s. Tab. 3). Ausführlicher wird dann auf die für die Konstruktion Informeller Tests bedeutsame Konsistenzanalyse eingegangen werden. Im übrigen verweisen wir auf den Beitrag von LANGFELDT (Kap. 3.1.).

Wie aus der Übersicht deutlich wird, gibt es also nicht *die* Reliabilität eines Tests, sondern eine Aussage über die Reliabilität eines Tests ist nur sinnvoll im Zusammenhang mit der Methode ihrer Bestimmung.

Wie angekündigt, wenden wir uns jetzt der *Berechnung des Reliabilitätskoeffizienten bei Informellen Tests,* der sog. *Konsistenzanalyse,* zu. Aus pragmatischen Gründen (geringer Zeitaufwand, einfache Berechnung, nur einmalige Testdurchführung) wird bei Informellen Tests in der Regel der Koeffizient der inneren Konsistenz berechnet. Innere Konsistenz eines Tests bedeutet, daß alle Aufgaben dasselbe messen, die Items des Tests sind also

Tab. 3: Übersicht über die Methoden zur Bestimmung der Reliabilität

Methode	Bezeichnung des Rel.-Koeffizienten
Wiederholte Testung (Retest) mit der gleichen Testform zu verschiedenen Zeitpunkten	Stabilitätskoeffizient
Retest mit einer Parallelform des Tests zum gleichen Zeitpunkt	Äquivalenzkoeffizient
Retest mit einer Parallelform des Tests zu verschiedenen Zeitpunkten	Stabilitäts- und Äquivalenzkoeffizient
Testhalbierung	Koeffizient der inneren Konsitenz (Halbierungskonsistenz)
Konsistenzanalyse	Koeffizient der inneren Konsistenz

homogen. Auf eben dieser Annahme basiert auch die zur Berechnung der inneren Konsistenz zu verwendende Methode von KUDER-RICHARDSON. Streng genommen ist eine Bestimmung des Konsistenzkoeffizienten nur bei solchen Tests oder solchen Testteilen sinnvoll, deren Aufgaben alle das gleiche Lernziel messen sollen. Erfaßt ein Test verschiedene Lernziele und wird die innere Konsistenz für den Gesamttest ermittelt, dann führt dies zwangsläufig zu einer unrealistischen Einschätzung der Reliabilität dieses Tests. Man sollte diesen Tatbestand bei der Reliabilitätsbestimmung im Auge behalten.

Bei der Konsistenzanalyse geht man davon aus, daß ein Test aus n Elementen (= n Aufgaben) besteht und bestimmt dann die Äquivalenz dieser Aufgaben. Anstelle der für dieses Verfahren von KUDER-RICHARDSON angegebenen Formel 20 verwendet man bei Informellen Tests eine vereinfachte Formel 21:

$$r = \frac{k}{k-1}\, 1 - \frac{M\,(k-M)}{k\,s^2}$$

Es bedeuten:

r = Reliabilitätskoeffizient
M = Mittelwert der Rohpunktwerte
k = Zahl der Testaufgaben
s = Standardabweichung der Rohpunktwerte

Wir wollen die Berechnung an einem (fiktiven) Beispiel verdeutlichen.

Tab. 4: Rohwertpunkte von 24 Schülern bei einem Test mit 35 Aufgaben, bei denen für jede richtige Lösung ein Punkt, für jede falsche Lösung null Punkte vergeben wurden.

Rohwerte (RW)	f	f . RW
10	1	10
12	1	12
14	2	28
17	3	51
19	2	38
22	4	88
23	4	92
25	2	50
28	2	56
29	1	29
30	1	30
32	1	32
	24	516

$$M = \frac{f \cdot RW}{N^*} = \frac{516}{24} = 21.5$$

* N = Zahl der Schüler

Die Standardabweichung wird nach einer vereinfachten Formel von JENKINS (zit. nach GAUDE & TESCHNER) berechnet.

Dazu werden die Rohpunkte des untersten Sechstels der getesteten Schüler von den Rohpunkten des obersten Sechstels subtrahiert und durch die halbe Zahl der getesteten Schüler dividiert:

$$s = \frac{1/6\, N_O \cdot RW - 1/6 N_U \cdot RW}{\frac{N}{2}}$$

Für das oberste Sechstel (24/6 = 4 Schüler) ergibt sich:
32 + 30 + 29 + 28 = 119 (= $1/6\, N_O \cdot RW$).

Für das unterste Sechstel finden wir:

$10 + 12 + 28 = 50 \; (= 1/6 \; N_U \cdot RW).$

$$s = \frac{119 - 50}{12} = 5.75$$

Der Reliabilitätskoeffizient r errechnet sich nun durch Einsetzen der gefundenen Werte in die oben angegebene Formel 21:

$$r = \frac{35}{35 - 1} \; 1 - \frac{21.5 \; (35 - 21.5)}{35 \cdot 5.75^2}$$

$= 1.03 \; (1 - .25)$

$= .77$

Der Koeffizient r schwankt zwischen 0 und 1.00. Je größer der Wert, um so reliabler ist der betreffende Test. Der in unserem (fiktiven) Beispiel errechnete Wert von .77 dürfte für einen Informellen Test hinreichend sein, da hier nicht die strengen Forderungen wie bei einem standardisierten Test gestellt werden können. Erhält man jedoch einen zu niedrigen Wert, etwa kleiner als .50, dann ist der Test nicht sehr zuverlässig und sollte vor einer gründlichen Überarbeitung nicht eingesetzt werden.

4.2.7. Die Validität des Tests

Ein Test ist dann valide (gültig), wenn er das mißt, was er messen soll. Bei der Überprüfung der Validität eines Tests gilt es also festzustellen, ob der Test dasjenige Merkmal einer Person, das ich messen will, auch tatsächlich mißt.

Es gibt verschiedene Arten der Validität, die sich, wie die Reliabilität, hinsichtlich der Methode ihrer Bestimmung und ihrer Aussage unterscheiden. Für Informelle Tests ist in der Regel nur das Konzept der inhaltlichen bzw. der curricularen Validität von Bedeutung.

Eine ausführliche Erörterung der Validitätsbestimmung und der damit zusammenhängenden Probleme findet sich im Kapitel 3.1.4. Aus diesem Grunde erübrigen sich weitere Ausführungen an dieser Stelle.

4.2.8. Die Normierung

„Allgemein versteht man unter ‚Norm' einen Vergleichswert, an dem man sich bei der Beurteilung einer Leistung orientiert" (Lienert 1969). Normen sind also Werte, die als Vergleichsmaßstab für die individuelle Leistung eines Schülers benutzt werden können.

Einige solcher u. U. für den Lehrer wichtigen Vergleiche seien genannt:

— der Vergleich einer individuellen Leistung mit der Leistung anderer Schüler,

— der Vergleich der Leistung eines Schülers in einem Test mit seiner Leistung in einem anderen Test.

Die Erstellung von Normen, wie sie etwa bei standardisierten Tests vorliegen, erfordert erhebliche Aufwendungen an Material und Personal. Der Lehrer wäre hoffnungslos überfordert, sollte er etwa den von ihm entwickelten Test an einer repräsentativen Stichprobe normieren. Andererseits mag es ihm wünschenswert erscheinen, die oben angeführten Vergleiche anzustellen. Zu Vergleichen innerhalb seiner Klasse könnte er noch die Rohwerte des Tests heranziehen, bei Vergleichen etwa mit Parallelklassen gerät er jedoch bereits in Schwierigkeiten. Mit Hilfe der Rohwerte lassen sich die Leistungen eines Schülers in verschiedenen Tests jedoch überhaupt nicht vergleichen.

Ein Verfahren, das dem Lehrer die angeführten Vergleiche ermöglicht, ihn aber andererseits zeitlich nicht übermäßig belastet, ist die Berechnung von *Prozentrangnormen*. Er braucht sich hierbei keine Gedanken über die Normalverteilung der Daten usw. zu machen, denn Prozentränge können auch bei schiefen Verteilungen errechnet werden.

Die Skala der Prozentränge reicht von 0 bis 100. Der Prozentrang kennzeichnet die relative Position einer individuellen Schülerleistung in bezug auf die Leistungen einer Gruppe von Schülern, der er angehört. Ein Prozentrang von 40 bedeutet, daß die Leistung dieses Schülers die Leistungen von 40 % der Schüler der Vergleichsgruppe übertrifft, während seine Leistung von den übrigen 60 % der Gruppe überboten wird. Hat ein Schüler in einem Englischtest einen Prozentrang von 50, in einem Deutschtest einen solchen von 70 erreicht, dann weiß der Lehrer, daß die *relative* Leistung dieses Schülers im Fach Deutsch besser ist als im Fach Englisch. Allein aufgrund der Rohwerte beider Tests hätte er nur schwerlich diese Information bekommen. Ein direkter Vergleich von Leistungen aus verschiedenen Tests ist erst mit Hilfe von Standardwerten (vgl. T-Werte) möglich.

Zur Berechnung der Prozentränge geht man wie folgt vor (s. a. Tab. 5):

(1) Die Rohwerte werden sinnvollerweise zu Klassen zusammengefaßt (Spalte 1).

(2) Es wird ausgezählt (Strichliste), wieviele Schüler auf die einzelnen Rohwertklassen entfallen (Spalte 2).

(3) Die kumulativen Häufigkeiten (f_{cum}) werden berechnet. Dazu werden die Häufigkeiten (f) von unten beginnend summiert, also $1 + 2 = 3 + 3 = 6$ usw. Der Wert in der obersten Reihe der Spalte 3 muß gleich der Gruppengröße ($N = 50$) sein.

(4) Die Prozentwerte werden für jede Reihe berechnet:

$$f_{cum}\% = \frac{f_{cum}}{N} \cdot 100$$

Beispiel für die Klasse (20—24): $f_{cum}\% = \frac{1}{50} \cdot 100 = 2$

Beispiel für die Klasse (45—49): $f_{cum}\% = \frac{18}{50} \cdot 100 = 36$

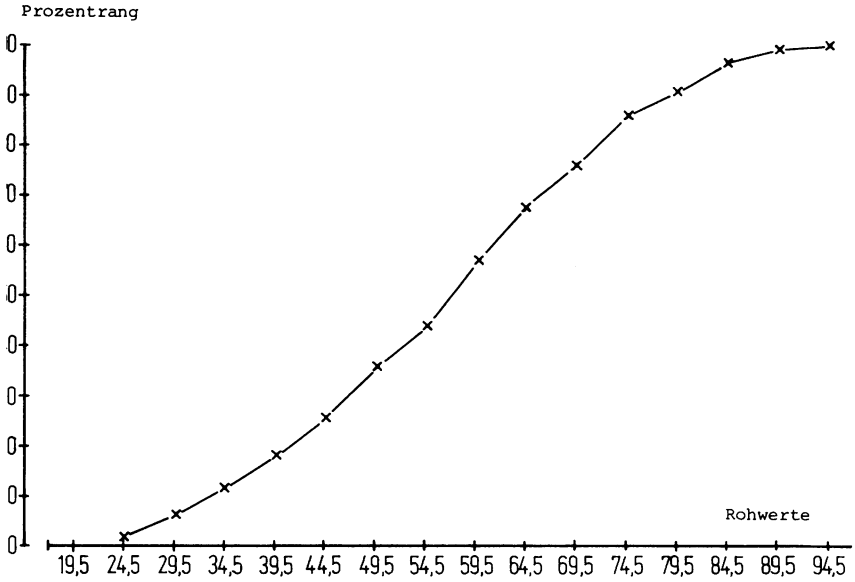
Tab. 5: Daten zur Berechnung der Prozentränge

1	2	3	N = 50 4
Klassen (Rohwerte)	Häufigkeiten f	Cumulative Häufigkeiten (f_{cum})	$f_{cum\%}$
90 - 94	/ 1	50	100
85 - 89	/ 1	49	98
80 - 84	/// 3	48	96
75 - 79	// 2	45	90
70 - 74	𝍸 5	43	86
65 - 69	//// 4	38	76
60 - 64	𝍸 5	34	68
55 - 59	𝍸 // 7	29	58
50 - 54	//// 4	22	44
45 - 49	𝍸 5	18	36
40 - 44	//// 4	13	26
35 - 39	/// 3	9	18
30 - 34	/// 3	6	12
25 - 29	// 2	3	6
20 - 24	/ 1	1	2

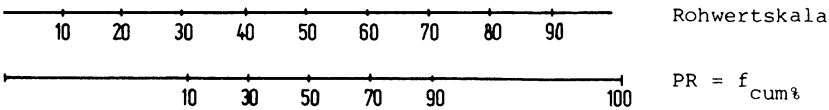
(5) Die Prozentwerte werden als Summenprozentkurve (Ogive) dargestellt. Auf der x-Achse werden die Rohwerte eingetragen, auf der y-Achse die Prozentwerte. Die Koordinatenschnittpunkte aus Rohwert und Prozentwert werden durch eine geglättete Kurve verbunden.
Es ist zu beachten: Die in Spalte 4 eingetragenen Prozentwerte entsprechen jeweils der oberen Klassengrenze; z. B. entspricht dem Rohwert 29 der Prozentwert 6 (streng genommen entspricht dieser Prozentwert dem Rohwert 29.5, da die Klasse von 24.5 bis 29.5 reicht).

(6) Die Summenprozentkurve liefert uns nun zweierlei Information:
a) zu jedem beliebigen Rohwert (Punktzahl) kann der entsprechende Prozentrang abgelesen werden,
b) umgekehrt kann zu jedem Prozentrang die zugehörige Maßzahl aufgesucht werden.

Abb. 6: Summenprozentkurve



Die beschriebenen Prozentränge dürfen nicht verwechselt werden mit dem Prozentsatz richtiger Lösungen in einem Test. So kann ein Schüler in einem leichten Test etwa 75 % aller Aufgaben richtig gelöst haben und trotzdem nur den Prozentrang 25 haben, weil eben sehr viele Schüler eine große Zahl von Aufgaben richtig beantworteten. Dagegen kann er bei einem schweren Test, bei dem er 50 % der Aufgaben löste, einen Prozentrang von 90 erreichen. Bei der Interpretation ist eine besondere Eigenschaft der Prozentränge zu beachten. Die Prozentrangnormen lassen nämlich individuelle Testunterschiede im Mittelbereich der Skala stärker erscheinen, als sie es tatsächlich sind, während die Unterschiede in extremen Bereichen stark nivelliert werden (s. u.).



Wie man sieht, ist der Unterschied bezüglich der Rohwerte zwischen den Prozenträngen 30 und 50 geringer als der zwischen den Prozenträngen 90 und 100, obwohl im ersteren Fall der numerische Prozentrangunterschied doppelt so groß ist. Diese Eigenschaft der Prozentskala muß unbedingt berücksichtigt werden, will man nicht zu falschen Schlußfolgerungen kommen (etwa bei dem Vergleich zweier Schüler). Des weiteren ist zu beachten, daß

mit Prozenträngen nicht algebraisch operiert werden darf, man kann also etwa keine Mittelwerte aus mehreren Prozenträngen bestimmen.

Die beschriebenen Nachteile lassen sich leicht vermeiden, wenn man die Prozentränge mit Hilfe eines von McCall entwickelten Verfahrens in Standardwerte — die sog. T-Werte — umrechnet (s. dazu Lienert 1969). Für die Umwandlung von Prozenträngen in T-Normen steht die Tabelle 6 zur Verfügung. Die T-Werte können wie andere Standardwerte (z. B. z, C, IQ) interpretiert und algebraisch verwendet werden.

Tab. 6: Tranformation von Prozenträngen (PR) in T-Werte (nach LIENERT 1969, S. 490)

PR	T	PR	T
0	20	54	51
0	21	58	52
0	22	62	53
0	23	66	54
0	24	69	55
1	25	73	56
1	26	76	57
1	27	79	58
1	28	82	59
2	29	84	60
2	30	84	60
3	31	86	61
3	32	88	62
4	33	90	63
5	34	92	64
7	35	93	65
8	36	95	66
10	37	96	67
12	38	96	68
13	39	97	69
16	40	98	70
18	41	98	71
21	42	99	72
24	43	99	73
27	44	99	74
31	45	100	75
34	46	100	76
38	47	100	77
42	48	100	78
46	49	100	79
50	50	100	80

Abschließend sei noch eine Methode angeführt, die es erlaubt, auch ohne Berechnung von Prozenträngen zu T-Standardwerten zu gelangen. Dieses Verfahren, das besonders bei kleinen Gruppen zu empfehlen ist, erlaubt eine direkte Umwandlung von Rangplätzen in T-Werte.

Bei der *Rangreihen-Umwandlung in T-Standardwerte* nach TREML, zit. bei HELLER (1973, S. 246), geht man wie folgt vor:

Nach der Durchführung eines Tests (beispielsweise in einer Klasse mit 20 Schülern) werden für jeden Schüler die Rohpunktwerte ermittelt. Sodann wird jedem Schüler aufgrund seines Rohpunktwertes eine Rangzahl zugeordnet. Der Schüler mit dem höchsten Rohpunktwert erhält den Rang 1, der mit dem zweithöchsten den Rang 2 usw. Dem Schüler mit dem niedrigsten Rohpunktwert wird der Rang 20 zugeordnet. Zur Ermittlung des T-Wertes bedient man sich der Tabelle 7. In der äußeren linken Spalte findet sich die Gruppengröße, in der untersten Zeile der (absolute) Rangplatz. Der jeweils zugehörige T-Standardwert steht im Schnittpunkt der Koordinaten Gruppengröße und Rangplatz. Bei einer Klasse mit 20 Schülern erhält der Schüler mit dem Rangplatz 1 somit den T-Wert 68, der mit Rangplatz 2 den T-Wert 64, der mit Rangplatz 12 den T-Wert 48 usw. (s. S. 221).

4.2.9. Literaturverzeichnis

Bloom, B. S.; Hastings, J. Th. u. *Madaus, G. F.:* Handbook on formative and summative evaluation of student learning. New York, McGraw-Hill, 1971.

Bloom, B. S.: Taxonomy of educational objectives. Handbook 1, Cognitive domain. New York, Mc Kay, 1956.

Ebel, R. L.: Measuring educational achievement. Englewood Cliffs, Prentice-Hall, 1965.

Gronlund, N. E.: Measurement and evaluation in teaching. New York, McMillan, 1968.

Gaude, P. u. *Teschner, W. P.:* Objektivierte Leistungsmessung in der Schule. Frankfurt am Main, Diesterweg, 1971.

Heller, K.: Intelligenzmessung. Villingen, Neckar-Verlag, 1973.

Ingenkamp, K.: Die deutschen Schultests. Weinheim, Beltz, 1962.

Lienert, G,: Testaufbau und Testanalyse. Weinheim, Beltz, 1969.

Nunnally, J. C.: Educational measurement and evaluation. New York, McGraw-Hill, 1964.

Nunnally, J. C. u. *Ator Nancy Almand:* Educational measurement and evaluation. New York, McGraw-Hill, 1972.

Smith, E. R. u. *Tyler, R. W.:* Appraising and recording student progress. New York, Harper & Row, 1942.

7: RANGREIHEN-UMWANDLUNG IN "T"—STANDARDWERTE nach F. TREML

67 60 57 54 51 49 46 43 39 33
67 61 57 54 52 50 48 45 42 38 32
67 61 58 55 53 51 49 48 45 42 38 32
67 62 59 56 54 52 50 48 46 44 41 37 32
67 62 59 56 54 52 51 49 47 45 43 40 37 32
67 62 59 57 55 53 51 50 48 46 45 43 40 37 32

68 63 60 57 55 54 52 51 49 47 46 44 42 40 37 32
68 63 60 58 56 54 53 51 50 48 47 45 44 42 39 37 32
68 63 60 58 56 55 53 52 50 49 48 46 45 43 41 39 36 31
68 64 61 59 57 55 54 52 51 50 48 47 46 44 43 41 39 36 31
68 64 61 59 57 56 54 53 52 50 49 48 47 45 44 42 40 38 36 31

69 64 61 59 58 56 55 53 52 51 50 49 47 46 45 43 42 40 38 35 31
69 64 62 60 58 56 55 54 53 51 50 49 48 47 46 44 43 42 40 38 35 31
69 64 62 60 58 57 55 54 53 52 51 50 49 48 46 45 44 43 41 40 38 35 31
69 65 62 60 58 57 56 55 53 52 51 50 49 48 47 46 45 44 43 41 39 37 35 30
69 65 62 60 59 57 56 55 54 53 52 51 50 49 48 47 46 45 43 42 41 39 37 35 30

69 65 62 61 59 58 56 55 54 53 52 51 50 49 48 47 46 45 44 43 42 40 39 37 34 30
69 65 63 61 59 58 57 56 55 54 53 52 51 50 49 48 47 46 45 44 43 42 40 39 37 34 30
70 65 63 61 60 58 57 56 55 54 53 52 51 50 49 48 48 47 46 45 44 43 42 40 38 37 34 30
70 66 63 61 60 58 57 56 55 54 53 52 51 51 50 49 48 47 46 45 44 43 42 41 40 38 36 34 30
70 66 63 62 60 59 57 56 55 54 54 53 52 51 50 49 48 47 47 46 45 44 43 42 41 40 38 36 34 30

70 66 63 62 60 59 58 57 56 55 54 53 52 51 51 50 49 48 47 47 46 45 44 43 42 41 39 38 36 34 30
70 66 64 62 60 59 58 57 56 55 54 53 52 52 51 50 49 49 48 47 46 45 44 44 43 42 41 39 38 36 34 30
70 66 64 62 61 59 58 57 56 55 54 54 53 52 51 51 50 49 48 47 47 46 45 44 43 42 41 40 39 38 36 33 29
70 66 64 62 61 59 58 57 57 56 55 54 53 52 52 51 50 49 49 48 47 46 46 45 44 43 42 41 40 39 37 36 33 29
70 66 64 63 61 60 59 58 57 56 55 54 53 53 52 51 50 50 49 48 48 47 46 45 45 44 43 42 41 40 39 37 35 33 29

70 66 64 63 61 60 59 58 57 56 55 54 54 53 52 51 51 50 49 49 48 47 47 46 45 44 44 43 42 41 40 39 37 35 33 29
70 67 64 63 61 60 59 58 57 56 55 55 54 53 52 52 51 50 50 49 48 48 47 46 46 45 44 43 42 41 40 39 38 37 35 33 29
70 67 64 63 61 60 59 58 57 56 56 55 54 53 53 52 51 51 50 49 49 48 47 47 46 45 45 44 43 42 41 40 39 38 37 35 33 29
71 67 65 63 61 60 59 58 57 57 56 55 54 54 53 52 52 51 50 50 49 48 48 47 47 46 45 44 44 43 42 41 40 39 38 37 35 33 29
71 67 65 63 62 60 59 58 58 57 56 55 55 54 53 53 52 51 51 50 49 49 48 48 47 46 46 45 44 44 43 42 41 40 39 38 36 35 33 29

71 67 65 63 62 61 60 59 58 57 56 55 55 54 53 53 52 52 51 50 50 49 49 48 47 47 46 45 45 44 43 43 42 41 40 39 38 36 35 33 29
71 67 65 63 62 61 60 59 58 57 57 56 55 54 54 53 52 52 51 51 50 49 49 48 48 47 47 46 45 45 44 43 42 42 41 40 39 38 36 35 32 29
71 67 65 63 62 61 60 59 58 57 57 56 55 55 54 53 53 52 51 51 50 50 49 49 48 47 47 46 46 45 44 44 43 42 41 41 40 39 37 36 34 32 29
71 67 65 63 62 61 60 59 58 57 57 56 55 55 54 53 53 52 52 51 51 50 49 49 48 48 47 47 46 45 45 44 43 43 42 41 40 39 38 37 36 34 32 29

Zahlen in der linken Kolonne bezeichnen die Gruppengröße (Gg)
Ordnungszahlen in der untersten Zeile sind Rangplätze (Rp)
(So erhält ein Schüler mit dem 2. Rangplatz in einer Klassengruppe von N = 36 einen linearen Skore von T = 66)

4.3. Einsatz standardisierter Schulleistungstests

Ralf Horn

4.3.0. Vorbemerkung

Da es seit einiger Zeit keinen systematischen Überblick mehr über die auf dem Markt vorhandenen standardisierten Schulleistungstests gibt (INGENKAMP 1962), bereitet es einige Schwierigkeiten, sich vollständige Informationen über diese Tests zu beschaffen. Es kommt hinzu, daß gerade in den letzten Jahren eine Fülle von Verfahren publiziert worden ist. Es ist zu erwarten, daß auch weiterhin eine größere Anzahl von Verfahren veröffentlicht werden wird; daher ist es kaum möglich, einen vollständigen Überblick über die Verfahren zu geben.

Für eine erste Auseinandersetzung mit den inhaltlichen Gegebenheiten und den methodischen Ansätzen ist es dagegen sinnvoll, sich mit den verschiedenen Verfahren zu beschäftigen, die sich innerhalb einer Kategorie, etwa dem Lesen, ohnehin nur geringfügig unterscheiden. Daher sollen in der Folge die methodischen Ansätze für die Beurteilung der Leistung in den verschiedenen Fachbereichen ausführlicher besprochen werden.

Eine Einteilung der Testverfahren nach Fachbereichen kann nach inhaltlichen Kriterien leicht vorgenommen werden. Dabei gibt es innerhalb der verschiedenen Kategorien unterschiedlich viele Testverfahren. Da wir uns hier lediglich mit den methodischen Ansätzen befassen, können diese Einschränkungen in der Verwendungsmöglichkeit weggelassen werden. Eine Vorstellung davon, wie viele Verfahren es in den einzelnen Bereichen gibt, vermittelt die Tabelle am Ende dieses Aufsatzes (vgl. Abschn. 4.3.7.).

Es muß darauf hingewiesen werden, daß auch für die verschiedenen Schultypen unterschiedlich viele Schulleistungstests zur Verfügung stehen. Am meisten Verfahren gibt es zur Zeit für den Bereich der Grundschule, während für Realschule und Gymnasium wesentlich weniger Verfahren zur Verfügung stehen.

Diesem Unterschied in der Zahl der verfügbaren Verfahren entsprechen deutliche Unterschiede in der Einstellung der Lehrer zu Testverfahren. Diese Einstellung ist bei Grundschullehrern am positivsten und bei Gymnasiallehrern am negativsten (HAASE 1972).

Da es für den Bereich der Grundschule am meisten Verfahren gibt, ist es nicht überraschend, daß viele Verfahren die Grundfertigkeiten Lesen, Schreiben und Rechnen überprüfen. Mit diesen Verfahren beschäftigen wir uns zunächst.

4.3.1. Lesetests

Bei allen Verfahren, die dieser Kategorie zugeordnet werden können, wird überprüft, ob die Schüler in der Lage sind, Informationen aus einzelnen Sätzen oder Abschnitten zu entnehmen. Dabei variiert die Komplexität der Abschnitte von einem oder zwei Sätzen (2. Schuljahr) bis zu kleineren Geschichten (7. bis 9. Schuljahr, Hauptschule).

Das Verständnis des Gelesenen wird dadurch überprüft, daß zu dem vorgegebenen Text Fragen gestellt werden, die von den Schülern beantwortet werden müssen. Diese Fragen gehören ohne Ausnahme zur Kategorie ‚Verständnis' der Taxonomie von BLOOM, da sie das Umsetzen von einer Art der Aufgaben an einem Beispiel zeigen. Es handelt sich dabei um eine Beispielaufgabe aus dem Test Verständiges Lesen, VL 5—6, (1965):

„Professor Einstein, der berühmte Gelehrte, erzählt von einer Bahnfahrt durch Amerika folgendes Erlebnis: Er hatte allein in seinem Abteil gesessen und gelesen, dann seine Brille abgesetzt und über das Gelesene nachgedacht. Tief in Gedanken versunken ging er dann in den Speisewagen. Als er zur Speisekarte griff, bemerkte er, daß er seine Brille vergessen hatte.
Deshalb bat er einen Neger, der am gleichen Tisch saß, ihm die Karte vorzulesen. Der aber schüttelte verlegen lächelnd den Kopf und flüsterte: „Leider habe ich auch nicht lesen gelernt."

Bei den ersten beiden Aufgaben, die nun folgen, sind die richtigen Lösungen bereits angekreuzt. Sieh Dir diese Aufgabe genau an. Suche selbst die Lösungen der nächsten Aufgabe und mache jedesmal ein Kreuz durch den danebenstehenden Buchstaben.

1. Aus der Geschichte ergibt sich, daß Einstein
   A ... in der Schule nicht lesen gelernt hatte.
   B ... sich mit dem Neger einen Scherz erlaubte.
   C ... sehen wollte, ob der Neger lesen konnte.
   X ... ohne Brille nicht lesen konnte.
2. Aus der Geschichte erfahren wir,
   A ... daß Einstein gute Augen hatte.
   X ... daß nicht alle Menschen lesen können.
   C ... daß die Speisekarte schlecht gedruckt war.
   D ... daß Einstein den Neger gut leiden mochte.
3. Der Neger nahm an, daß Einstein
   A ... gar nicht lesen konnte.
   B ... seine Brille vergessen hatte.
   C ... blind war.
   D ... ihn zum Mittagessen einladen wollte.
4. Die beste Überschrift für diese kleine Geschichte wäre:
   A ... Der ertappte Schulschwänzer.
   B ... Der berühmte Professor.
   C ... Eine schwere Beleidigung.
   D ... Ein lustiger Irrtum."

Es ist noch nachzutragen, daß die Korrelation zwischen der schulischen Beurteilung der Leseleistung und den Testergebnissen der Lesetests relativ eng ist (r = 0,74 zwischen dem Ergebnis des LT 2 und der Deutschzensur), daß aber trotzdem nur etwa die Hälfte der Varianz aufgeklärt wird.

4.3.2. Rechtschreibtests

Die Verfahren zur Erfassung der Rechtschreibleistung spielen in der Grundschule eine große Rolle, weil dort die Diagnose und die Therapie der Legasthenie in den letzten Jahren besonders intensiv betrieben wird.

Dabei werden Intelligenz- und Rechtschreibtests häufig verwendet, während Lesestörungen seltener mit in die Diagnose einbezogen werden.

Bei den Rechtschreibtests lassen sich zwei Grundtypen unterscheiden: einmal der Test, der auf eine reine Leistungsfeststellung abzielt und zweitens die Verfahren, die eine qualitative Auswertung ermöglichen und damit Ansätze für die Therapie der Minderleistungen liefern.

Der methodische Ansatz ist jedoch bei beiden Grundtypen zunächst gleich. Es ist offenkundig, daß es eine ganze Reihe von Wörtern gibt, die den Schülern aufgrund des häufigen Vorkommens keine Schwierigkeiten machen. Bei einem Diktat, das vom Lehrer vorgelesen wird, sind diese Wörter, da sie für den Zusammenhang eines Textes notwendig sind, ebenfalls relativ häufig vorhanden, aber für den eigentlichen diagnostischen Zweck wenig interessant. Bei allen Rechtschreibtests brauchen die Schüler diese einfachen Wörter wie „der, die, das, und" usw. nicht zu schreiben, sie sind dort bereits vorgedruckt. Die Schüler schreiben also nur die eigentlichen „kritischen" Wörter, die auch diagnostisch interessant sind. Der Vorteil gegenüber dem Lehrerdiktat liegt auf der Hand. In kürzerer Zeit kann eine größere Anzahl von „kritischen" Wörtern von den Schülern geschrieben werden und die Wörter, die ohnehin von fast allen Schülern richtig geschrieben werden, belasten die Schüler nicht.

Auch für diesen Testtyp ein Beispiel. Es wurde dem Diagnostischen Rechtschreibtest, DRT 4—5 (1969) entnommen.

„In der ..... müssen wir vieles lernen.
Wir ..... Deutsch, Rechnen, Turnen und ..... Fächer. Am liebsten gehen wir jedoch alle ....."

Beim Vorlesen des vollständigen Textes durch den Lehrer werden von den Schülern der Reihe nach folgende Wörter eingesetzt: Schule, haben, andere, wandern. Dabei handelt es sich nicht um kritische Wörter, sondern bei diesem Beispiel lernen die Schüler das Vorgehen beim Test kennen.

Neben dem Diagnostischen Rechtschreibtest, DRT 4—5, gibt es noch zwei weitere Verfahren, die als diagnostische Verfahren in der zweiten und

dritten Klasse eingesetzt werden können (DRT 2 und DRT 3). Zu diesen beiden Verfahren gibt es ein entsprechendes Trainingsprogramm zur Behebung der festgestellten Fehler. Die Ansichten über dieses Programm gehen auseinander, weil eine empirische Bestätigung für die Möglichkeit der Übertragung der von MÜLLER beim DRT 2 und DRT 3 aufgefundenen Fehlerklassen aussteht. MEIS mußte bei der Konstruktion des DRT 4—5 feststellen, daß die von MÜLLER vorgeschlagene Typologie nicht übernommen werden konnte (Beiheft zum DRT 4—5, S. 17). Weiter konnte bestätigt werden, daß der sinnvollste Weg zur Beseitigung von Rechtschreibfehlern darin besteht, ein intensives Training durchzuführen. Die Möglichkeit, mit Hilfe von einsichtigen Regeln die Rechtschreibung zu verbessern, wird dagegen skeptisch beurteilt.

4.3.3. Rechentests

Testverfahren für das Fach Rechnen gibt es relativ wenig. Das liegt zum großen Teil daran, daß die Notwendigkeit für die Konstruktion derartiger Verfahren weniger offenkundig ist. Rechenaufgaben lassen sich im Gegensatz zu anderen Aufgaben aus den Unterrichtszielen leicht ableiten und sind objektiv auswertbar.

Auch bei Rechentests gibt es zwei Grundtypen, von denen einer der reinen Leistungsfeststellung dient (etwa der Rechentest für 2. Klassen, RT 2) und der andere Ansätze zu einer Therapie liefert (Diagnostischer Rechentest für 3. Klassen, DRE 3).

Gegenüber den Aufgaben, die vom Lehrer einer Klasse gestellt werden, unterscheiden sich die im Test verwendeten Aufgaben lediglich dadurch, daß sie nach Schwierigkeit und Trennschärfe ausgelesen und wiederholt überprüft wurden. Auf eine Darstellung von Beispielaufgaben kann daher verzichtet werden.

4.3.4. Allgemeine Schulleistungstests

Neben den Verfahren, die auf *einen* Bereich beschränkt sind, gibt es „Omnibusverfahren", die die Schulleistungen in mehreren Fachgebieten erfassen. Derartige Verfahren können überall dort eingesetzt werden, wo ein möglichst umfassender Überblick über die gesamte Leistungsfähigkeit des Schülers benötigt wird.

Eine konkretere Vorstellung von derartigen Verfahren erhält man, wenn man ein Verfahren, das zu dieser Gruppe gehört, etwas ausführlicher analysiert. Als Test dieses Typs bietet sich der Allgemeine Schulleistungstest für 2. Klassen, AST 2, an. Dieser Test besteht aus 6 Untertests und kann in zwei Unterrichtsstunden durchgeführt werden. Die einzelnen Untertests sind:

Wortschatz
Leseverständnis (aufgebaut wie die bereits besprochenen Lesetests)
Zahlenrechnen
Textaufgaben
Rechtschreiben (abweichend von den oben beschriebenen Rechtschreibtests wird von den Schülern nicht verlangt, daß sie „kritische" Wörter schreiben, sondern sie sollen die richtige Schreibweise eines Wortes unter vier vorgegebenen und drei falschen Antworten erkennen.
Die Korrelation dieses Untertests mit den entsprechenden Noten liegt bei .68. Die Korrelation zum Diagnostischen Rechtschreibtest, DRT 2, liegt in der gleichen Größenordnung (Beiheft AST 2, S. 13).
Sachwissen

Die übrigen Verfahren, die noch dieser Kategorie zuzuordnen sind, sind ähnlich aufgebaut. Eine Ausnahme stellt lediglich die Schulleistungstestbatterie für Lernbehinderte SBL I und SBL II (1972) dar. Da diese Verfahren am Ende der ersten bzw. am Ende der zweiten Klasse eingesetzt werden und zwischen den leistungsschwachen Schülern differenzieren sollen, werden dabei nur Lesen, Rechtschreiben und Rechnen überprüft. Die Abweichungen, die gegenüber den bereits geschilderten Methoden zur Erfassung dieser Leistungen auftreten, sind nicht so tiefgreifend, daß sie gesondert behandelt werden müßten.

4.3.5. Fremdsprachentests

Das Erfassen fremdsprachlicher Leistungen mit Hilfe von standardisierten Verfahren unterliegt einer Reihe von Beschränkungen, die nur schwer überwunden werden können. Dazu gehört vor allem, daß die eigentliche sprachliche Produktion mit den üblichen Papier- und Bleistiftverfahren kaum zu erfassen ist.

Die Autoren von Fremdsprachentests behelfen sich damit, daß sie Situationen mit Hilfe von bildlichen Darstellungen anbieten und dazu Fragen stellen. Daneben ist es natürlich mit geringerer Schwierigkeit möglich, das Textverständnis zu überprüfen.

Das kann mit ähnlichen Methoden geschehen, wie sie bei den Lesetests besprochen wurden.

Wie die Aufgaben eines Tests etwa aussehen und welche Aspekte erfaßt werden, läßt sich am Beispiel eines Tests am besten zeigen. Es handelt sich dabei um den Englisch-Einstufungstest 6+ (1973). Dieser Test besteht aus zwei Teilen, die einen umfassenden Überblick über den Leistungsstand der Schüler nach zwei Jahren Unterricht ermöglichen. Dabei sind die Aufgaben so ausgewählt, daß die Testergebnisse nicht von dem verwendeten Lehrbuch abhängen.

Teil I erfaßt „Spelling" und „Vocabulary". Dabei sind die Aufgaben folgendermaßen aufgebaut:

1) Spelling

Susan is Mrs. Brown's d .....
A augther
B aughter
C oughter
D ougther

2) Vocabulary

Hamburg is in the ..... of Germany.
A south
B north
C west
D east

Teil II des Tests erfaßt „Structure", „Pronunciation" und „Listening Comprehension". Da es sich bei diesen Untertests um neue Ansätze handelt, derartige Leistungen zu erfassen, sei auch hierzu jeweils eine Beispielaufgabe angeführt.

3) Structure

Mr. Smith was born in Newcastle. So he comes ..... Newcastle.
A out of
B out
C from
D of

4) Pronunciation

Bei diesem Untertest soll der Schüler den Laut bei den vorgegebenen Antworten anstreichen, der dem im zugehörigen Satz gleich ist.

Daddy, *gi*ve me the milk please.
A he *li*ves
B a*li*ve
C *li*fe
D wi*f*e

5) Listening Comprehension

Der Schüler hat dabei nur die vier Antworten vor sich. Der zugehörige Text wird vom Lehrer vorgelesen.

I must get some fruit. Yes, go and get some .....
A beans
B cabbages
C peas
D pears

Die anderen zur Zeit erhältlichen Testverfahren erfassen die fremdsprachlichen Leistungen auf ähnliche Weise. Allerdings meistens nicht so präzise auf die verschiedenen sprachlichen Aspekte bezogen wie im obigen Beispiel.

4.3.6. Tests für verschiedene Fächer

Neben den bereits besprochenen Verfahren für die verschiedenen Fachbereiche gibt es noch einige andere Verfahren, die spezielle Fächer behandeln. Da es von den meisten dieser Tests nur ein Verfahren pro Fach gibt, werden sie hier summarisch besprochen.

Das Konstruktionsverfahren von Schulleistungstests läßt sich ohne größere Schwierigkeiten auf alle Fächer übertragen. Einheitlich werden dabei Mehrfachauswahlaufgaben verwendet. Für ein bestimmtes Fachgebiet müssen zuerst die entsprechenden Lernziele aufgestellt werden und nach der Aufgabenkonstruktion kann dann die Erprobung und Eichung beginnen. Es gibt zur Zeit Testverfahren für die Fächer Erdkunde, Geschichte und Naturlehre. Diese Verfahren weisen alle den gleichen Antworttypus auf und unterscheiden sich auch im Prinzip wenig in der Aufgabenstellung voneinander. Eine konkretere Vorstellung vermittelt eine Beispielaufgabe, die aus dem Erdkundetest Deutschland, ETD 5—7 (1971), entnommen wurde.

Ein Kompaß dient zur
A Zeitmessung
B Messung von Luftdruck
C Bestimmung der Meerestiefe
D Messung der Luftfeuchtigkeit
E Feststellung von Himmelsrichtungen

Vollständige Information über das Angebot lieferbarer Verfahren für den Einsatz in der Schule enthalten die Verzeichnisse der Verlage, die Tests veröffentlichen. Die Anschriften lauten:

Beltz Test GmbH, 694 Weinheim, Postfach 167
Marhold, C., 1 Berlin 19, Hessenallee 12
Reinhardt, E., 8 München 38, Schalterfach
Verlag f. Psychologie, 34 Göttingen, Postfach 414
Westermann, G., 33 Braunschweig, Postfach 3320
Wolf, L., 84 Regensburg, Postfach 112

4.3.7. Übersicht über die zur Zeit verfügbaren Tests nach Testkategorien und Klassenstufen

Klasse	1	2	3	4	5	6	7	8	9
1. Lesetests		Lesetest LT 2	Sinnverstehendes Lesen SVL 3 Lesen 3	Verständiges Lesen VL 5-6 Lesen 4			Verständiges Lesen VL 7 - 9		
2. Rechtschreibtests	Rechtschreibtest RST 1	Diagn. Rechtschreibtest DRT 2	Diagn. Rechtschreibtest DRT 3	Diagn. Rechtschreibtest DRT 4-5 Rechtschreibtest RST 4 +				Rechtschreibtest RST 8 Rechtschreibungstest RT	
3. Rechentests		Rechentest RT 2	Diagn. Rechentest DRE 3		Bruchrechentest BRT 6			Rechentest RT 8 +	
4. Allgemeine Schulleistungstets	Schulleistungstestbatterie für Lernbeh. SBL I	Schulleistungstestbatterie f. Lernbeh. SBL II Allgem. Schulleistungstest AST 2	Allgemeiner Schulleistungstest AST 3 Kombinierter Schultest KS 3	Allgem. Schulleistungstest AST 4 Kombiniert. Schultest KS 4	Kombin. Schult. KS 5			Schulabschluß u. Berufseintrittstest SABET 8 +	
5. Fremdsprachentests				Hamburger Englisch-test HET 6 + Diagn. Englisch Leistungstest ELT 6-7 Englisch-Einstufungstest 6 +				Französischer Wortschatztest FWS 9-12	
6. Tests für verschiedene Fächer					Erdkundetest Deutschland ETD 5-7			Geschichtstest Neuzeit GTN 8-10 Naturlehretest NLT 9	

5. Subjektive Verfahren der Leistungsbeurteilung in der Schule

Einleitender Kommentar

Daß die Leistungsbeurteilung voller Probleme steckt, wurde bereits einleitend betont. In ganz besonderem Maße gilt diese Feststellung im Hinblick auf die subjektiven Verfahrensansätze, also jene Methoden, die dem Lehrer unmittelbar zur Verfügung stehen. Das Repertoire für schulische Zwecke der Leistungsbeurteilung verfügbarer Methoden umfaßt zwar alle in diesem Buch behandelten Verfahrensmodi, gleichwohl gewinnen die sog. subjektiven Verfahrensweisen, wie (freie oder gebundene) Verhaltensbeobachtung, Beurteilungstechniken (z. B. Rating-Skalen), schriftliche oder mündliche Prüfungen bzw. die verschiedenen Formen der Notengebung (Zensurierung), im Rahmen der praktischen Schülerbeurteilung nach wie vor herausragende Bedeutung. Die Erziehungs- bzw. Sozialwissenschaftler mögen die Ursache hierfür im Traditionalismus unseres Bildungssystems suchen (und finden), die Schulpraktiker werden dagegen die ‚Schuld' den Wissenschaftlern zuschieben mögen, die offenbar nicht in der Lage sind, genügend brauchbare und hinreichend abgesicherte Verfahren für die Praxis der Schülerbeurteilung bereitzustellen. Der Verfasser, der sich beiden ‚Lagern' verpflichtet weiß, gesteht jedem der beiden Standpunkte eine gewisse Berechtigung zu und ist nüchtern genug, weder wissenschaftliche Allheilmittel noch umwälzende Neuerungen der Beurteilungspraxis in absehbarer Zeit zu erwarten. Trotz erkennbarer Fortschritte gerade auf testdiagnostischem Gebiet wird den subjektiven Methoden der Schülerbeurteilung auch in der Zukunft große Bedeutung zukommen. Deren Einsatzmöglichkeiten so schnell und gut wie möglich zu verbessern, hilft m. E. den Betroffenen mehr als vorläufig nicht realisierbaren Wunschvorstellungen nachzuhängen. Außerdem erblicken wir in den beiden Verfahrensansätzen der ‚objektiven' und ‚subjektiven' Beurteilung prinzipiell sich ergänzende, keine inkompatiblen Methoden der Schülerbeurteilung.

Für Unterrichtsanalysen mannigfacher Art (z. B. Interaktionsanalysen, Analysen des Lehrerverhaltens, didaktische Analysen usw.) oder für Zwecke der Beobachtung und Beurteilung des Schülerverhaltens (Mitarbeit, Konzentration, Arbeitshaltung, Lernmotivation, emotionales und soziales Verhalten u. ä.) bieten sich die sog. *Beobachtungs-* und *Beurteilungstechniken* (Observational Techniques bzw. Ratingverfahren, Contentanalysen, Q-Techniken) als adäquate Methoden an (vgl. HELLER et al. 1974). Der Beitrag von LANGHORST zeigt eine Reihe von Möglichkeiten auf, diese Verfahren im Rahmen des Unterrichts sinnvoll einzusetzen, wobei der phänomengetreuen Verhaltens*beschreibung* und Verhaltens*beurteilung* mit

Hilfe von Kategorienlisten oder Schülerbeobachtungsbögen sowie sog. rating scales besondere Aufmerksamkeit geschenkt wird. Ein eigener Abschnitt befaßt sich mit den häufigsten *Beurteilungsfehlern*, deren Kenntnis für eine vorurteilsfreie und treffsichere Schülerbeurteilung gleichermaßen von Bedeutung ist.

Im folgenden Beitrag von FINGERHUT und LANGFELDT werden Probleme der *Notengebung in der Schule* erörtert, wobei sich die Autoren zum Teil auf eigene empirische Erhebungen stützen können. Zunächst werden „die instrumentellen Eigenschaften der Noten als Indikatoren für bestimmte Leistungen von Schülern" untersucht, d. h. die Urteils- und Meßwertfunktion von Schulzensuren analysiert. Entsprechende Varianzanalysen bestätigten erneut die mangelnde Objektivität, Reliabilität und Validität der Schulnoten, wenngleich die Autoren vor unkritischen Versuchen warnen, „die Noten den testtheoretischen Modellen anzupassen", da diese „den pädagogischen Absichten des Unterrichtens und Erziehens widersprechen können". Diese Feststellung sollte freilich — nach übereinstimmender Meinung der Autoren und des Herausgebers — den Lehrer nicht dazu verleiten, eine (weiterhin) unkritische Notengebung zu praktizieren! Die Arbeit von FINGERHUT und LANGFELDT bietet zahlreiche Anregungen und erste Ansätze zu einer wirksamen Reform der Zensierung von Schülerleistungen. Abschließend wird auf mehr oder weniger systematische Zusammenhänge zwischen Schulnoten und Intelligenz- bzw. Persönlichkeitsmerkmalen des Schülers sowie bestimmten soziokulturellen Variablen hingewiesen.

Die beiden letzten Beiträge dieses Sammelbandes sind dem Problem der Aufsatzbeurteilung gewidmet. Nirgendwo sonst zeigt sich die Problematik der Notengebung so deutlich, wie bei der Bewertung von Aufsätzen oder aufsatzähnlichen Schülerarbeiten. Die verschiedenen *Einflüsse auf die Beurteilung von Schüleraufsätzen* sind Gegenstand einer breitangelegten Versuchsreihe, über die NICKEL und WIECZERKOWSKI hier berichten. Während für die Variablen „Geschlecht" und „Einbettung des Beurteilungsmaterials in Vergleichsserien" keinerlei Einflußwirkung auf die Aufsatzbeurteilung nachzuweisen war, konnten die „Informationen über die Ausgangssituation der Schüler" und deren „allgemeines Leistungsverhalten im Unterricht", „das Ausmaß an Lehr- und Beurteilungserfahrungen der Bewerter" sowie eine Reihe verschiedener „Sprachkriterien" (Länge des Aufsatzes, Sprachrichtigkeit vs. fehlerhafte Darstellung, attributiver Stil, Originalität der Einfälle, Differenziertheit des sprachlichen Ausdrucks und Flüssigkeit bzw. Abgeschlossenheit des Handlungsablaufs) als Einflußgrößen der Aufsatzbeurteilung und damit auf die Höhe der Aufsatzzensur eindeutig belegt werden. „Diese Ergebnisse dürften insgesamt die Bemühungen um die Erarbeitung einheitlicher und zuverlässiger Bewertungskriterien mit dem Ziel einer größeren Objektivität und Reliabilität der Aufsatzbeurteilung we-

sentlich unterstützen." Darüber hinaus stellt die Untersuchung ein Paradigma für empirische Untersuchungen gleicher oder ähnlicher Fragestellung dar, nicht zuletzt im Hinblick auf notwendige weitere Forschungsarbeiten auf dem Gebiet der Aufsatz- bzw. (allgemeinen) Leistungsbeurteilung durch Notengebung.

Die *mangelnde Übereinstimmung von Aufsatzbeurteilungen* und *Vorschläge für eine Vereinheitlichung* derselben sind Ausgangspunkt und Ziel der Ausführungen von WENDELER, der sich vor allem auf angelsächsische Untersuchungen (z. B. im Rahmen der englischen 11+ Prüfungen) bezieht. Sein Beitrag vermittelt dem Leser nicht nur einen guten Überblick zur Gesamtproblematik der Aufsatzbeurteilung, die Zusammenstellung der wichtigsten Bewertungskriterien sowie praktikable Verbesserungsvorschläge können unmittelbar zu einer befriedigenderen Beurteilungspraxis führen. Trotz einer Reihe unbestrittener Schwierigkeiten und Probleme im Hinblick auf die derzeitigen Möglichkeiten der Objektivierung von Lehrerurteilen ist nach den Ausführungen WENDELERS Fatalismus fehl am Platze. Dieses Fazit mag alle diejenigen Leser ermutigen, die aufgrund der vorhergehenden (zahlreichen) Problematisierungen vielleicht einen gegenteiligen Eindruck — zu Unrecht — gewonnen haben.

5.1. Beobachtung und Beurteilung des Schülerverhaltens im Unterricht

Erich Langhorst

5.1.1. Beobachtungsnotwendigkeit und Beurteilungspraxis

Für diagnostizierende Psychologen ist es zur Selbstverständlichkeit geworden, eine individuelle Testuntersuchung durch eine Verhaltensbeobachtung zu komplettieren. Außer den eigentlichen Meßergebnissen wie „Prozentrang" oder „Leistungsprofil" interessieren hier Anstrengungsbereitschaft, Ausdauer, Konzentration, Verhalten bei Erfolg und Mißerfolg, Erregung, kognitiver Stil und andere leistungsfördernde wie -hemmende Faktoren. Deren Berücksichtigung erlaubt vielfach erst eine adäquate Interpretation der normbezogenen Testbefunde.

Von noch größerer Bedeutung ist die Verhaltensbeobachtung in der Schule, wo der Lehrer — analog zum testenden Psychologen — die Schülerleistung im Kontext der Schülerpersönlichkeit beurteilen und — anders als der Psychologe — den Schüler zum optimalen Leistungsverhalten erziehen soll. Die Wirkungen seiner verhaltensändernden Maßnahmen kann der Lehrer dann wieder mit Hilfe von Verhaltensbeobachtungen kontrollieren. So ergibt sich ein Kreisprozeß aus Beobachtung und Hilfestellung. „Gerade diese Vertiefung der erzieherischen Einwirkung durch die Beobachtungstätigkeit ist deren eigentliche Rechtfertigung" (Thomae 1970, S. 47).

Da die Art und Weise, wie sich der Schüler im Unterricht verhält, von sehr großer Bedeutung für seinen Lernerfolg ist, wird denn auch vom Lehrer häufig verlangt, daß er das Verhalten seiner Schüler sachlich und richtig beschreibt und beurteilt: Vor allem die Eltern erwarten in dieser Hinsicht ausführliche Informationen (siehe Elternsprechtage); auf den Zeugnissen müssen „Betragen", „Führung", „Ordnung", „Beteiligung am Unterricht", „Mitarbeit", „Aufmerksamkeit", „Fleiß" oder andere Verhaltensaspekte in freier oder gebundener Form beurteilt werden; bei Überweisungen auf Realschulen, Gymnasien oder Sonderschulen ist die Erstellung eines „Gutachtens" notwendig.

Vielerorts versucht man, mit „Schülerbögen" die Entwicklung der Schüler zu erfassen. In Bayern z. B. wird jeder Schüler gemäß der VSO nach dem 2., 4., 6. und 8. Schülerjahrgang und vor Überweisung auf eine andere Schule mit Hilfe des nachfolgend wiedergegebenen Formulars zusammenfassend beurteilt, und zwar durch Unterstreichung der zutreffenden Begriffe und gegebenenfalls durch Hinzufügung weiterer Merkmale. Aus Platz-

gründen sind hier die Auswahlantwortkategorien jeweils nur bei der 1., 3., 5. usw. Verhaltensdimension angeführt:

1. Auffassungsgabe
   (sehr rasch — rasch — vorschnell — bedächtig — langsam — unsicher)
2. Beobachtungsgabe
3. Gedächtnis
   (besonders gutes Gedächtnis — gutes Gedächtnis — gutes Behalten von Zahlen, Gedichten, Sachstoffen)
4. Fähigkeiten
5. Phantasie
   (selbständig — beherrscht — blühend — mäßig — wenig anregbar — phantasiearm — auf welchen Gebieten?)
6. Gewandtheit im Gebrauch der Sprache
7. Arbeitsverhalten
   (selbständig — wenig ausdauernd — gründlich — zuverlässig — schwankend — ungenau — unselbständig)
8. Arbeitsweise
9. Aufmerksamkeit und Konzentration
   (stark konzentrierbar — beherrscht — ausdauernd — unauffällig — ablenkbar — zerstreut — konzentrationsschwach)
10. Einstellung zur Lernarbeit
11. Stimmung
    (froh — ausgeglichen — ernst — launenhaft — ausgelassen — gedrückt — mißmutig)
12. Besondere Neigungen
13. Handgeschicklichkeit
    (groß — sicher — durchschnittlich — fahrig — unsicher — gering)
14. Soziales Verhalten
15. Körperliche Entwicklung und Besonderheiten
    (Übergewicht — Untergewicht — groß — mittel — klein — Sehschwäche — Hörschwäche — Sprachstörung — Ängstlichkeit vor Prüfung — auffällige Reaktion bei schlechten Leistungen — Selbstüberschätzung — keine Auffälligkeiten)

Wenn nach diesem oder einem ähnlichen Kategorienschema Beurteilungen vorgenommen werden sollen, setzt man als selbstverständlich voraus, daß der Lehrer ein geschulter Beobachter ist und um die Fehlerquellen beim Beobachten und Beurteilen weiß. Das ist nicht der Fall. Es fehlt sogar eine wissenschaftliche Fundierung schulischen Beobachtens und Urteilens. „Seit Beginn der Unterrichtsforschung hat wohl selten ein Vorgang von so weiter Verbreitung und so einschneidenden Konsequenzen für die Betroffenen so wenig Beachtung durch die Wissenschaft gefunden wie die Persönlichkeitsbeurteilung in der Schule" (Ulich & Mertens 1973, S. 11). Noch prekärer ist die Situation, wenn man berücksichtigt, daß die bislang gewonnenen wenigen Erkenntnisse bei der Aus- und Weiterbildung von Lehrern kaum oder nur ungenügend verwertet werden.

Dieser Beitrag versucht, in gedrängter Form Möglichkeiten aufzuzeigen, wie die im großen und ganzen subjektive und unsystematische Beobachtungs- und Beurteilungspraxis in der Schule verbessert werden kann. Zu diesem Zweck werden nur solche Methoden vorgestellt, die m. E. für den Lehrberuf von vorrangiger Bedeutung sind. Eine systematische Darstellung der wissenschaftlichen Beobachtungs- und Beurteilungstechniken samt ihrer Probleme ist an dieser Stelle nicht möglich. Darüber informieren u. a. HASEMANN (1964), GRAUMANN (1966), v. CRANACH & FRENZ (1969), THOMAE (1970), TENT (1971) und HELLER et al. (1974).

Zunächst gilt es festzuhalten, daß zur wissenschaftlichen „Beobachtungsmethode prinzipiell drei — möglichst auch zeitlich zu diskriminierende — Schritte gehören: die *Beobachtung i. e. S.*, die *Beschreibung* und die *Beurteilung*" (HELLER 1974, S. 29 ff.). Ohne den ausdrücklichen Vorsatz, diese strikte Sequenz bei der Beobachtung und Beurteilung eines Menschen einzuhalten, beobachten wir allzuleicht verzerrt und urteilen voreilig. Unsere Reaktionen auf die Mitmenschen sind für gewöhnlich sehr gefühlsmäßig und vorschnell bewertend (= Vorurteile). Schon nach kurzem „Kennenlernen", z. T. aufgrund des ersten Eindrucks, „wissen" wir, ob dieser Mensch da „vertrauenswürdig", „gutmütig", „umgänglich", „egoistisch", „geltungssüchtig", „intelligent" usf. ist. Diese prompte Einschätzung und Bewertung mag in gewissen Lebenssituationen und auch Berufen (etwa bei einem Vertreter) die erfolgreichste Methode sein, im Erziehungsprozeß ist sie unverantwortlich. Dort müssen die Schüler genügend häufig, objektiv und zuverlässig, d. h. methodisch abgesichert, beobachtet werden, bevor Verhaltenskonstanten ermittelt werden können.

5.1.2. Beobachtung

Am Anfang eines objektivierten und damit kontrollierbaren Urteilsprozesses steht — wie angedeutet — eine zureichende Sammlung von Beobachtungsdaten, die mit Hilfe verschiedener Beobachtungsarten gewonnen werden können. Man unterteilt die Beobachtungsformen vielfach in „Gelegenheitsbeobachtungen" und „systematische Beobachtungen". Die *systematischen Beobachtungen* unterscheiden sich von den Gelegenheitsbeobachtungen durch genau festgelegte Beobachtungspläne (bezüglich Beobachtungssituation, -aspekt, -zeit und -dauer), ein Höchstmaß an Konzentration auf das zu Beobachtende, ein vorauslaufendes Beobachtungstraining, das die Beschreibung bzw. Kodierung mitumfaßt, und Zuverlässigkeits- wie Gültigkeitskontrollen. Die *Gelegenheitsbeobachtung* dagegen ereignet sich zufällig oder, wenn beabsichtigt, weniger methodenstreng und nicht frei von anderen Aufgaben des Beobachters. Sie ist die schulübliche und, indem sie der systematischen Beobachtung angenähert wird, die via regia der Schülerbeurteilung.

Der erste Schritt zur Effektivitätssteigerung der Gelegenheitsbeobachtungen ist der Vorsatz, sie häufiger und gezielter einzusetzen, etwa in der Form:

„Ich will in den nächsten Stunden auf den Schüler X achten, den ich als unauffälligen Schüler noch sehr wenig kenne." (= „offene", nicht aspekteingeengte Beobachtung) —
„Ich will in den nächsten Stunden auf das motorische Verhalten des Schülers X achten. Ist er wirklich motorisch so unruhig, wie seine Eltern es sagten?" (= Beobachtung unter einem bestimmten Aspekt) —
„Ich will in der nächsten Zeit darauf achten, wie sich Schüler X bei seinen Zornanfällen verhält." (= Ereignissammlung, „event-sampling") —

Für solche Beobachtungen ist es von höchster Wichtigkeit, die Bedingungen, unter denen sich das Beobachtete zeigt, mitzuerfassen (= kontrollierte Beobachtung). Bei den wissenschaftlichen, systematischen Beobachtungen ist eine hinreichende Bedingungskontrolle geradezu unerläßlich (vgl. NICKEL 1972, S. 76 ff.).

Als Variante des Typus' „systematische Beobachtung", der die schulische Gelegenheitsbeobachtung angeglichen werden kann, kommt der *fraktionierten Beobachtung* („time sampling") besondere Bedeutung zu. Bei ihr setzt der Beobachter „Dauer, Abstand und Anzahl der Intervalle fest, um eine Art repräsentative Zeitstichprobe zu gewinnen" (GRAUMANN 1973, S. 32). Je nach Situation können zwischen den einzelnen Beobachtungsstichproben Tage oder nur Minuten liegen. Im letzten Fall wird (etwa bei einer wissenschaftlichen Untersuchung) durch eine „Einwegscheibe" die einstündige Spielphase eines Kindes von mehreren Mitarbeitern in der Weise beobachtet, daß jeweils drei Minuten lang beobachtet und drei Minuten lang protokolliert bzw. pausiert wird. Analoge (fraktionierte) Beobachtungen kann der Lehrer im Unterricht, während der Schulpausen usf. durchführen.

Eine andere, traditionsreiche Beobachtungsart, welche die Kontrolle der äußeren Bedingungen und das „Sehen" der relevanten Verhaltensweisen stark erleichtert, ist die *Beobachtung in standardisierten Situationen:* Der Versuchsleiter arrangiert zu Beobachtungszwecken eine „ergiebige" Situation, die beliebig oft wiederholt werden kann, und gewinnt so mit der Zeit brauchbare Kriterien beim Vergleich verschiedener Personen in eben dieser „Modellsituation". Die Versuche von GOTTSCHALDT (1954), bei denen normal intelligente und debile Kinder mit einer begrenzten Anzahl von Vierkanthölzern einen bis zur Decke des Versuchsraumes reichenden Turm bauen sollten, und die Versuche von HECKHAUSEN u. Mitarbeitern, bei denen Kinder im Wettstreit mit dem Versuchsleiter durchlöcherte Holzscheiben auf eine Stange zu schieben hatten, sind bekanntgewordene standardisierte Beobachtungssituationen (s. WASNA 1970, BITTMANN 1973). Da in unseren Schulen die Lernsituationen z. T. erheblich standardisiert sind, bieten sich

Beobachtungsgelegenheiten an, die den Beobachtungen in „Modellsituationen" sehr nahe kommen, z. B.:

Beobachtung von Schülern beim Vortrag vor der Klasse.
Beobachtung von Schülern beim Schreiben eines Aufsatzes (Diktates).
Beobachtung von Schülern bei einem bestimmten Ballspiel.

5.1.3. Beschreibung

Nach der Erörterung der Beobachtungsmöglichkeiten gewinnt die Frage nach der Kodierung und Speicherung der Beobachtungsdaten größte Bedeutung. Bei wissenschaftlichen Untersuchungen handelt es sich um die Phase des Protokollierens bzw. der Deskription, ohne die kontrollierbare Auswertungen nicht möglich sind. HELLER (a. a. O.) bezeichnet die Beobachtungs*deskription* als „das Kernstück der Beobachtungsmethode und zugleich ihren problematischsten Teil". Schon die Übersetzung des Gesehenen und Gehörten in die Sprache hat ihre Probleme: Wortbedeutungen und Sprachschatz sind individuell sehr verschieden. Das gilt besonders, wenn außer dem „Was" des beobachteten Verhaltens auch das „Wie" beschrieben werden soll, d. h. wenn die Deskription nicht nur auf dem „verbalen", sondern auch auf dem „adverbialen Niveau" versucht wird [1]. Die Aussage „Der Schüler X lacht häufig, wenn ein Mitschüler eine falsche Antwort gibt" kann zwar durch die nähere Kennzeichnung „... lacht häufig, und zwar sehr laut und beleidigend ..." wesentlich informativer gefaßt werden, — ist aber weniger objektiv.

Bei dieser Art der Attribuierung handelt es sich letztlich um Einschätzungen bestimmter Ausprägungsgrade von Verhaltensweisen (etwa: sehr laut, laut, ... sehr leise) und deren eventuelle Auswirkungen (in unserem Beispiel „beleidigend"). Wie angedeutet, erfolgen derartige Charakterisierungen von Verhaltensweisen z. T. nach individuellen Maßstäben. Diese dürfen nicht übersehen, sollten aber auch nicht überbewertet werden, etwa mit der Konsequenz des Verzichtes auf adverbiale Beschreibungen.

Unabhängig von der Problematik der Verbalisierung der Beobachtungen besteht die der Speicherung. Genügt es, wenn der Lehrer sich vornimmt, das jeweils Beobachtete zu behalten, zumindest das sich wiederholt Zeigende? Oder sollte er sich die Zeit nehmen, Beobachtungsnotizen zu machen? Untersuchungsergebnisse, die es erlauben, diese Frage zu beantworten, liegen nicht vor. Die Entscheidung ist dem Lehrer überlassen. DONAT (1970) emp-

---

[1] *Graumann* (1960) unterscheidet folgende Modi der Beschreibung (S. 90 ff.):

1. den *verbalen* Modus	(A. löst das Problem)
2. den *adverbialen* Modus	(A. löst das Problem schnell)
3. den *adjektivischen* Modus,	(A. ist sehr intelligent)
4. den *substantivischen* Modus	(Die Intelligenz von A. ist ...)

fiehlt den Lehrern, die „sachlichen Beobachtungen in sachlicher Form schriftlich festzuhalten ... Das geschieht am besten mit Hilfe eines Beobachtungsheftes" (S. 121). Einmal im Jahr sollten dann die Beobachtungsnotizen in einen von demselben Autor entworfenen „Beobachtungsbogen", der in diesem Fall ein Formular der Schule wäre, übertragen werden, und zwar unter fünf Gesichtspunkten (S. 126 f.):

„1. die äußere Erscheinung des Kindes
(Körperbeschaffenheit, Kleidung und Körperpflege, Bewegungen und Sprechweise);
2. die Familienverhältnisse
(In wirtschaftlicher und geistig-seelischer Hinsicht: Anregungen, Hilfen, Hemmungen, Nachwirkungen früherer Erlebnisse);
3. das Verhalten im Unterricht
(Allgemeines Arbeits- und Aufmerksamkeitsverhalten, Auffassungs- und Denkweise, Begabungen, Fähigkeiten und Fertigkeiten);
4. das Allgemeinverhalten
(Triebe und Strebungen, Interessen und Neigungen, Lebensgrundstimmung, Selbstgefühl, Gemeinschaftsgefühl, Willensartung) [2];
5. die Leistungen
(Evtl. notwendige Begründungen und Erläuterungen zum Verständnis der einzelnen Leistungen des Kindes)."

Der Autor betont, daß es ihm „nur um sachliche Feststellungen und nicht um Deutungen oder Wertungen" gehe (S. 125). In solche gleitet man aber leicht und unbemerkt über, wenn man etwa Strebungen, Interessen, Neigungen oder das Sebstgefühl „beschreibt". Diese sind Verhaltensursachen und werden aus Verhaltensbeobachtungen erst *erschlossen*. Sie gehören also entsprechend unserer Trias (Beobachtung i. e. S., Beschreibung, Beurteilung bzw. Deutung) in den dritten und abschließenden Teil. Das mag zunächst überraschen. Aber: Wie gelange ich zu Angaben über das „Selbstgefühl"? Ich folgere aus bestimmten Verhaltensweisen (etwa: ruhiges Reagieren bei Problemen und Mißerfolgen) auf ein sicheres Selbstgefühl. Nun führen solche Schlußfolgerungen leicht zu Irrtümern, da das Verhalten vielfach mehrdeutig ist. Das angeführte „ruhige Reagieren" ist in Wirklichkeit vielleicht die Folge eines gehemmten und übersteuerten Verhaltens. Ein weiteres Beispiel möge die Notwendigkeit, Beschreibung und Deutung auseinanderzuhalten, unterstreichen: Schon die Feststellung „Werner ging ängstlich zur Tafel" enthält eine die Beschreibung überspringende Deutung. Aus Symptomen (etwa: langsames und zögerndes Gehen) wird auf „ängstlich" geschlossen. Vielleicht aber dachte Werner intensiv über das Anzuschreibende nach und ging deshalb langsam und zögernd. Ängstlich war er gar nicht.

[2] Diese Begriffe entstammen der deutschen Charakterologie der Vorkriegszeit und werden in der gegenwärtigen Psychologie z. T. nicht mehr verwendet.

Nun ist das Schlußfolgern bzw. Deuten nicht immer so problematisch wie in den genannten Beispielen. Aufgrund vieler Beobachtungen kann man syndromatisch feststellen, daß Schüler X in der Schule „sehr ängstlich" ist. Während der Einzelbeobachtung selbst und während ihrer sprachlichen Formulierung sind Deutungen oder Schlußfolgerungen *(Inferenz*-Kategorien) aber verfrüht.

Die Gefahr, unbemerkt auf der Stufe der Beobachtung und Beschreibung Verhaltensursachen „mitzusehen" und zu vermerken, kann durch spezielle *„Zuordnungsverfahren"* vermieden werden. Bei solchen „gebundenen" Formen der Deskription müssen die beobachteten Verhaltensweisen bestimmten, vorformulierten Verhaltenskategorien zugeordnet und dort angestrichen werden.

Ein von PUCKET (1928) entwickeltes Aufzeichnungsschema zur Registrierung der Schülermitarbeit macht deutlich, wie objektiv und urteilsfrei, aber notwendigerweise auch eingeengt eine standardisierte Beobachtungsbeschreibung sein kann (zit. nach SCHULZ et al. 1971, S. 664):

Schüler meldete sich.
Schüler meldete sich und wurde vom Lehrer drangenommen.
Schüler meldete sich, kam dran und gab eine Ein-Wort-Antwort.
Schüler meldete sich, kam dran und gab eine befriedigende Antwort.
Schüler meldete sich, kam dran und gab eine gute Antwort.
Schüler meldete sich, kam dran und gab eine sehr gute Antwort.
Schüler kam dran, ohne sich gemeldet zu haben.
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine Ein-Wort-Antwort.
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine befriedigende Antwort.
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine gute Antwort.
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine sehr gute Antwort.
Schüler kam dran, ohne sich gemeldet zu haben, und gab keine Antwort.
Schüler stellte eine Frage.
Schüler sprach ohne Aufforderung des Lehrers.

Bei Schülerbeobachtungen mit Hilfe dieser oder ähnlicher Listen wird fortlaufend die jeweils zutreffende Verhaltenskategorie signiert. Eine solche — freilich auch nicht immer rein deskriptive — Protokollierung (wann ist eine Antwort „gut" oder „sehr gut"?) kann unter den derzeitigen schulischen Bedingungen nur ein zusätzlicher Beobachter durchführen und ist deshalb in der Unterrichtspraxis — anders als in der Unterrichtsforschung — kaum praktikabel.

Praxisnäher ist das Kategoriensystem von BALES (1972), das mit zwölf verbalisierten Verhaltensvarianten die Interaktionen in einer Gruppe zu registrieren erlaubt (S. 154):

1. Zeigt *Solidarität,* bestärkt den anderen, hilft, belohnt.
2. *Entspannte Atmosphäre,* scherzt, lacht, zeigt Befriedigung.
3. *Stimmt zu,* nimmt passiv hin, versteht, stimmt überein, gibt nach.

4. *Macht Vorschläge,* gibt Anleitung, wobei Autonomie des anderen impliziert ist.
5. *Äußert Meinung,* bewertet, analysiert, drückt Gefühle oder Wünsche aus.
6. *Orientiert,* informiert, wiederholt, klärt, bestätigt.
7. *Erfragt Orientierung,* Information, Wiederholung, Bestätigung.
8. *Fragt nach Meinungen,* Stellungnahmen, Bewertung, Analyse, Ausdruck von Gefühlen.
9. *Erbittet Vorschläge,* Anleitung, mögliche Wege des Vorgehens.
10. *Stimmt nicht zu,* zeigt passive Ablehnung, Förmlichkeit, gibt keine Hilfe.
11. *Zeigt Spannung,* bittet um Hilfe, zieht sich zurück.
12. *Zeigt Antagonismus,* setzt andere herab, verteidigt oder behauptet sich.

Den einzelnen Gruppen teilte BALES geschulte Beobachter zu, die während der Beobachtungszeit die jeweils zutreffende Verhaltensklasse für jedes Gruppenmitglied notierten. Im Bereich der Forschung hatte das Verfahren von BALES eine Auslöserfunktion. Es regte zur Konstruktion ähnlicher Beobachtungs-, Beurteilungs- und Registriertechniken an. Für die Schulpraxis dagegen fehlt es weithin an analogen Klassifizierungsmöglichkeiten des Schülerverhaltens. Es wäre m. E. vordringlich, folgende Verhaltensbereiche mit Zuordnungsverfahren (Kategorienlisten) zu erfassen: Auffassungstempo, Lern- und Leistungsmotivation, Mitarbeit im Unterricht, Aufmerksamkeit, Ausdauer, Erregbarkeit (emotionale Stabilität), Arbeitstempo, Arbeitssorgfalt und Verhalten zu Lehrern und Mitschülern. Jede dieser Beobachtungsdimensionen müßte — wie angedeutet — durch operational definierte Verhaltensklassen differenziert beschrieben werden. ROTH (1969) sah die Notwendigkeit, „Verhaltensskalen" aufzustellen, schon früh. Als Beispiel für eine solche Skala stellte er Verhaltensweisen zusammen, die die Charakterisierung bzw. Beurteilung der „Über-, Ein- und Unterordnung" fundieren helfen:

„a) Sucht bei jeder Gelegenheit im Mittelpunkt zu stehen, ist ärgerlich, wenn er eine Nebenrolle zugewiesen bekommt.
b) Ordnet sich den Stärkeren unter, sucht aber die Schwächeren zu beherrschen.
c) Ordnet sich ein, verteidigt sich aber energisch, wenn er angegriffen wird oder seine Rechte geschmälert werden.
d) Sucht mit allen auszukommen und vermeidet Konflikte.
e) Ist von den Rädelsführern in der Klasse abhängig, ja läuft ihnen nach.
f) Ordnet sich immer ein und unter.
g) Wird ständig von anderen verführt, ist äußerst suggestibel und nachgiebig" (S. 36).

Einen systematischen Ausbau dieses Ansatzes hält ROTH aber nicht für realisierbar. „Solche Beobachtungsskalen in den Beobachtungsbogen einzubauen, würde diesen ins Unübersichtliche ausweiten, die Beispiele können aber den Beobachter anregen, in dieser Weise seine Beobachtungen zu objektivieren" (S. 36). Nun — bei einem derart detaillierten „Beobachtungsbogen", wie ROTH ihn vorstellte (über 100 Einzelaspekte), ist eine Operationa-

lisierung der einzelnen Merkmale selbstverständlich ausgeschlossen. Möglich ist aber die Ausarbeitung von Verhaltensskalen für besonders wichtige Verhaltensbereiche, z. B. für die oben genannten. Dem Klassenlehrer stünde auf diese Weise ein brauchbares und zuverlässiges Beobachtungsgut für die Beurteilung des Schülerverhaltens zur Verfügung, vergleichbar den Untersuchungsbefunden in der ärztlichen Praxis. Neben der Funktion, test- oder schulleistungsdiagnostische Urteile (z. B. Zensuren) hinreichend abzusichern, käme derartigen Beobachtungsdaten auch eine besondere Bedeutung für die Aufdeckung von Verhaltens- und Leistungsstörungen zu.

Probleme solcher Daten*auswertungen* — die eigentliche Schülerbeurteilung — sollen im folgenden erörtert werden.

5.1.4. Beurteilung

Zur Beurteilung zählen hier *zusammenfassende, abstrahierende Verhaltensbeschreibungen* (geringe Mitarbeit — motorisch sehr unruhig — leicht anregbar — unzuverlässig — leicht entmutigt), *Deutungen des Verhaltens* (vermutlich überfordert — verläßt sich zu sehr auf die Hilfe seiner Eltern — die mangelhafte Konzentration ist vermutlich ein Überforderungssymptom) und der *Vergleich* der Beobachtungsergebnisse *mit Milieueinflüssen und Lebenslaufdaten* (die schlechten Rechtschreibleistungen stehen wohl im Zusammenhang mit dem schlechten häuslichen „Sprachmilieu"). Ein derartiges Beurteilen der Schüler ist ein erziehungs- und unterrichtsimmanenter Prozeß. Er muß möglichst fundiert (viele Beobachtungsdaten), bewußt (Beobachtungen, Verbalisierungen und Urteile trennend) und vorsichtig ablaufen. Zumal die Deutungen und Rückführungen des Schülerverhaltens auf bestimmte Erfahrungen sollten stets in der Schwebe gehalten werden. Es unterlaufen leicht Fehlschlüsse. THOMAE (1970) bringt instruktive Beispiele für „Kurzschlußprozesse" bei Beurteilungen. Hier ein Modell-Beispiel des Autors: „Man geht von einem Detail der Beobachtungsdaten — z. B. ‚ängstliches Verhalten im Klassenverband' — aus, kombiniert dieses Detail mit Informationen über die soziale und persönliche Entwicklung des Falles — z. B. ‚der Vater hat das Kind öfter hart gestraft' oder ‚auch die Mutter hatte eine ängstliche Art' — und kommt von hier aus unmittelbar zu einer ‚Diagnose' — z. B. ‚neurotische Fehlentwicklung' bzw. ‚konstitutionell verankerte Fehlreaktion'" (S. 62). Derartige kurzgeschlossene Urteile fand THOMAE auch bei der Durchsicht von Gutachten aus Erziehungs- und schulpsychologischen Beratungsstellen. Sie sind nicht schulspezifisch, d. h. nicht auf den Schulbereich beschränkt.

Außer der Voreiligkeit bei Urteilsprozessen muß die *Einseitigkeit* derselben bewußt vermieden werden. Die „Persönlichkeitsbilder" der Schüler sind nicht selten auf die Aspekte der „Schulleistung" und „Schuldisziplin" ein-

geengt und somit eine unzureichende Basis für Verständnis und Hilfestellung. Eine Ausweitung der Beobachtungsrichtungen und damit auch der Urteilsgesichtspunkte kann durch Beurteilungsbögen wie den für bayerische Grund- und Hauptschulen oder den von BLEIDICK (1972) für Sonderschulen herausgegebenen erreicht werden. Der zuletzt genannte ist aspektenreicher und durchgearbeiteter, weshalb er hier vollständig wiedergegeben wird. *Persönlichkeitsdimensionen* nach BLEIDICK (a. a. O., S. 16 f.):

Körperliche Erscheinung	Intelligenz	Begabung	Leistungsbeeinflussende Faktoren
Physiognomie Ernährungsstand Erscheinung äußeres Gebaren usw.	1 extrem hoch 2 sehr hoch 3 hoch 4 durchschnittlich 5 niedrig 6 sehr niedrig 7 extrem niedrig	verbal praktisch musisch usw.	durch: Aktivität Anregbarkeit Steuerung Selbstgefühl Anpassungsweisen

Schulleistungen Les. Schr. Rech.	Entwicklungsstand	Auffassung	Arbeitsweise
1 extrem hoch 2 sehr hoch 3 hoch 4 durchschnittlich 5 niedrig 6 sehr niedrig 7 extrem niedrig	1 primitiv 2 einförmig 3 z. T. unentfaltet 4 differenziert 5 hoch integriert 6 zersplittert 7 aufgelöst	1 verständnislos 2 schwerfällig 3 langsam 4 leicht 5 oberflächlich 6 unkritisch 7 verblendet	1 untätig 2 stockend 3 langsam 4 stetig 5 ungenau 6 übereilt 7 ziellos

Gedächtnis	Konzentration	Sprache	soziale Einordnung
1 mangelhaft 2 zusammenhanglos 3 mechanisch 4 genau 5 oberflächlich 6 ungetreu 7 verworren	1 unkonzentriert 2 ablenkbar 3 unaufmerksam 4 konzentriert 5 überwach 6 verhaspelt 7 abgeblendet	1 kleinkindlich 2 dürftig 3 unsicher 4 altersgemäß 5 altklug 6 unkontrolliert 7 gestört	1 isoliert 2 einzelgängerisch 3 zurückhaltend 4 eingeordnet 5 dominant 6 anlehnungsbedürftig 7 gesellschaftsbedürftig

soziale und erziehliche Reaktion	Antrieb	Steuerung	Stimmung
1 übergefügig	1 apathisch	1 hemmungslos	1 traurig
2 fügsam	2 lahm	2 unbeherrscht	2 mißgestimmt
3 folgsam	3 mäßig aktiv	3 teilgesteuert	3 ernst
4 selbständig	4 aktiv	4 gesammelt	4 ausgeglichen
5 schwierig	5 sehr aktiv	5 beherrscht	5 froh
6 oppositionell	6 impulsiv	6 gespannt	6 heiter
7 aggresiv	7 umtriebig	7 verkrampft	7 lustig

Selbstgefühl	Anregbarkeit	Gemüt	Angepaßtheit
1 selbstzweiflerisch	1 stumpf	1 gemütlos	1 rücksichtslos
2 ängstlich	2 indolent	2 kalt	2 egozentrisch
3 unsicher	3 unempfindlich	3 nüchtern	3 unbekümmert
4 sicher	4 ansprechbar	4 gemütsfähig	4 bedachtsam
5 selbstbewußt	5 beeindruckbar	5 warm	5 bemüht
6 großspurig	6 sensibel	6 weich	6 versiert
7 überheblich	7 übererregbar	7 selbstverschwenderisch	7 hektisch

Halt	Daseinstechnik und Daseinsthematik
1 haltlos	Hauptanliegen
2 labil	Hauptnöte
3 unsicher	Lebensstil
4 gefestigt	usw.
5 selbstsicher	
6 verrannt	
7 starr	

Die Persönlichkeitsdimensionen und Eigenschaftslisten dieses Schemas werden in einer Anleitung interpretiert. Dadurch gewinnen die vorgegebenen Ober- und Auswahlbegriffe an Eindeutigkeit. Sie werden zudem durch die Aufzählung weiterer Eigenschaftsbegriffe ergänzt. Anhand des aufgeführten Beurteilungsschemas und seines Kommentars kann — genügend Beobachtungsdaten vorausgesetzt — eine freie Charakteristik geschrieben werden. Es ist aber auch möglich, lediglich durch „Ankreuzen" und Ausfüllen der Leerstellen des Schemas einen Schüler zu beurteilen. Bei den Dimensionen, die BLEIDICK numeriert unterteilte, handelt es sich um sog. Schätzskalen *(rating scales)*. Es ist zu erwarten, daß sie in Zukunft stärker zum Einsatz gelangen, z. B. zwecks objektivierter Effektivitätskontrollen unterschiedlicher Unterrichtsformen.

Mit Hilfe von *Schätzskalen* können Beobachtungen unter Merkmalsbegriffe (Beschreibungs- und Interferenzkategorien) gebracht und eingestuft werden. Das geschieht selbstverständlich nicht frei von subjektiven Urteilstendenzen, kommt aber einer objektiven Beurteilung um so näher, je eindeutiger die unterschiedlichen Items beschrieben werden. Auf eben diese Eindeutigkeit, d. h. auf die Inter-Rater-Übereinstimmung hin, müssen die Beurteilungsskalen sehr kritisch geprüft werden. Bei *numerischen* Skalen werden die Beobachtungen lediglich in Zahlenwerte transformiert: Etwa: „1" bedeutet minimale und „7" maximale Ausprägung des jeweils definierten Merkmals (= unipolare Skala). Oder: „—3" ist der Negativpol und „+3" der Positivpol eines Merkmals (= bipolare Skala). Es ist offensichtlich, daß numerische Skalen in dem Maße zuverlässiger und treffsicherer werden, wie ihre Zielbereiche (etwa „Auffassung" oder „Selbststeuerung") und deren Unterteilungen klar definiert sind. Wenn das der Fall ist, handelt es sich um eine *Kombination von numerischen und verbalen* Skalen. Eine solche verwandte RYANS (1960) in seinem bekannt gewordenen Forschungsprojekt zur Erfassung der „Characteristics of teachers". Seine Beobachter hatten das Lehrer- und Schülerverhalten mit den Ziffern 1 bis 7 zu bewerten. Z. B. *Schülerverhalten:*

gleichgültig 1 2 3 4 5 6 7 lebhaft

Die einzelnen bipolaren Eigenschaftsdimensionen (gleichgültig-lebhaft, unsicher-selbstsicher usw.) wie auch deren Zwischenwerte wurden in den Instruktionen für die Beurteiler sorgfältig zu kennzeichnen versucht. Die „Begrenzungsanker" der Skala „gleichgültig-lebhaft" definierte RYANS wie folgt (nach TENT 1971, S. 867):

gleichgültig	lebhaft
1. Teilnahmslos.	1. Darauf bedacht, dranzukommen und teilzunehmen.
2. Spielte(n) den Gelangweilten.	2. Beachtete(n) den Lehrer aufmerksam.
3. War(en) nur mit halbem Herzen bei der Sache.	3. Arbeitete(n) konzentriert.
4. Unruhig.	4. Schien(en) eifrig mitzumachen.
5. Schweifende Aufmerksamkeit.	5. Sofort bereit, sich zu beteiligen.
6. Kam(en) nur langsam in Gang.	

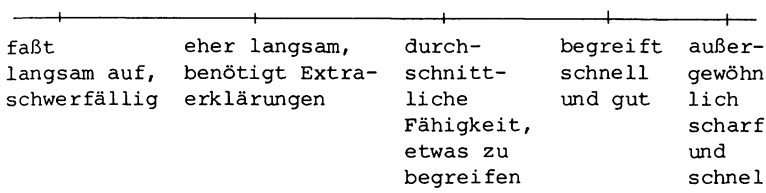
Rein formal ist die Zeugnisbenotung („1" = sehr gut = weit über gut hinausgehend, „2" = gut = wesentlich über dem Durchschnitt stehend, usw.)[3] dem Verfahren von RYANS ähnlich. Sie unterscheidet sich aber von diesem durch die abstraktere und vieldeutigere, d. h. unklare Definition der einzelnen Notenwerte.

Die Eindeutigkeit der Items entscheidet auch über den Wert der *graphischen* Skalen. Bei ihnen werden entlang einer vertikalen oder horizontalen Geraden die als zutreffend beurteilten Stellen markiert.

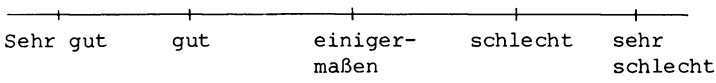
[3] Festgesetzt durch die „Ständige Konferenz der Kultusminister".

Die „Auffassung" läßt sich graphisch etwa so skalieren:

4)



Weniger beschreibend ist die nachfolgende Schätzskala "Wie konzentriert sich der Schüler im Unterricht?"



Der Vorteil der graphischen Skalen liegt darin, daß man an jeder beliebigen Stelle ankreuzen kann.

Wie wenig mit Skalen erreicht wird, die minimal präzisiert sind, zeigt folgendes Beispiel: Bei einer Fragebogenerhebung im Rahmen der „Wissenschaftlichen Arbeitsgemeinschaft für Jugendkunde" beantwortete ein Lehrer auf die Frage „Wie konzentriert sich der Schüler? Gut — normal — schlecht" 53,8 % seiner Schüler, ein anderer Lehrer 0 % seiner Schüler als „schlecht" (THOMAE 1960, S. 26).

Was nun den Einsatz von Beurteilungsskalen im schulischen Raum angeht, bedarf es sehr sorgfältiger Überlegungen, welche Urteilsdimensionen für Zeugnisse, Überweisungen und andere Zwecke, etwa bei Lern- und Erziehungsstörungen, notwendig sind. Sicherlich gilt der Grundsatz: Soviel wie nötig, so wenig wie möglich!

In einem zweiten Entwicklungsschritt müßte untersucht werden, wie die Oberbegriffe der Beurteilungsskalen (z. B. „Antrieb", „Ichstärke [5], „emotionale Stabilität", „erzieherische Reaktion", „Frustrationstoleranz", „Mitarbeit") und ihre Abstufungen zu definieren sind. Die weiter oben geforderten „Verhaltens-Zuordnungs-Skalen" wären dann objektive Unterlagen für die Eintragungen in die Beurteilungsskalen.

---

[4] Eine „descriptive graphic rating scale" aus einem Beurteilungsformular für Mitarbeiter (nach *Cronbach* 1970, S. 578).

[5] *Ulich* & *Mertens* (1973, S. 128 ff.) beschrieben kürzlich ihren Versuch einer „Entwicklung eines Beurteilungsverfahrens zur Einschätzung von ‚Ichstärke'". Die psychoanalytisch orientierte Operationalisierung der Ichstärke ist — wie die Autoren selbst erkennen — noch einseitig, im Zusammenhang mit der sorgfältigen Auswahl und Skalierung der einzelnen Indikatoren (Dimensionen) der Ichstärke sicherlich aber sehr lehrreich.

Diese Anregungen sind selbstverständlich Höchstforderungen im Sinne einer Idealkonstruktion, die auch bei nicht 100%iger Erfüllung ihren Wert darin haben können, daß sie die Misere der „Ist-Lage" deutlicher machen und Anstöße in Richtung auf die „Soll-Lage" geben können.

Wie subjektiv derzeit geurteilt wird, zeigen z. B. die sog. Kopfnoten der Zeugnisse. Hier einige „Gelegenheitsbeobachtungen" des Verf.:

1. Der Klassenlehrer eines Gymnasiums (OII) erklärte seinen Schülern, daß jeder in „Betragen" und „Ordnung" eine „2" bekäme. Er wisse nicht, wonach er urteilen solle. Über die Noten in „Fleiß" und „Aufmerksamkeit" könne diskutiert werden. — Sie streuten zwischen „3" und „1".
2. Der Leiter einer Schule für schwer erziehbare Kinder erklärt: „Warum haben die (die überweisende Schule) dem Jungen in Führung eine „6" gegeben. Eine „3" hätte auch genügt. Dann hätten wir schon gewußt, was los ist."
3. Vierter Schülerjahrgang: „. . .Sein zuverlässiger Fleiß und sein angenehmes, natürliches Wesen gefielen sehr."

5.1.5. Beurteilungsfehler

Die Erfolge bei den Bemühungen, den einzelnen Schüler — in welcher Form auch immer — richtig zu beurteilen, hängen nicht zuletzt auch davon ab, ob sich der Lehrer seiner subjektiven Urteilstendenzen bewußt ist und sie unter Kontrolle zu halten vermag. Ein kurzer Aufweis der subjektiven Fehlerquellen bei der Beobachtung und Beurteilung von Schülern ist deswegen wohl unerläßlich und im Rahmen unserer Thematik auch üblich; siehe HASEMANN (1964), v. CRANACH & FRENZ (1969), CRONBACH (1970), THOMAE (1970), ULICH & MERTENS (1973).

In welchem Ausmaß die Wahrnehmung von Personen emotionale Stellungnahmen, Wünsche, eigene Erfahrungen und Probleme usf. aktualisieren kann, zeigt folgendes „Lehrexperiment", das ich mit Pädagogikstudenten (ab 4. Sem.) durchführte:

In einem verdunkelten Raum wurde das Foto eines 28jährigen Mannes gezeigt. Die Versuchsinstruktion lautete: „Der Mann sitzt Ihnen während einer längeren Zugfahrt gegenüber und liest. Hin und wieder schauen Sie zu ihm hin. Welchen Eindruck haben Sie von ihm? Schreiben Sie bitte ohne Auswahl auf, wie Sie auf ihn reagieren, an was Sie denken, was Ihnen ‚einfällt'!"

Aus den Protokollen einige Beispiele, die jeweils von einem anderen Studierenden stammen:

„Könnte ein Theologe sein, der dauernd überzeugen möchte." —
„Er ist bestimmt ein begeisterter Sportler." —
„Wahrscheinlich glaubt er an ein höheres Wesen, ohne konfessionell gebunden zu sein. . . . Dackelbesitzer." —
„Liebt Gemütlichkeit und behagliche Unordnung; . . . kann auch öfters über den Durst trinken." —

„Sitzt mit übereinander geschlagenen Beinen; dies macht für mich einen Mann allerdings unsympathisch." —
„Man muß ihm mit derselben Sicherheit, in der er steht, oder vorgibt zu stehen, begegnen." —
„Er wird nicht gerade temperamentvoll sein und bei einer Gesellschaft nicht der Wortführer, — keine Stimmungskanone." —
„Helle, nicht allzu breitrandige Brille soll Gebrechen möglichst unauffällig zeigen." —
„Geht gerne ins Konzert." —
„Er könnte Lehrer sein." —
„Er wird manchmal Temperamentsausbrüche haben." —

Diesen Versuch, der die subjektiven Interpretationstendenzen bei der Personwahrnehmung stark begünstigt, habe ich in vielen Varianten durchgeführt und immer wieder die sehr große Streuung der „Betrachterreaktionen" aufzeigen können. Nachdem einer Gruppe von Studenten mit einem derartigen Experiment die Subjektivität ihrer „Eindrücke" demonstriert worden war, wurde sie gebeten, beim nächsten Versuch zurückhaltender zu sein. Jeder sollte nur solche Urteile aufschreiben, deren er sich relativ sicher war. Dann beobachtete die Gruppe 45 Minuten lang ein 9jähriges Mädchen, das sich mit dem VL unterhielt, an die Tafel schrieb und zeichnete (eine Frau und eine Wabenmusterfortsetzung) und den GOTTSCHALDTschen Turmbauversuch durchführte. Die Übungsteilnehmer stimmten nach dem Bericht einer studentischen Auswertergruppe in folgendem größtenteils überein:

„Sorgfältiges Bauen und Schreiben, kleine Tafelschrift und Zeichnungen, hohes Sprachniveau, beim Bauen des hohen Turmes etwas ängstlich, reagierte schnell und richtig auf die ‚Denkanstöße' des VL, schnelle Auffassung."

In diesem Bereich relativ hoher Übereinstimmung und in anderen weniger übereinstimmend beurteilten Verhaltens- und Leistungsbereichen kam es z. T. zu konträren Aussagen:

„gehemmt" — „nicht gehemmt"; „technisch bewandert" — „im Umgang mit der Technik hilflos"; „befangen" — „fröhlich und unbeschwert"; „sagt, was sie denkt" — „redet nach dem Mund"; „gutes Erkennen des Wabenmusters" — „erkannte das Wabenmuster schlecht"; „sehr intelligent" — „nicht intelligent".

Die folgenden Urteile der Beobachter spiegeln vielleicht zeitbedingte Einstellungen wider:

„wahrscheinlich autoritär erzogen", „ist typisch als Mädchen erzogen", „könnte aus gehobener Unterschicht kommen", „sie wird später vielleicht dazu neigen, keine eigene Kreativität an den Tag zu legen".

In den Protokollen der beiden „Lehrversuche" stößt man auf sehr verschiedenartige Urteilsfehler. Der bekannteste und meist zitierte Beobachtungs- und Urteilsfehler ist der *Hof-* oder *Halo-Effekt*, der „Erzfeind jeder

objektiven Beobachtung" (MEDLEY & MITZEL). Er kommt dadurch zustande, daß die wertende Einstellung zu einem Menschen dessen Eigenschaften positiv oder negativ einfärbt. THORNDIKE (1920) nannte die Tendenz, Teilurteile über eine Person aufgrund des allgemeinen Gesamteindruckes, den man von ihr hat, abzugeben „halo effect" (HASEMANN 1964, S. 34). Andere Autoren erwähnen als Ursache des Halo-Effektes außer dem Gesamteindruck auch hervorstechende Eigenschaften oder besondere Charakteristika der zu beurteilenden Person. Die positive oder negative Überstrahlung erfolgt zumeist aufgrund von Sympathie oder Antipathie. In diesem Zusammenhang ist vielleicht für Pädagogen eine Untersuchung von LEWIS (1947) interessant: Er bat Lehrer, Eigenschaften ihrer Schüler einzuschätzen. Dabei ergab sich: „Die 0,74 % der als intellektuelle Genies eingestuften Kinder bewerteten die Lehrer in jeder Weise als außergewöhnlich. Dagegen erhielt eine Gruppe von Kindern, die als Problemfälle galten, kaum irgendwelche günstigen Urteile. Tatsächlich aber waren deren Schulleistungen durchaus hinreichend" (GUILFORD 1971, S. 140). FENNER (1973, S. 132) führt zum Nachweis des Halo-Effektes im Bereich der Schule die umfangreichen empirischen Untersuchungen von KEMMLER (1967) über „Erfolg und Versagen in der Grundschule" an und macht auf ein Teilergebnis aufmerksam, demzufolge „unangepaßte und daher den Lehrern unsympathische Schüler häufig als unbegabt, faul und desinteressiert beurteilt" wurden, „obwohl dieser Zusammenhang bei objektiven Messungen zum Beispiel mit Intelligenztests nicht" bestand.

Mit dem Halo-Fehler verwandt und von diesem nicht immer unterscheidbar ist der sog. *„logische Fehler"*. GUILFORD, auf den die Bezeichnung dieses Begriffes zurückgeht, beschreibt ihn so: „Man macht einen logischen Fehler bei Einschätzungen, wenn man dazu neigt, solche Wesenszüge ähnlich zu bewerten, die einem logisch ähnlich oder zusammenhängend erscheinen. Dies kann geschehen, wenn ein Beurteiler ... irgendwelche festen, aber falschen Vorstellungen von der Persönlichkeitsstruktur hat" (1971, S. 142).

Seit einiger Zeit werden die einschlägigen Probleme mehr unter dem Begriff der *„impliziten Persönlichkeitstheorie"* gefaßt. In Auswirkung dieser privaten Theorie der Eigenschaftskorrelationen erklärt es sich, „daß wenige Informationen über einen Mitmenschen durch Eigenschaftsschlüsse zu einem subjektiv abgerundeten Bild über ihn führen können" (HOFER 1970, S. 198). Als der zitierte Autor Lehrer mit Hilfe von 25 Eigenschaftsbegriffen Schüler beurteilen ließ, um so die implizite Persönlichkeitstheorie von Lehrern zu erfassen, folgerte er aus seinen Experimenten, „daß Lehrer ihr vereinfachtes Bild-Schema von Persönlichkeitszusammenhängen bei Schülern unabhängig von den tatsächlichen Zusammenhängen der Eigenschaften bei den beurteilten Schülern in die Beurteilung eingehen lassen" (S. 207). Diese Schlußfolgerungen überschreiten m. E. die Aussagekraft von Versuchen, in

denen mit einer begrenzten Anzahl von Eigenschaftsbegriffen experimentiert wurde. Damit ist nicht bestritten, daß bei Schülerbeurteilungen durch Lehrer implizite Persönlichkeitstheorien wirksam sind. Auch HÖHN (1967) stieß bei der Auswertung von Lehrerurteilen über „schlechte Schüler" auf solche Interpretationsmodelle. Sie resümiert: „Versuchen wir, das Wesentliche des Lehrerbildes vom schlechten Schüler zusammenzufassen, so ist vor allem hervorzuheben, daß für die Lehrer das Schulversagen in erster Linie ein Nichtwollen aus Faulheit und Interesselosigkeit heraus ist, erst in zweiter Linie ein Nichtkönnen aus Begabungsmangel." Die Wirkung des negativen Stereotyps „geht aber nicht so weit, daß es blind gegenüber der Realität machen würde. Da, wo die Lehrer konkrete schlechte Schüler ihrer eigenen Klasse beurteilen, ergibt sich ein vielfältigeres und differenzierteres Bild, das durchaus Platz für individuelle Eigenart läßt ... Die vorwiegend negative Schilderung, vor allem die Kombination von Faulheit und Dummheit, dominiert allerdings auch hier" (S. 103 f.). Da die Ergebnisse HÖHNS auf Untersuchungen aus den Jahren 1959 und 1960 basieren, repräsentieren sie vielleicht nicht mehr voll die Einstellungen der jetzt unterrichtenden Lehrer.

Neben den genannten subjektiven Urteilstendenzen, die selbstverständlich auch selegierend und interpretierend die Beobachtungen selbst (also schon die Wahrnehmung) verzerren, macht sich der *„Mildefehler"* („generosity error" oder „error of leniency") recht harmlos aus. Gemeint ist die Tendenz bestimmter Urteiler, überdurchschnittlich gut zu beurteilen. Andere Urteiler dagegen lassen sich als „harte" einstufen. Bei einschlägigen Untersuchungen — nicht mit Lehrern — konnte nachgewiesen werden, „daß Beurteiler recht konstant bei ihren milden oder harten Schätzungen verharren" (GUILFORD 1971, S. 139). Diese Fehler sind bekannt und durchsichtig, nicht ganz so der *„Kontrastfehler"* (MURRAY) und *„Ähnlichkeitsfehler"* (GUILFORD). Im ersten Fall besteht eine Tendenz, „der zu beurteilenden Person seiner eigenen Wesensart gegenteilige Eigenschaften oder gegenteilige Ausprägung von Merkmalen beizulegen", im zweiten Fall die fälschliche „Annahme des Beurteilers, die von ihm beurteilten Personen seien genau so geartet wie er selbst" (HASEMANN 1964, S. 35). Die Ähnlichkeitsfehler sind eine Variante der *„Projektionsfehler"* (FREUD). Bei diesen werden eigene Einstellungen, Wünsche, Probleme, Fehler usf. in andere Menschen verlegt, „hineingesehen". Der weiter oben berichtete Versuch mit den Porträteinschätzungen brachte Beispiele für Projektionen. Diese Mechanismen zählen zu den früh erkannten und oft beschriebenen Fehlern bei Personenbeurteilungen. Im Gegensatz dazu schenkte man erst in jüngerer Zeit jenen subjektiven Tendenzen bei der Beobachtung und Beurteilung von Menschen mehr Aufmerksamkeit, die aus der *Erwartungshaltung* resultieren. Bei dieser Forschungsrichtung ging es in der pädagogischen Psychologie primär um Erkenntnisse dar-

über, inwieweit sich die Erwartungen, die Lehrer in bestimmte Schüler setzen, erfüllen („self-fulfilling-prophecy", „Erwartungs-Effekt"). Der grundsätzliche Nachweis solcher Effekte konnte erbracht werden, z. B. in der viel beachteten und kritisierten „Pygmalion-Studie" von ROSENTHAL und JACOBSEN (vgl. ERLEMEIER & TISMER 1973). Bei diesen Experimenten wurde unter anderem beobachtet, daß Lehrer je nach positiver oder negativer Erwartung bei ihren Schülern mehr Positives oder Negatives wahrnehmen. Diese erhöhte Sensibilität für erwartungsgemäßes Schülerverhalten interessiert in unserem Zusammenhang. Sie ist allerdings keine Neuentdekkung der Psychologie. ZILLIG war schon 1928 aufgefallen, daß Lehrer in den Diktatheften guter Schüler eher Fehler übersehen als in den Heften schlechter Schüler.

Wenn nun gute Leistungen von Schülern mit einem ungünstigen „Leistungs-Halo" oder „sozialen Halo" weniger beachtet werden als bei guten Schülern, so werden sie auch weniger verstärkt. Das wiederum behindert den Lernprozeß. RÖHM (1973) berichtet über einen Einzelfall:

„Der Lehrer weiß aus den Akten, daß Ralf der nichteheliche Sohn einer berufstätigen Frau ist. Ralf fiel dem Lehrer bisher — abgesehen von einigen, wie ihm schien, ‚schrulligen Antworten' — kaum auf ... Nun kommt die Mutter zum ersten Mal zu ihm ... Der Lehrer lernt die Mutter, eine gebildete Frau, näher kennen. Er erfährt, daß man sich — anders als er dachte — viel um Ralf kümmert und daß der Vater der Mutter ein bekannter Hochschullehrer war ... Die veränderte Einstellung Ralf gegenüber aktivierte den bisher zurückhaltenden Jungen deutlich ... Antworten, die der Lehrer wie die Klasse früher als verschroben abtaten, erschienen dem Lehrer (und bald auch seinen Mitschülern) nun als kreativ und wert, daß man sich mit ihnen, auch wenn sie auf den ersten Blick unpassend schienen, beschäftigte" (S. 97).

Bei einem Schulexperiment, das MEICHENBAUM et al. (1969) durchführten, konnten bei den Lehrern ähnliche Erwartungsänderungen erzielt werden wie bei dem Lehrer von Ralf. Allein die Mitteilung eines Experten, bestimmte Schüler seien „Spätentwickler", bewirkte, daß die „umgestellten" Lehrer bei den aufgewerteten Schülern mehr gute Leistungen registrierten als vorher (nach HOFFMANN 1973, S. 111).

Ob der Lehrer die Vielzahl der subjektiven Urteilstendenzen, die hier nur ansatzweise diskutiert wurden, beim Umgang mit seinen Schülern in etwa kontrollieren und steuern kann, bedarf noch eingehender wissenschaftlicher Untersuchungen. Prinzipiell ist die Möglichkeit dazu solange anzunehmen, bis das Gegenteil erwiesen ist (woran ich nicht glauben kann). Entsprechende Anstrengungen sind schon deshalb lohnenswert, weil — erwiesenermaßen — im Fall einer gerechten Verhaltensbeurteilung sich der Schüler vom Lehrer besser verstanden fühlt. Daraus resultiert nicht nur eine größere

Aufgeschlossenheit seitens der Schüler, sondern im allgemeinen auch eine bereitwilligere Mitarbeit im Unterricht, was dem Lern- und Erziehungsprozeß vielfach zugute kommt.

5.1.6. Literaturverzeichnis

Bales, R. F.: Die Interaktionsanalyse: Ein Beobachtungsverfahren zur Untersuchung kleiner Gruppen. In: *König, R.*, Beobachtung und Experiment in der Sozialforschung. Köln 1972, 8. Aufl.

Bittmann, F.: Leistungsmotivation bei behinderten Kindern und ihre schulische Förderung. In: *Nickel, H.* u. *Langhorst, E.* 1973.

Bleidick, U.: Das sonderpädagogische Gutachten. Berlin 1972, 4. Aufl.

Cranach, M. v. u. *Frenz, H.-G.:* Systematische Beobachtung. In: Handbuch der Psychologie. 7. Bd.: Sozialpsychologie, 1. Halbband (hg. von *C. F. Graumann*), Göttingen 1969.

Cronbach, L. J.: Essentials of Psychological Testing. New York 1970, 3. Aufl.

Dieterich, R.: Einführung in die methodischen Grundlagen der Pädagogischen Psychologie. München 1972.

Donat, H.: Persönlichkeitsbeurteilung. Methoden und Probleme der Charaktererfassung im pädagogischen Bereich. München 1970, 2. Aufl.

Erlemeier, N. u. *Tismer, K.-G.:* Einstellungen und Erwartungen bei Lehrern und ihre Auswirkungen auf die Beurteilung und das Verhalten von Schülern. In: *Nickel, H.* u. *Langhorst, E.* 1973.

Fenner, H.-J.: Verfahren und Ergebnisse zur Objektivierung des Lehrerverhaltens. In: *Nickel, H.* u. *Langhorst, E.* 1973.

Gottschaldt, K.: Der Aufbau des kindlichen Handelns. Vergleichende Untersuchungen an gesunden und psychisch abnormen Kindern. Leipzig 1954, 2. Aufl.

Graumann, C. F.: Eigenschaften als Problem der Persönlichkeitsforschung. In: Handbuch der Psychologie. 4. Bd.: Persönlichkeitsforschung und Persönlichkeitstheorie (hg. von *Lersch, Ph.* u. *Thomae, H.*), Göttingen 1960.

Graumann, C. F.: Grundzüge der Verhaltensbeobachtung. In: *Meyer, E.*, Fernsehen in der Lehrerbildung. München 1966. — Wiederabdruck in: Pädagogische Psychologie, 1. Teil: Entwicklung und Sozialisation (hg. von *Graumann, C.-F.* u. *Heckhausen, H.*), Frankfurt a. M. 1973 (Fischer Taschenbuch).

Guilford, J. P.: Persönlichkeit. Logik, Methodik und Ergebnisse ihrer quantitativen Erforschung. (Übertragen von *Kottenhoff, H.* u. *Agrell, U.*), Weinheim 1971, 5. Aufl.

Hasemann, K.: Verhaltensbeobachtung und Verhaltensbeurteilung in der psychologischen Diagnostik. In: Handbuch der Psychologie. 6. Bd.: Psychologische Diagnostik (hg. von *Heiß, R.*), Göttingen 1964 (Auch als Sonderdruck erschienen).

Heller, K.: Intelligenzmessung. Zur Theorie und Praxis der Begabungsdiagnostik in Schule und Sonderpädagogik. Villingen 1973.

Heller, K. et al.: Planung und Auswertung empirischer Untersuchungen. Stuttgart 1974.

Hofer, M.: Zur impliziten Persönlichkeitstheorie von Lehrern. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 1970, 2, 197—209.

Hoffmann, Monika: Beobachten und Protokollieren von Verhalten. In: *Belschner, W.* et al.: Verhaltenstherapie in Erziehung und Unterricht. Stuttgart 1973.

Höhn, Elfriede: Der schlechte Schüler. Sozialpsychologische Untersuchungen über das Bild des Schulversagers. München 1973, 5. Aufl.

Nickel, H.: Entwicklungspsychologie des Kindes- und Jugendalters. Bd. I. Bern 1972.

Nickel, H. u. *Langhorst, E.* (Hg.): Brennpunkte der pädagogischen Psychologie. Bern/Stuttgart 1973.

Röhm, H.: Zur Problematik der Beurteilung von Schülern. Psychologie in Erziehung und Unterricht, 1973, *20*, 94—102.

Roth, H.: Pädagogische Psychologie des Lehrens und Lernens. Hannover 1969, 11. Aufl.

Schulz, W. et al.: Verhalten im Unterricht. Seine Erfassung durch Beobachtungsverfahren. (Deutsche Bearbeitung von: *Medley, D. M.* a. *Mitzel, H. E.:* Measuring Classroom Behavior by Systematic Observation. In: Handbook of Research on Teaching, Hrsg. *Gage, N. L.)* — Deutsch hg. von *Ingenkamp, K.*: Handbuch der Unterrichtsforschung, Teil I. Weinheim 1971, 2. Aufl.

Tent, L.: Schätzverfahren in der Unterrichtsforschung. (Deutsche Bearbeitung von: *Remmers, H. H.:* Rating Methods in Research on Teaching. In: Handbook of Research on Teaching, Hrsg. *N. L. Gage)* — Deutsch hg. von *Ingenkamp, K.:* Handbuch der Unterrichtsforschung, Teil I. Weinheim 1971, 2. Aufl.

Thomae, H.: Beobachtung und Beurteilung von Kindern und Jugendlichen. Basel 1960, 3. Aufl.; 1971, 10. Aufl.

Ulich, D. u. *Mertens, W.:* Urteile über Schüler. Zur Sozialpsychologie pädagogischer Diagnostik. Weinheim 1973.

Wasna, Maria: Die Entwicklung der Leistungsmotivation. Zielsetzungen normaler und debiler Kinder bei einer Turmbauaufgabe. München 1970.

Zillig, Maria: Einstellung und Aussage. Zeitschrift für Psychologie, 1928, *106*, 58 bis 106.

5.2. Leistungsbeurteilung durch Notengebung

Walter Fingerhut und Hans-Peter Langfeldt

5.2.1. Schulnoten als Urteile

Schulnoten sind Meßzahlen für erbrachte Leistungen der Schüler, die der Lehrer nach seinen Erfahrungen und Einschätzungen auf der Notenskala einstuft. Noten kommen also aufgrund eines Urteilsprozesses des Lehrers zustande und sind mit all den Mängeln behaftet, die man bei Urteilsprozessen nachgewiesen hat.

Zunächst muß man feststellen, daß der Urteilsprozeß sich in einer sozialen Situation vollzieht. Man kann vom Lehrer nicht erwarten, daß er sich von den momentanen situativen Einflüssen freimachen kann, um dann mit der Präzision eines physikalischen Meßinstrumentes die Leistung eines Schülers zu bewerten. Der soziale Bezug schlägt sich also in den Zensuren nieder (vgl. RÖHM 1973).

Nach allen bekannten Ergebnissen schließt die Leistungsbeurteilung eines Schülers außerdem die Beurteilung seiner Persönlichkeit mit ein (vgl. HÖHN 1967, HOFER 1969, ULICH & MERTENS 1973). So ist es nicht verwunderlich, wenn extreme Kritiker der Beurteilungspraxis davon sprechen, daß die Noten mehr über den Lehrer aussagen als über den Schüler.

Noten als Urteile unterliegen den bekannten Beurteilungsfehlern. Diese Fehler sind im Rahmen sozialpsychologischer Lehrbücher häufig beschrieben worden, so daß es an dieser Stelle genügt, nur die wichtigsten kurz anzudeuten (in Anlehnung an HASEMANN 1964, S. 826 ff.):

Der Fehler des milden Urteils („generosity error")

Damit ist allgemein die Neigung eines Beurteilers gemeint, negative Urteile zu vermeiden und möglichst positive Beurteilungen abzugeben. Dieser Neigung wird besonders dann nachgegeben, wenn Personen beurteilt werden, die dem Beurteiler sympathisch sind.

Der Fehler der zentralen Tendenz („error of central tendency")

Darunter wird die Neigung verstanden, Extremurteile zu vermeiden und eher solche Urteile abzugeben, die weder positiv noch negativ akzentuiert sind. Dieser Fehler tritt z. B. auf, wenn der Beurteiler glaubt, daß ihm der Beurteilte noch zu wenig bekannt ist.

Der Halo-Effekt

Damit ist die Tendenz gemeint, die Beurteilung einzelner Merkmale unberechtigterweise auf andere Merkmale des Beurteilten auszudehnen („Wer lügt, der stiehlt auch").

Der logische Fehler („logical error")

Damit bezeichnet man die Neigung eines Beurteilers, diejenigen Merkmale gleichwertig zu beurteilen, die er für logisch zusammengehörig hält. Wenn dieser Zusammenhang empirisch aber nicht besteht, kommt der Beurteiler aufgrund solcher Überlegungen zu einem Fehlurteil.

Eine systematische Darstellung der Urteilsfehler findet sich bei SIXTL (1967, S. 259 ff.). Eine Übersicht über die Fehler speziell bei Lehrerurteilen gibt KLEITER (1973).

Aus den geschilderten Einflüssen auf die Urteilsbildung folgt also, daß Noten als Urteile keine exakten Abbildungen von Schulleistungen sein können. Dies kann den Lehrern jedoch nicht zum Vorwurf gemacht werden. In jeder Unterrichtsstunde wird von ihnen eine Vielzahl schneller und unabhängiger Entscheidungen und Beurteilungen verlangt, die sie nur bewältigen können, wenn sie bestimmte Urteils- und Verhaltensstrategien entwikkeln. Diese notwendige Bildung von Stereotypen verhindert aber exakte Urteile. Ein guter Beurteiler sollte „in bezug auf Menschen kognitiv komplex sein", „keine falsche implizierte Persönlichkeitstheorie haben", „für die einzuschätzenden Dimensionen empfänglich sein" und er sollte „sensitiv sein für kleinste verbale und nichtverbale Schlüsselreize" (ARGYLE 1972, S. 161 f.). Diesen Anforderungen an einen guten Beurteiler scheint niemand gewachsen zu sein, denn auch unter wesentlich günstigeren Bedingungen sind Urteile über Personen ungenau und fehlerhaft (vgl. das Sammelreferat von MERZ 1963).

Demnach besteht wenig Aussicht, die Beurteilungen von Lehrern zu verbessern. Verbessert werden sollten dagegen die Möglichkeiten zur objektiven Leistungserfassung, damit die Lehrer nicht länger fast ausschließlich auf ihre subjektive Beurteilung angewiesen sind. Dazu gehört im weiteren auch, daß Schulnoten nicht länger solche folgenschwere Konsequenzen haben sollten wie bisher.

5.2.2. Schulnoten als Meßwerte

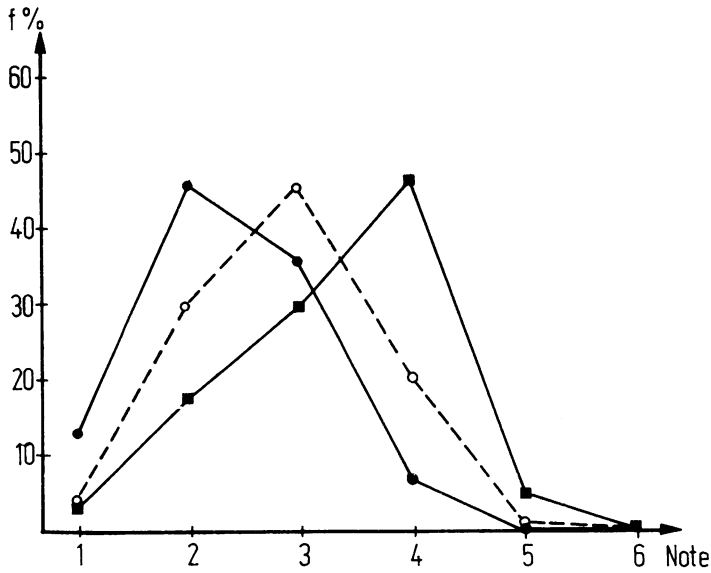
Schulnoten als Meßwerte des Merkmals „Schulleistung" werden in unserem Schulsystem auf einer sechsstufigen Skala abgebildet. Diese Skala ist eine Rangskala, wobei „1" (sehr gut) die beste und „6" (ungenügend) die schlechteste Leistung charakterisieren. Da Noten nur Rang- oder Ordinalskalenniveau besitzen, ist die Berechnung der Verteilungsparameter Mittelwert und Streuung eigentlich unzulässig. Bei hinreichend großen Stichproben wird jedoch angenommen, daß die Notenskala sich dem Intervallskalenniveau annähert (LIENERT & HOPP 1964, S. 195; TENT 1969, S. 55 f.).

Um Noten statistisch verarbeiten zu können, wird das Verfahren der Flächentransformation vorgeschlagen (Orlik 1961), wodurch z. B. Noten verschiedener Lehrer miteinander verglichen werden können.

5.2.2.1. Verteilungsform von Schulnoten

In ihren beiden Veröffentlichungen (Lienert & Hopp 1964; Hopp & Lienert 1965) berichten die Autoren über die Verteilungsform von Gymnasialzensuren. Da die Daten an einer großen Stichprobe (N um 1 000) gewonnen wurden, kann geschlossen werden, daß die Verteilungsformen *fachtypisch* sind. Es bilden sich drei Gruppen von Fächern heraus: Fächer mit milder (musische Fächer und Religion), mit mittlerer (Nebenfächer) und mit strenger Beurteilung (Hauptfächer). Abbildung 1 zeigt die relativen Häufigkeiten der Noten für Religion, Erdkunde und Latein als Beispiele für die Verteilungsformen bei mild, mittel und streng benoteten Fächern. Die Verteilungen sind entsprechend linksschief, nahezu symmetrisch oder rechtsschief.

Abb. 1: Relative Häufigkeitsverteilung der Noten in den Fächern -●- Religion, -○- Erdkunde und -■- Latein an der Unterstufe des Gymnasiums.
(nach Daten von HOPP & LIENERT 1965, 32, Tab. 3)



Die Rangreihe hinsichtlich der Strenge der Gymnasialzensuren, die üblicherweise mit einem musischen Fach beginnt und mit Latein (oder einer anderen Fremdsprache) endet, scheint relativ stabil zu sein, wie einige weitere Untersuchungen belegen (z. B. HÖGER 1964; WEINGARDT 1964; KNOCHE 1969; MÜLLER-FOHRBRODT & DANN 1971). Ähnlich strukturierte Rangordnungen der Fächer fand WEISS (1971) bei nahezu 4 000 vierzehnjährigen Schülern verschiedener österreichischer Schultypen. Auch im Primarbereich zeigt sich eine deutliche Rangordnung der Fächer. Die Daten von FERDINAND & KIWITZ (1971) können dies verdeutlichen:

Tabelle 1

Durchschnittszensuren mit Streuungswerten von 1898 Schülern und Schülerinnen des vierten Schuljahres (aus FERDINAND & KIWITZ 1971, S. 182, Tab. II)

Fach	Durchschnitts-zensur	Streuung
Führung	1,85	o,58
Hausl.Fl.	2,64	0,93
Beteil. am Unt.	2,85	0,92
Lesen	2,74	0,89
Mündl. Ausdruck	2.95	0,87
Heimatk.	2,99	0,89
Aufsatz	3,18	0,90
Rechnen	3,19	1,00
Rechtschr.	3,41	1,22

Die Tabelle zeigt ein weiteres durchgängiges Ergebnis: Je numerisch kleiner die Durchschnittszensur, desto geringer ist die Streuung. Dies bedeutet, daß eine milde Beurteilung eine undifferenzierte Beurteilung mit einschließt. Die Untersuchungsbefunde von FUNKE (1972) und von ZIMMERMANN (1968) an lernbehinderten Sonderschülern bestätigen diesen Sachverhalt. Sowohl bei Sonderschülern als auch bei Grundschülern wird „Betragen" am mildesten und undifferenziertesten und „Deutsch schriftlich" am strengsten und differenziertesten beurteilt.

Schulnoten verteilen sich auf der Notenskala also nicht normal. Vielmehr sind fachtypisch schiefe Verteilungen festzustellen. Diese Tatsache führte häufig zu der Forderung, Lehrer sollten zur Verbesserung ihrer Notengebung dafür sorgen, daß ihre Noten pro Klasse normalverteilt seien. GÖLLER (1966, S. 24 ff.) beispielsweise gibt konkrete Anleitungen, wie die Normalverteilung der Noten bei jeder Klassenarbeit hergestellt werden kann. FERDINAND & KIWITZ (1971) schlagen dagegen eine Modifikation vor, bei der — im Vergleich zur Normalverteilung — die Extremwerte („sehr gut"

bzw. „nicht genügend") an Häufigkeit zu- und die mittleren Werte abnehmen.

Wir sind jedoch der Meinung, daß die Forderung nach Normalverteilung (oder anderen fest vorgegebenen Verteilungen) der Schulnoten pädagogisch nicht zu verantworten ist. Normalverteilungen können — falls überhaupt — nur in großen Stichproben Gültigkeit besitzen. Das Anpassen der Noten an die Normalverteilung in (statistisch gesehen) kleinen Schulklassen ist nicht zu rechtfertigen. Das Festhalten an dem Normalverteilungskonzept bei jeder Klassenarbeit hat z. B. zur Folge, daß der schlechteste Schüler immer die Note „ungenügend" erhält, und zwar unabhängig von einer etwaigen *objektiven* Leistungssteigerung, solange die übrigen Schüler der Klasse sich auch verbessern. Im Extremfall erhält ein objektiv guter Schüler immer eine schlechte Note, wenn er das Pech hat, in einer Klasse von noch besseren Schülern zu sein.

Bei Anwendung der Normalverteilung wird die Note eines Schülers eher von den Leistungen seiner Mitschüler als von seiner eigenen Leistung bestimmt. Wenn er selbst eine bessere Note erhalten will, muß er dafür sorgen, daß seine Mitschüler nicht im gleichen Maße eine Leistungssteigerung vollbringen wie er selbst. Dies könnte zu übersteigertem Wettbewerb innerhalb der Klasse führen. Wahrscheinlicher ist aber, daß ein Schüler resigniert, wenn er trotz Anstrengung weiterhin schlechte Noten erhält. Diese fehlende Motivation kann dann den weiteren schulischen Werdegang entscheidend beeinflussen.

Das Abweichen der Schulnoten von einer Normalverteilung ist allenfalls unter statistischen Gesichtspunkten ärgerlich. Pädagogisch gesehen ist die Forderung nach Normalverteilung der Noten unverantwortlich. Sie zwingt den Lehrer unter Umständen, die tatsächlichen Verhältnisse in seiner Klasse zu mißachten.

5.2.2.2. Objektivität von Schulnoten

Die Forderung nach Objektivität von Schulnoten bedeutet, daß die gleiche schulische Leistung von verschiedenen Lehrern gleich beurteilt wird. Eine Reihe empirischer Untersuchungen konnte zeigen, daß Schulnoten diese Forderung nicht erfüllen. Hauptgegenstand der entsprechenden Forschung war bezeichnenderweise die Aufsatznote. Ein kurzer Überblick findet sich in Ingenkamp (Hrsg. 1971, besonders Kapitel III). Da die Problematik der Aufsatzbeurteilung außerdem Gegenstand zweier Kapitel dieses Buches ist (vgl. Kap. 5.3 und 5.4.) wird auf eine detaillierte Darstellung an dieser Stelle verzichtet. Es kann jedoch als durchgängiges und gesichertes Ergebnis festgehalten werden: Aufsatzzensuren sind nicht objektiv, d. h. gleiche Aufsätze werden von verschiedenen Beurteilern unterschiedlich benotet.

Mangelhafte Objektivität ist nicht nur bei der Beurteilung von Aufsätzen festgestellt worden, sondern auch bei Noten in solchen Fächern, in denen eindeutige Kriterien für richtige und falsche Lösungen vorliegen, wie etwa in Rechnen und Rechtschreiben.

Im Rahmen seiner Untersuchung zur Objektivität von Aufsätzen ließ WEISS (1965) auch Beurteilungen hinsichtlich der Kategorie „Rechtschreibung" abgeben. Die auftretenden Divergenzen zeigen, daß selbst die Beurteilung der Rechtschreibung — wenn auch in geringerem Ausmaße — nicht objektiv vorgenommen wurde.

Die Benotung von Klassenarbeiten ist dadurch zu beeinflussen, daß man den Beurteilern (Lehrern) unterschiedliche Informationen über den entsprechenden Schüler gibt. Dies gilt sowohl für Aufsätze (FERDINAND 1971) als auch für Rechenarbeiten (WEISS 1966).

CARTER (1952) konnte zeigen, daß Rechennoten abhängig vom Geschlecht des beurteilenden Lehrers und des beurteilten Schülers unterschiedlich ausfallen können. Mädchen erhielten bessere Rechennoten als Jungen, obwohl sich die Leistungen in einem objektiven Rechentest nicht unterschieden. Lehrerinnen erteilten bessere Noten als Lehrer.

Erziehungsschwierige Kinder (definiert durch schlechte Kopfnoten) erhalten bei gleichen Intelligenztestleistungen im Durchschnitt im gesamten Zeugnis schlechtere Zensuren als nicht erziehungsschwierige Kinder (FERDINAND 1962). Dies stimmt mit den Ergebnissen einer Untersuchung von HADLEY (1971) überein, wonach beliebte Schüler im Vergleich zu objektiven Schultests überschätzt und unbeliebte Schüler unterschätzt sowie entsprechend benotet werden.

Objektivität von Schulnoten müßte auch bedeuten, daß gleiche Leistungen in verschiedenen Klassen gleich benotet werden. INGENKAMP (1965 und 1971) konnte zeigen, daß dies für Rechennoten nicht zutrifft. Indirekt läßt sich dies auch für andere Schulleistungen schließen: FIPPINGER (1967) stellte bei Schülern in voll gegliederten Schulen durchschnittlich bessere objektive Schulleistungen fest als bei Schülern in wenig gegliederten Schulen. Man kann aber nicht annehmen, daß sie auch durchschnittlich bessere Zensuren erhalten, denn Lehrer orientieren sich im allgemeinen an klasseninternen Maßstäben.

Objektivität von Schulnoten ist also nicht gegeben. Dies ist vor allem in der Verquickung mindestens zweier Funktionen der Noten begründet. Einerseits soll eine Note die Leistung eines Schülers erfassen oder messen, andererseits ist mit der Note gleichzeitig eine Bewertung oder Beurteilung eben dieser Leistung verbunden. Es kann kein Zweifel daran gelassen werden, daß die Schulleistung *objektiv erfaßt und registriert* werden muß. Dies kann jedoch keinesfalls heißen, daß objektiv gleiche Leistungen nach pädagogischen Intentionen auch gleich *interpretiert und bewertet* werden müs-

sen. Eine Objektivierung von Schulnoten wäre allenfalls zu erreichen, wenn man auf die pädagogische Funktion (vgl. FLITNER 1966; WEISS 1969), welche die Noten sowieso nicht ausreichend erfüllen können, verzichten würde. Nur durch eine strikte Trennung von *Leistungsmessung* und *Leistungsbewertung* ließen sich einige Schwierigkeiten hinsichtlich der Objektivität verringern.

5.2.2.3. Reliabilität von Schulnoten

Die derzeitigen Versetzungsordnungen und Selektionspraktiken setzen die Konstanz der Schulleistung und damit der Noten voraus. Von einem guten Schüler wird erwartet, daß er ein guter bleibt, ebenso von einem schlechten Schüler, daß er ein schlechter bleibt. Entsprechend diesen Erwartungen werden sie versetzt bzw. nicht versetzt oder es wird ihnen der Besuch einer weiterführenden Schule empfohlen bzw. davon abgeraten.

Die Korrelation zwischen aufeinanderfolgenden Noten oder Zeugnissen ist ein Maß für die Konstanz der Noten. Diese Korrelation entspricht der Berechnung einer Retest-Reliabilität. Unabhängig davon, ob es sinnvoll ist, eine hohe Wiederholungs-Reliabilität der Noten zu fordern, gibt ein solcher Koeffizient doch Auskunft über die praktizierte Notengebung. TENT (1965, S. 588) teilt für die Grundschule überraschend hohe Koeffizienten mit (0,76 bis 0,86). WEINGARDT (1964) fand für die gesamte Gymnasialzeit mittlere Retest-Reliabilitätskoeffizienten in der Größenordnung von 0,3 bis 0,5. Aufgrund verschiedener Untersuchungen (unter anderem RANK 1962; HÖGER 1964; SCHULTZE 1964; TENT 1965 und 1969) kann allgemein gesagt werden: Bei zeitlich kurzem Abstand (etwa einem Jahr) erhält man relativ hohe Retest-Koeffizienten. Mit zunehmendem Zeitabstand sinken sie in mittlere bis niedrige Bereiche ab. Interessant ist in diesem Zusammenhang, daß die Koeffizienten über einen längeren Beobachtungszeitraum in ihrer Höhe periodisch schwanken (TENT 1969, S. 72). Die Korrelationen zwischen dem ersten Halbjahreszeugnis und dem letzten des Vorschuljahres liegen im allgemeinen niedriger als die Korrelationen zwischen den beiden Zeugnissen eines Schuljahres. Diese Schwankungen könnten mit dem Lehrerwechsel zu Beginn eines neuen Schuljahres erklärt werden.

ASCHERSLEBEN (1971) korrelierte die Noten je vier aufeinander folgender Diktate und Rechenklassenarbeiten miteinander. Die beiden so erhaltenen durchschnittlichen Koeffizienten werden von ihm als Parallel-Retest-Reliabilität interpretiert. Die erhaltenen Koeffizienten liegen bei den meisten von ihm untersuchten Klassen unter 0,6, was er als Minimalforderung ansieht. Diktatnoten erwiesen sich in diesem Sinne als reliabler als die Noten der Rechenarbeiten.

Es ist selbstverständlich, daß Noten die Schulleistung zuverlässig erfassen sollten. Die Reliabilitätskonzepte der klassischen Testtheorie werden dem Problem nicht gerecht. Paralleltest-Reliabilität oder Konsistenz-Reliabilität sind bei herkömmlichen Klassenarbeiten oder Noten für mündliche Leistungen überhaupt nicht praktikabel. Außerdem ist die Retest-Reliabilität eines Meßinstrumentes oder Meßwertes nur dann ein sinnvolles Maß für die Güte des Meßinstrumentes bzw. Meßwertes, wenn der gemessene Sachverhalt konstant bleibt. Die Forderung nach hoher Retest-Reliabilität von Noten setzt also voraus, daß die schulischen Leistungen der Schüler sich nicht ändern. Dies widerspricht aber offensichtlich den Intentionen der Schule.

Unter diesem Gesichtspunkt erscheinen hohe, empirisch ermittelte Retest-Koeffizienten überraschend und bedenklich. Das klassische Reliabilitätskonzept ist zwar geeignet, die instrumentellen Mängel von Noten zu beschreiben. Mit dieser Feststellung kann die Forderung nach absoluter Anpassung der Noten an dieses Konzept jedoch *nicht* verbunden werden.

5.2.2.4. Validität von Schulnoten

Als drittes testtheoretisches Gütekriterium soll die Validität von Noten diskutiert werden. Die Frage danach, *was* Schulnoten eigentlich „messen" bzw. welcher Sachverhalt der Angabe einer Maßzahl zwischen eins und sechs zugrunde liegt, scheint zunächst banal zu sein: Schulnoten sind Maßzahlen für die vom Lehrer vorgenommene Einschätzung der schulischen Leistungen, wie sie in Klassenarbeiten, Prüfungen oder während des Unterrichtes wahrgenommen werden. Von daher besitzen Schulnoten zunächst logische Validität bezüglich des Merkmals „Schulleistung".

Lehrer unterscheiden sich aber hinsichtlich der Aspekte, die sie in ihr subjektives Konzept von Schulleistung einfließen lassen und wie sie die verschiedenen Aspekte gewichten. Mit anderen Worten: Nimmt man als Kriterium für die Validität von Noten das subjektive Konzept „Schulleistung" des benotenden Lehrers, so muß man mit großen interindividuellen Differenzen rechnen. Daß damit den Lehrern kein Vorwurf gemacht werden kann, läßt sich leicht einsehen, wenn man berücksichtigt, welche Funktionen Noten erfüllen sollen. Zielinski (1973) gibt einen Kanon von zehn Funktionen an:

1. Die Rückmeldefunktion für den Lehrer
2. Die Rückmeldefunktion für den Schüler
3. Die Berichtsfunktion
4. Die Anreizfunktion
5. Die Disziplinierungsfunktion
6. Die Sozialisierungsfunktion
7. Die Klassifikationsfunktion

8. Die Selektionsfunktion
9. Die Zuteilungsfunktion
10. Die Chancenausgleichsfunktion.

Ähnliche, wenn auch weniger differenzierte Funktionsangaben finden sich auch bei DOHSE (1963) und FLITNER (1966). Zusätzlich wird die pädagogische Funktion genannt (WEISS 1969, S. 203 f.), nach der die Note ein Erziehungsmittel sein soll, den guten Schüler zu belohnen, den schlechten anzuspornen oder den faulen zu bestrafen.

Es ist leicht einzusehen, daß Lehrer überfordert wären, wollten sie alle diese Aspekte in die Notengebung einfließen lassen. Außerdem wird die Aussagekraft der Noten eingeschränkt, wenn zu dem eingangs diskutierten Messungsaspekt zusätzlich Bewertungsaspekte in die Zensuren eingehen. „Eine einzige Ziffer soll Feststellungs- und Bewertungsakt adäquat ausdrücken. Das ist eine unmögliche Forderung" (INGENKAMP 1968, S. 412).

Es kommt hinzu, daß auch die Aufgaben der Schule, wie sie in den Bildungsplänen festgelegt sind, meistens so vage formuliert sind, daß die Lehrer auch von daher keine Aufklärung darüber erhalten, welche konkreten Kriterien der Leistungsbeurteilung zugrunde gelegt werden sollen. Das Problem der validen Leistungsbeurteilung ist also auch eng mit den Problemen der Curriculumforschung verbunden (vgl. Kap. 4.1. in diesem Buch).

Damit wird deutlich, wie schwierig es ist, Aussagen über die Validität von Schulnoten zu machen, solange man so wenig über das Kriterium „Schulleistung" weißt. Auf neuere Ansätze zur Lösung dieses Problems wurde in Kapitel 2.1. eingegangen. Die dort referierten Befunde sind als Beitrag zur Bestimmung der inhaltlichen Validität von Schulnoten zu verstehen.

Häufig bestimmt man die inhaltliche Validität von Noten durch die Korrelation mit objektiven, standardisierten Schultests, die den Kriterien der „logischen" bzw. „psychologischen" und der „kurrikularen Validität (SÜLLWOLD 1964, S. 368 ff.) genügen. SÜLLWOLD nennt für eine Reihe amerikanischer Schultests Koeffizienten in der Größenordnung um 0,7. Dies entspricht den Werten bei deutschen Schultests (Beltz Verlag 1968).

Je nach der Betrachtungsweise kann jedoch ein beobachteter Zusammenhang zwischen Noten und Schultests als Validitätskoeffizient sowohl der Tests als auch der Noten interpretiert werden. Hier liegt noch eine grundsätzliche Schwierigkeit des Vorgehens. Bei einer solchen Vorgehensweise hängt die inhaltliche Validität von Noten von der Validität der Schultests ab (vgl. dazu den Beitrag von LANGFELDT, S. 114 ff.).

Schulnoten sollen nicht nur aktuelle schulische Leistungen kennzeichnen, sondern auch Prognosen zukünftiger Leistungen ermöglichen. Dazu müßten die Lehrer aber Informationen mit einbeziehen, die über die weitere schulische Entwicklung der Schüler Auskunft geben. Da solche allgemeingültigen Schülermerkmale, die eine Extrapolation von momentaner auf zukünfti-

ge Schulleistung gestatten würden, nicht bekannt sind, müssen notwendigerweise subjektive Momente des Lehrers einfließen. Es verwundert daher nicht, daß die Vorhersage-Validität von Noten recht niedrig ausfällt. Die Korrelationen zwischen den vor dem Übergang auf das Gymnasium erteilten Grundschulnoten und dem Gymnasialerfolg nach fünf Jahren liegen in der Größenordnung von r = 0,3. Das gleiche gilt für die Korrelationen zwischen Aufnahmeprüfung und Gymnasialerfolg (SCHULTZE 1964). Demgegenüber korrelieren Intelligenztestwerte mit dem späteren Schulerfolg auf dem Gymnasium zwischen r = 0,4 und r = 0,8 (HITPASS 1963; GEBAUER 1965; BURGER 1967). Diese empirisch gefundene prognostische Überlegenheit von Intelligenztests gegenüber Grundschulnoten ist aufgrund der besseren testtheoretischen Gütekriterien von Intelligenztests zu erwarten. Allerdings wird die Vorhersage-Validität von Noten durch Aufnahmeprüfungen stärker eingeschränkt als die von Intelligenztestwerten, so daß die empirisch gefundenen Differenzen überhöht sein können (JANSSEN 1972).

Prinzipiell ist fraglich, inwieweit uns mit einer besseren prognostischen Validität von Schulnoten wirklich geholfen wäre. Eine hohe Vorhersage-Validität beinhaltet ja die Konstanz des Merkmals Schulleistung. Man würde mit der Forderung nach hoher Validität, die ja im Zusammenhang mit der Übertrittsauslese immer noch eine Rolle spielt, dem Anspruch der Schule auf Förderung der Schüler entgegenwirken, da eine Verbesserung der Schulleistungen, vor allem bei den schwächeren Schülern, zu einem Absinken der Vorhersage-Validität führen müßte.

Diese Problematik verdeutlicht die Untersuchung von STEINKAMP (1971). In einer Umfrage sahen sich nur fünf von 69 Lehrern des zweiten Schuljahres außerstande, potentielle Oberschüler ihrer Klasse zu benennen. Anders ausgedrückt: 64 von 69 Klassenlehrern „wußten" schon im zweiten Schuljahr, welche ihrer Schüler für das Gymnasium geeignet waren. In diesem Zusammenhang sei an den „Pygmalion-Effekt" erinnert (ROSENTHAL & JACOBSON 1971), der besagt, daß Schüler objektiv bessere Leistungen erbringen, wenn der Lehrer ein positives Vorurteil über diese Schüler hat. Das ist offensichtlich darauf zurückzuführen, daß Lehrer sich gegenüber solchen Schülern so verhalten, daß ihre Erwartungen bestätigt werden müssen (BROPHY & GOOD 1970; RUBOVITZ & MAEHR 1971).

5.2.3. Schulnoten als Variablen

5.2.3.1. Methodische Fragen

Wie schon dargestellt, sind Schulnoten Maßzahlen auf einer Rangskala. Falls die Noten im Zusammenhang mit anderen Variablen analysiert werden sollen, ist eine statistische Transformation notwendig, um mindestens Intervallskalenniveau zu erreichen (ORLIK 1961).

In einschlägigen Untersuchungen gibt man sich häufig nicht mit den Noten zufrieden, so wie sie vom Lehrer im Zeugnis erteilt werden. In Anbetracht der instrumentellen Schwächen der Noten wird versucht, weitere Maßzahlen zu gewinnen. So entwickelte beispielsweise WEISS (1964 a) eine „Schulleistungszahl" als Summe aus gewichteten Einzelnoten. Die Gewichte wurden dabei durch Expertenrating aufgestellt und auf Brauchbarkeit überprüft. TODT (1966) definiert eine „allgemeine Schulleistung" durch das arithmetische Mittel der Einzelnoten eines Zeugnisses. Vereinzelt wurden zu den Schulnoten noch zusätzliche Lehrerurteile oder Klassenarbeiten hinzugezogen (so z. B. bei FIPPINGER 1966). SCHMITZ (1964) legt einen „allgemeinen Schulleistungsstand" fest, der durch den mittleren z-Wert aus einem Diktat und zwei Rechenarbeiten definiert wird.

In den genannten Untersuchungen ist im allgemeinen von Zusammenhängen der Schul*leistung* mit bestimmten Variablen die Rede. Es ist jedoch zu beachten, daß es sich im strengen Sinne um Zusammenhänge der Schul*noten* handelt.

5.2.3.2. Schulnoten und Intelligenz

Seit der Einführung der Intelligenztests ist die Beziehung zwischen Schulnoten und Intelligenztestleistungen häufig Gegenstand empirischer Forschung gewesen. FIPPINGER (1966) gibt in diesem Zusammenhang einen geschichtlichen Überblick über die Ergebnisse solcher Untersuchungen. Er zeigt, daß die Korrelationen zwischen Schulnoten und Intelligenztestleistungen im Laufe der Zeit in ihrer Höhe abgenommen haben. So wurden in der Zeit vor 1930 Koeffizienten um 0,8 gefunden. Inzwischen sind die Koeffizienten in mittlere und niedrige Bereiche abgesunken.

Es ist selbstverständlich, daß der Zusammenhang zwischen Noten und Intelligenztestergebnissen (IQ) u. a. vom zugrunde liegenden Intelligenzkonzept des verwendeten Tests abhängt. In diesem Sinne ist es nur folgerichtig, wenn LÖSCHENKOHL (1973) nach einer ausführlichen Analyse einschlägiger Arbeiten feststellt, daß von einem allgemeinen Zusammenhang zwischen Schulleistung (erfaßt durch Noten) und Intelligenz nicht gesprochen werden kann. Vielmehr ist zu berücksichtigen, welcher Test in welcher Situation an welcher Stichprobe durchgeführt wurde.

Die Korrelationen zwischen Schulnoten und IQ schwanken in verschiedenen neueren Untersuchungen (nach 1960) zwischen 0,2 und 0,6. Koeffizienten dieser Größenordnung werden mitgeteilt von TODT (1966) bei Unterprimanern, HÖGER (1964) bei Gymnasiasten der Klassen sechs bis neun, SEITZ & LÖSER (1969) bei 17jährigen Gymnasialschülern, SIMONS (1969) bei Sextanern, SEITZ (1971 und 1970) bei Schülern der sechsten bzw. dritten Volksschulklasse und von SCHWARZ (1967) bei Erstkläßlern. Durch die der-

zeitige Auslesepraxis zu den weiterführenden Schulen sind Erhebungen im vierten Grundschuljahr von besonderer Bedeutung. Wie die Arbeiten von KOHL (1964), SCHMITZ (1964), TENT (1965 und 1969), FIPPINGER (1966) und LANGFELDT & FINGERHUT (1972) zeigen, bewegen sich auch hier die Koeffizienten in den genannten Grenzen. Bei der Bewertung solcher Befunde ist allerdings zu berücksichtigen, daß sie häufig an Stichproben gewonnen wurden, in denen Jungen und Mädchen vertreten waren. Die Ergebnisse von AMELANG & VAGT (1970) deuten jedoch an, daß bei Mädchen die entsprechenden Korrelationskoeffizienten höher ausfallen als bei Jungen.

Obwohl die Untersuchungen an unterschiedlichen Stichproben mit unterschiedlichen Verfahren durchgeführt wurden, ergänzen sie sich insofern, als man feststellen kann: Unabhängig vom Schultyp und unabhängig vom verwendeten Testverfahren ist der Zusammenhang zwischen Schulnoten und Intelligenztestleistungen gering bis allenfalls mittelmäßig. Die gemeinsame Varianz, ausgedrückt durch den Determinationskoeffizienten r^2, beläuft sich auf 5 % bis 40 %. Eine ausführlichere Darstellung der aufgezeigten Zusammenhänge findet sich in dem Beitrag von GAEDIKE auf S. 47 ff.

5.2.3.3. Schulnoten und soziale Herkunft

Die soziale Auslese innerhalb unserers Schulsystems ist häufig dokumentiert worden. Da diese Auslese hauptsächlich aufgrund der Zeugnisse geschieht, muß ein bedeutsamer Zusammenhang zwischen sozialer Schicht und Schulnoten angenommen werden.

Eine empirische Erhebung von PETRAT (1964) zeigt deutlich die ständische Gliederung unseres Schulsystems. Mit zunehmendem Klassenjahrgang nimmt im Gymnasium der Anteil der Kinder mit höherem Sozialstatus zu, in der Realschule der Anteil der Kinder mit mittlerem und in der Volksschule der Anteil der Kinder mit niedrigerem Sozialstatus. Dementsprechend konnte SCHULTZE (1964) an der Zusammensetzung von sechsten Gymnasialklassen zeigen, daß die Anteile der Schüler verschiedener Sozialschichten sich nahezu reziprok zu den Anteilen in der Gesamtbevölkerung verhalten. Nach einer Längsschnittuntersuchung, in deren Verlauf die schulische Entwicklung von Gymnasiasten von der Sexta bis zur Oberprima beobachtet wurde, konnte HITPASS (1967) eindeutig feststellen, daß die Zugehörigkeit zur sozialen Unterschicht trotz gleicher intellektueller Ausgangsbedingungen die Erfolgswahrscheinlichkeit auf dem Gymnasium vermindert. Die Befunde von GIESEN et al. (1967) stützen diese Ergebnisse. Die Autoren konnten nachweisen, daß auf dem Gymnasium — gemessen an ihren intellektuellen Leistungen — die Schüler der höheren Sozialschichten zu 80 % über- und die der niederen Sozialschichten zu 46 % unterrepräsentiert sind (siehe noch HELLER 1970).

Auf dem Gymnasium verringert sich nicht nur der Anteil von Schülern der unteren Sozialschichten, vielmehr ist auch ein „Schereneffekt" schulischer Leistungen wiederholt nachgewiesen worden. Damit ist der Tatbestand gemeint, daß mit zunehmender Schuldauer die Diskrepanz in den schulischen Leistungen (Noten) zwischen den Kindern der unteren und oberen Schichten zunimmt, obwohl zu Beginn der Gymnasialzeit keine signifikanten Unterschiede hinsichtlich der Intelligenztestleistungen vorhanden waren (z. B. SIMONS 1973).

Die soziale Auslese wird jedoch nicht vom Gymnasium allein betrieben. PETRAT (1964) weist beispielsweise entschieden darauf hin, daß sie mindestens schon im zweiten Grundschuljahr beginne. FERDINAND (1969) konnte dies bestätigen. Die Selektion der Kinder aus den unteren Schichten setzt sich bis zur Sonderschule für Lernbehinderte fort. FERDINAND & UHR (1973) fanden, daß von 265 normal intelligenten Kindern (IQ zwischen 90 und 116), die eine Lernbehindertenschule besuchten, 98 % Arbeiterkinder waren, also der Unterschicht angehörten.

Für diese Chancenunterschiede wurden umfangreiche Erklärungsversuche unternommen. Es muß in diesem Zusammenhang wiederum auf das Kapitel „Determinanten der Schulleistung" (S. 46 ff. in diesem Buch) verwiesen werden. Ein bedeutsames Merkmal zur Erklärung des relativen Schulversagens von Unterschichtkindern ist die Sprache. Formal zeigt sich dies auch in den Zeugnissen. In der Volksschule werden im Fach Deutsch die schlechtesten Noten erteilt. Schulversager scheitern hauptsächlich an der Rechtschreibung (KEMMLER 1967, FUNKE 1972).

MILLER (1970) zeigt, daß das Merkmal der sozialen Zugehörigkeit — definiert durch den Beruf des Vaters — zu global ist. Familiäre Verhältnisse und Erziehungsumwelt, die nicht mit dem Vaterberuf korrelierten, klärten etwa doppelt soviel Varianz auf. Damit ist *„als eigentliche Quelle der Schulleistungsvariation eben nicht der Beruf des Vaters, sondern die Institution Schule anzusehen."* (SIMONS 1973, S. 266).

5.2.3.4. Schulnoten und Persönlichkeit

Zusammenhänge zwischen Persönlichkeitsmerkmalen (außer Intelligenz), die durch Fragebogen meßbar sind, und Schulnoten sind eher unsystematisch und nur sporadisch untersucht worden.

ZIELINSKI (1967) fand bei Schülern der vierten Klasse nur sehr geringe negative Korrelationen zwischen *Ängstlichkeit* und der Deutsch- bzw. Rechennote ($r = -0{,}14$ bzw. $-0{,}21$). Da diese Beziehungen nicht linear sind, können die Korrelationskoeffizienten den tatsächlichen Zusammenhang unterschätzt haben. In einem varianzanalytischen Versuchsplan konnte Bärbel TEWES (1971) bei zehnjährigen Volksschülern deutlichere Beziehungen zwischen Ängstlichkeit und den Durchschnittsnoten aus Rechen- bzw. Deutsch-

klassenarbeiten feststellen. Die Autorin kommt zum Ergebnis: „Hohe Ängstlichkeitswerte sind in beiden Fächern Indikator für unterdurchschnittliche Leistungen, niedrige Ängstlichkeitswerte hingegen für durchschnittliche oder bessere Leistungen (TEWES 1971, S. 118). In Extremgruppenvergleichen konnte gezeigt werden, daß schlechte Volks- und Realschüler (Notendurchschnitt aus Deutsch und Rechnen schlechter als 4,5) signifikant höhere Angstwerte erreichen als gute Schüler (Durchschnitt 2,5 oder besser); vgl. NICKEL & SCHLÜTER (1970), NICKEL et al. (1973) sowie Seite 66 ff. in diesem Buch.

ENTWISTLE & CUNNINGHAM (1968) konnten an 13jährigen Schülern einen schwachen Zusammenhang zwischen Schulnoten und *Neurotizismus* feststellen. Höhere Neurotizismuswerte sind eher mit schlechten Noten verbunden. Einen durchgängigen Zusammenhang zwischen Schulnoten und *Extraversion* konnten die Autoren dagegen nicht finden. U. TEWES (1973) konnte ebenfalls keinen durchgängigen Zusammenhang zwischen Extraversions- und Neurotizismuswerten mit den Noten in Diktaten und Rechenarbeiten nachweisen.

Umfangreichere Untersuchungen legten SEITZ & LÖSER (1969) und SEITZ (1970 und 1971) vor. Nach deren Ergebnissen kann allgemein gesagt werden, daß Schulnoten schwach mit *Erwartung auf Mißerfolg* und mit *sozialer Zurückgezogenheit* korrelieren. Schüchterne Schüler, die sich selbst wenig zutrauen, erhalten tendenziell schlechtere Noten. Die genannten Zusammenhänge sind aber gering, so daß der Meinung TENTS (1969, S. 135 ff.) zugestimmt werden kann, nach welcher der Zusammenhang zwischen Schulnoten und *Leistungsmotivation* (d. h. Erwartung von Erfolg oder Mißerfolg) noch unklar ist. Allenfalls kann man annehmen, daß die Beziehungen zwischen Leistungsmotivation und Noten mit zunehmenden Lebensalter etwas enger werden (TENT 1969, S. 138, Tab. 9.5.1.). Die Befunde von SADER & SPECHT (1967) weisen dagegen eine umgekehrte Tendenz auf. In diesem Zusammenhang kann noch auf eine Untersuchung von TODT (1966) hingewiesen werden, in der die *Interessen* von Unterprimanern im sprachlichen Zweig höher mit den Noten korrelierten als Intelligenztestleistungen.

Insgesamt kann festgehalten werden, daß es einzelne Hinweise auf Zusammenhänge von Noten und Persönlichkeitsmerkmalen gibt. Die bisherigen Befunde ergeben jedoch kein klares Bild. Dieses Problem muß anscheinend mit spezifischeren Untersuchungsplänen und präziseren Verfahren als bisher angegangen werden.

5.2.4. Zusammenfassung

Schulnoten sind Urteile. Sie unterliegen daher den bekannten Urteilsfehlern. Durch diese Fehler werden die instrumentellen Eigenschaften der Noten

als Indikatoren für bestimmte Leistungen von Schülern von vornherein negativ beeinflußt.

Es zeigen sich fachtypische Häufigkeitsverteilungen der Noten. Im testtheoretischen Sinne muß den Noten eine ausreichende Objektivität und Reliabilität abgesprochen werden. Ihre Validität ist unklar. Es sollte jedoch deutlich geworden sein, daß unkritische Versuche, die Noten den testtheoretischen Modellen anzupassen, den pädagogischen Absichten des Unterrichtens und Erziehens widersprechen können. Die mangelnden Qualitäten der Noten sind nicht den Lehrern anzulasten, vielmehr einer Institution, die von ihnen Unmögliches verlangt. Es muß jedoch von jedem einzelnen Lehrer erwartet werden, daß er sich seine Praktiken der Notengebung kritisch bewußt macht.

Schulnoten variieren in mehr oder weniger systematischen Zusammenhängen mit der Intelligenz des Schülers, seiner sozialen Herkunft und bestimmten Merkmalen seiner Persönlichkeit. Auf ein Bild zusammengedrängt könnte man vielleicht sagen, daß eine wohlerzogene, jedoch selbstbewußte intelligente Tochter eines Akademikers alle Chancen hat, in unserem Schulsystem die besten Noten zu erhalten.

5.2.5. Literaturverzeichnis

Amelang, M. u. *Vagt, G.:* Warum sind die Schulnoten von Mädchen durch Leistungstests besser vorherzusagen als diejenigen von Jungen? Z. f. Entw. Psychol. u. Päd. Psychol., 1970, *2*, 210—220.

Argyle, M.: Soziale Interaktion. Köln, Kiepenheuer & Witsch, 1972.

Aschersleben, K.: Untersuchungen zur Reliabilität von Schulnoten. Schule u. Psychol., 1971, *18*, 147—154.

Beltz-Verlag: Deutsche Schultests. Informationsbroschüre und Gesamtverzeichnis. Weinheim, Beltz, 1968.

Brophy, J. E. u. *Good, T. L.:* Teacher's communication of differential expectations for children's classroom performance: Some behavioral data. J. of. Educ. Psychol., 1970, *61*, 365—374.

Burger, R.: Psychologische Übertrittsbegutachtung für ein Gymnasium. Psychol. Rdschau., 1967, *18*, 145—154.

Carter, R. S.: How invalid are marks assigned by teachers? J. of Educ. Psychol., 1952, *43*, 218—228.

Dohse, W.: Das Schulzeugnis. Sein Wesen und seine Problematik. Weinheim, Beltz, 1963.

Entwistle, N. J. u. *Cunningham, Shirley:* Neuroticism and school attainment — a linear relationship? Brit. J. of. Educ. Psychol., 1968, *38*, 123—132.

Ferdinand, W.: Der Lehrer mag mein Kind nicht leiden. Schule u. Psychol., 1962, *9*, 361—369.

Ferdinand, W.: Über Schulreife und Schulleistung IQ-äquivalenter Kinder aus unterschiedlichem sozialen Milieu. Z. Entw. Psychol. u. Päd. Psychol., 1969, *1*, 190—199.

Ferdinand, W.: Das Vorurteil des Lehrers über die Leistungsfähigkeit bestimmter Schüler im Spiegel von Aufsatzzensuren. Schule u. Psychol., 1971, *18*, 92—95.

Ferdinand, W. u. *Kiwitz, H.:* Über die Häufigkeitsverteilung der Zeugnisnoten 1 bis 6. In: *Ingenkamp, K.* (Hrsg.) 1971, 178—185.

Ferdinand, W. u. *Uhr, R.:* Sind Arbeiterkinder dümmer — oder letztlich nur „die Dummen"? Psychol. in Erz. u. Unterr., 1973, *20,* 31—35.

Fippinger, F.: Intelligenz und Schulleistung. Erziehung und Psychologie, 1966, *41.*

Fippinger, F.: Empirische Untersuchung zur Leistung von Schülern aus voll und wenig gegliederten Schulen. Schule u. Psychol., 1967, *14,* 96—103.

Flitner, A.: Das Schulzeugnis im Lichte neuerer Untersuchungen. Z. f. Päd., 1966, *12,* 511—538.

Funke, E. H.: Grundschulzeugnisse und Sonderschulbedürftigkeit. Berlin, Marhold, 1972.

Gebauer, T.: Vergleichende Untersuchung über den Voraussagewert von Aufnahmeprüfung und Testuntersuchung für den Erfolg auf weiterführenden Schulen. In: *Ingenkamp, K.* (Hrsg.) 1965, 97—141.

Giesen, H., Ullrich, M. u. *Tent, L.:* Elternberuf, Leistungspotential und Bildungschancen. Psychol. Beitr., 1967, *10,* 541—563.

Göller, A.: Zensuren und Zeugnisse. Stuttgart, Klett, 1966.

Hadley, T. S.: Feststellungen und Vorurteile in der Zensurierung. In: *Ingenkamp, K.* (Hrsg.) 1971, 134—141.

Hasemann, K.: Verhaltensbeobachtung. In: *Heiss, R.* (Hrsg.) 1964, 807—836.

Heiss, R. (Hrsg.): Psychologische Diagnostik. Handbuch der Psychologie Band 6. Göttingen, Hogrefe, 1964.

Heller, K.: Aktivierung der Bildungsreserven. Bern/Stuttgart, Huber/Klett, 1970.

Hitpass, J.: Bericht über eine 6jährige Bewährungskontrolle von Aufnahmeprüfung und Testprüfung. Schule u. Psychol., 1963, *10,* 211—218.

Hitpass, J.: Verlaufsanalyse des schulischen Schicksals eines Sextaner-Jahrganges von der Aufnahme bis zur Reifeprüfung. Schule u. Psychol., 1967, *14,* 371—378.

Hofer, M.: Die Schülerpersönlichkeit im Urteil des Lehrers. Weinheim, Beltz, 1969.

Höger, D.: Analyse der Intelligenzstruktur bei männlichen Gymnasiasten der Klassen 6—9 (Untersekunda — Oberprima). Psychol. Forsch., 1964, *27,* 419 bis 474.

Höhn, Elfriede: Der schlechte Schüler. München, Piper, 1967.

Hopp, Anna-Dorothea u. *Lienert, G. A.:* Eine Verteilungsanalyse von Gymnasialzensuren. Schule u. Psychol., 1965, *12,* 139—150.

Ingenkamp, K.: Die Jahrgangsklasse im Bild der Tempelhofer Untersuchungen. In: *Ingenkamp, K.* (Hrsg.) 1965, 255—283.

Ingenkamp, K.: Möglichkeiten und Grenzen des Lehrerurteils und der Schultests. In: *Roth, H.* (Hrsg.) 1968, 407—431.

Ingenkamp, K.: Sind Zensuren aus verschiedenen Klassen vergleichbar? In: *Ingenkamp, K.* (Hrsg.) 1971, 156—163.

Ingenkamp, K. (Hrsg).: Schulkonflikt und Schülerhilfe. Weinheim, Beltz, 1965.

Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim, Beltz, 1971.

Janssen, J. P.: Kritische Bemerkungen zu Validitätsstudien mit den Prädiktoren „Schulnoten" und „Intelligenztests". Diagnostica, 1972, *18,* 26—37.

Kemmler, Lilly: Erfolg und Versagen in der Grundschule. Göttingen, Hogrefe, 1967.

Kleiter, E.: Über Referenz-, Interaktions- und Korrelationsfehler im Lehrerurteil. Bildung und Erziehung, 1973, *26,* 100—117.

Kohl, G.: Eine vergleichende Begabungs- und Leistungsmessung in 4. Klassen der Dortmunder Volksschulen unter dem Gesichtspunkt der Auslese für weiterführende Schulen. Schule u. Psychol., 1964, *11*, 180—186.

Knoche, W.: Jungen, Mädchen, Lehrer und Schulen im Zensurenvergleich. Weinheim, Beltz, 1969.

Langfeldt, H. P. u. *Fingerhut, W.:* Der Beitrag biographischer Daten von Schülern und Lehrern zur Vorhersage von Schulnoten. 28. Kongreß DGfP, Saarbrücken, 1972. (Kongreßbericht im Druck).

Lienert, G. A. (Hrsg).: Bericht des 23. Kongresses der DGfP Göttingen, Hogrefe, 1963.

Lienert, G. A. u. *Hopp, Anna-Dorothea:* Über die Interkorrelation von Gymnasialzensuren. Schule u. Psychol., 1964, *11*, 193—206.

Löschenkohl, E.: Gibt es einen allgemeinen faßbaren Zusammenhang zwischen Schulleistung und Intelligenz? Psychol. in Erz. u. Unterr., 1973, *20*, 145—155.

Merz, F.: Die Beurteilung unserer Mitmenschen als Leistung. In: *Lienert, G. A.* (Hrsg.) 1963, 32—51.

Miller, G. W.: Factors in school achievement and social class. J. of Educ. Psychol., 1970, *61*, 260—269.

Müller-Fohrbrodt, Gisela u. *Dann, H. D.:* Zum Problem der Notengebung: Selbstbeurteilung von Zeugnisnoten. Z. Entw. Psychol. u. Päd. Psychol., 1971, *3*, 241—252.

Nickel, H. u. *Schlüter, P.:* Angstwerte bei Hauptschülern und ihr Zusammenhang mit Leistungs- sowie Verhaltensmerkmalen, Lehrerurteil und Unterrichtsstil. Z. f. Entw. Psychol. u. Päd. Psychol., 1970, *2*, 125—136.

Nickel, H., Schlüter, P. u. *Fenner, H. J.:* Angstwerte, Intelligenztest- und Schulleistungen sowie der Einfluß der Lehrerpersönlichkeit bei Schülern verschiedener Schularten. Psychol. in Erz. u. Unterr., 1973, *20*, 1—13.

Nickel, H. u. *Langhorst, E.* (Hrsg): Brennpunkte der Pädagogischen Psychologie. Bern/Stuttgart, Huber/Klett, 1973.

Orlik, K.: Ein Beitrag zu den Problemen der Metrik und der diagnostischen Valenz schulischer Leistungsbeurteilungen. Z. exp. angew. Psychol., 1961, *8*, 400 bis 408.

Petrat, G.: Soziale Herkunft und Schullaufbahn. Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt, 1964.

Rank, Therese: Schulleistung und Persönlichkeit. München, Barth, 1962.

Röhm, H.: Zur Problematik der Beurteilung von Schülern. Psychol. in Erz. u. Unterr., 1973, *20*, 77—88.

Rosenthal, R. u. *Jacobson, Lenore:* Pygmalion im Unterricht. Weinheim, Beltz, 1971.

Roth, H. (Hrsg.): Begabung und Lernen. Stuttgart, Klett, 1968.

Rubovitz, P. C. u. *Maehr, M. L.:* Pygmalion analysed: Toward an explanation of the Rosenthal-Jacobson findings. J. Pers. Soc. Psychol., 1971, *19*, 197—203.

Sader, M. u. *Specht, Heike:* Leistung, Motivation und Leistungsmotivation. Arch. ges. Psychol., 1971, *119*, 90—130.

Schmitz, G. F.: Grundschulleistung, Intelligenz und Übertrittsauslese. Erziehung u. Psychol., 1964, *29*.

Schultze, W.: Über den Voraussagewert der Auslesekriterien für den Schulerfolg an Gymnasien. Max Traeger Stiftung, Forschungsberichte. Frankfurt, 1964.

Schwarz, Elisabeth: Schulreife, Intelligenz und Schulleistung im ersten Schuljahr. Schule u. Psychol., 1967, *14*, 233—245.

Seitz, W.: Über den Zusammenhang von Persönlichkeitseigenarten, Schulnoten und HAWIK-Leistungen bei Volksschülern. Psychol. Beitr., 1970, *13*, 579—602.

Seitz, W.: Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztestleistungen bei Volksschülern. Z. exp. angew. Psychol., 1971, *18*, 307 bis 336.

Seitz, W. u. *Löser, G.:* Über die Beziehung von Persönlichkeitsmerkmalen zu Schul- und Intelligenztestleistungen bei Gymnasial-Schülern. Z. exp. angew. Psychol., 1969, *10*, 651—678.

Simons, H.: Intelligenz und Schulleistung. Schule u. Psychol. 1969, *16*, 307—318.

Simons, H.: Intelligenz- und Schulleistungen bei Arbeiter- und Akademikerkindern auf der Unterstufe des Gymnasiums. In: *Nickel, H.* u. *Langhorst, E.* (Hrsg.) 1973, 260—273.

Sixtl, F.: Meßmethoden der Psychologie. Weinheim, Beltz, 1967.

Steinkamp, G.: Die Rolle des Volksschullehrers im schulischen Selektionsprozeß. In: *Ingenkamp, K.* (Hrsg.) 1971, 256—276.

Süllwold, F.: Schultests. In: *Heiss, R.* (Hrsg.) 1964, 352—384.

Tent, L.: Das Leistungsprüfsystem (LPS) von *Horn, W.* bei Schülern des vierten Schuljahres. Zum Problem der Auslese für weiterführende Schulen, ein vorläufiger Bericht. Psychol. Beitr., 1965, *8*, 564—595.

Tent, L.: Die Auslese von Schülern für weiterführende Schulen. Göttingen, Hogrefe, 1969.

Tewes, Bärbel: Zusammenhänge zwischen Ängstlichkeit und Leistungsverhalten bei Schulkindern. ZeF, 1971, *5*, 107—118.

Tewes, U.: Emotionalität und Schulleistung: Einige Angaben zur Validität der HANES (KJ). Diagnostica, 1973, *19*, 40—45.

Todt, E.: Untersuchungen zur Vorhersage von Schulnoten. Psychol. Forsch., 1966, *26*, 32—51.

Ulich, D. u. *Mertens, W.:* Urteile über Schüler. Weinheim, Beltz, 1973.

Weingardt, E.: Korrelation und Voraussagewert von Zeugnisnoten bei Gymnasiasten. Erziehung u. Psychol., 1964, *31*.

Weiss, R.: Die Berechnung einer „Schulleistungszahl". Schule u. Psychol., 1964 a, *11*, 114—121.

Weiss, R.: Über den Zusammenhang zwischen Schulleistung und Intelligenz. Schule u. Psychol., 1964 b, *11*, 321—333.

Weiss, R.: Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. Schule u. Psychol., 1965, *12*, 252—269.

Weiss, R.: Über die Zuverlässigkeit der Ziffernnoten bei Rechenarbeiten. Schule u. Psychol., 1966, 13, 198—207.

Weiss, R.: Vor- und Nachteile der Leistungsbeurteilung durch Ziffernnoten. Schule u. Psychol., 1969, *16*, 198—207.

Weiss, R.: Über die Strenge der Benotung in verschiedenen Unterrichtsgegenständen. In: *Ingenkamp, K.* (Hrsg.) 1971, 832—843.

Zielinski, W.: Beziehungen zwischen Ängstlichkeit, schulischer Aktivität, Intelligenz und Schulleistung bei 9- bis 11jährigen Volksschülern. Schule u. Psychol., 1967, *14*, 265—273.

Zielinski, W.: Die Beurteilung von Schülerleistungen. Funkkolleg Pädagogische Psychologie, Studienbegleitbrief 12, S. 5—33. Weinheim, Beltz, 1973.

Zimmermann, K. W.: Über die Beziehungen zwischen Schulnoten von lernbehinderten Sonderschülern. Z. f. Heilpäd., 1968, *19*, 465—471.

5.3. Einflüsse auf die Beurteilung von Schüleraufsätzen Ergebnisse einer quasi-experimentellen Versuchsreihe

Horst Nickel und Wilhelm Wieczerkowski

5.3.1. Einleitung und Problemstellung

Seit dem Beginn einer empirischen Überprüfung der Leistungsbeurteilung in Schulen wurde in zunehmendem Maße auf die unzureichende Objektivität und Reliabilität unserer Schulzensuren verwiesen (AHRENS 1964, SCHRÖTER 1965 und 1973, INGENKAMP 1972). Die „Ungerechtigkeit" der Notengebung wird in den letzten Jahren auch immer stärker in der Öffentlichkeit unter Beteiligung aller Betroffenen diskutiert, und es werden Wege nach neuen verbesserten Bewertungsformen gesucht (LAURIEN 1973, LANDSHEERE 1973).

In ganz besonderem Maße richtet sich die an unserem Bewertungssystem geübte Kritik gegen die Beurteilungspraxis von Schul- und Prüfungsaufsätzen. Spätestens seit den aufsehenerregenden Ergebnissen von ULSHÖFER (1948/49), der feststellen mußte, daß ein und derselbe Reifeprüfungsaufsatz von verschiedenen Lehrern mit allen Stufen der Notenskala von 1 bis 6 bewertet wurde, haben sich zahlreiche Autoren kritisch mit der bisher üblichen Praxis der Aufsatzbeurteilung auseinandergesetzt (KARNICK 1959, FALK 1962, JÖRG 1964, SCHRÖTER 1965 und 1971, WEBER 1969).

Einen ersten empirischen Versuch zur Analyse der Bedingtheit dieser unterschiedlichen Benotung von Schüleraufsätzen im deutschen Sprachraum unternahmen KÖTTER & GRAU (1965); schon einige Jahre zuvor bemühten sich DIEDERICH u. a. (1961) im angelsächsischen Sprachgebiet darum, mit Hilfe eines faktorenanalytischen Ansatzes verschiedene Beurteilungskriterien unterschiedlicher Bewerter zu ermitteln. Auf die Praxis der Aufsatzbeurteilung hatten diese Arbeiten jedoch wenig Einfluß. Die pädagogischen Bemühungen um die Erarbeitung einheitlicher Bewertungskriterien blieben meistens im intuitiven Vorgehen verhaftet und führten eher zu einer noch größeren Vielfalt, Uneinheitlichkeit und Subjektivität, als daß sie diese zu überwinden vermochten (vgl. WEBER 1969).

Während im Bereich anderer schulischer Leistungsbeurteilung durch den Einsatz formeller und informeller Tests eine gewisse Objektivität und damit Einheitlichkeit und Gerechtigkeit der Beurteilung angestrebt wird, ist ein solches Vorgehen bei der Beurteilung von Aufsätzen jedoch nicht oder lediglich in äußerst beschränktem Umfang und auch nur für Teilaspekte der sprachlichen Leistung möglich. Um so notwendiger muß es deshalb erscheinen, nach Wegen zu einer größeren Objektivierung und Vereinheitlichung der Aufsatzbewertung zu suchen. Eine wichtige Voraussetzung dafür stellt

zweifellos die Kenntnis und Analyse der dabei wirksamen und relevanten Bedingungsvariablen dar.

Aufgabe der vorliegenden Untersuchungsreihe war es daher, in einer quasi-experimentellen Situation den Einfluß verschiedener vermutlich bedeutsamer Variablen auf die Bewertung von Aufsätzen zu erforschen. Zu diesem Zweck wurden drei aufeinander aufbauende Untersuchungen durchgeführt, deren verschiedene Fragestellungen sich aus der Kenntnis über die bei Beurteilungsvorgängen allgemein wirksamen Faktoren sowie aus den bisherigen praktischen Erfahrungen bei der Aufsatzbewertung ableiteten. Diese Versuchsreihe ist im weiteren zunächst in ihrer chronologischen Abfolge dargestellt. Das erscheint u. a. auch deshalb sinnvoll, weil sich dadurch für den Leser zugleich ein Einblick in den Aufbau und die fortschreitende Weiterentwicklung ursprünglicher Fragestellungen ergibt.

5.3.2. Die erste Untersuchung [1]

5.3.2.1. Ausgangsproblem und Fragestellungen

Eine hauptsächliche Variationsquelle für die mangelnde Übereinstimmung verschiedener Lehrer bei der Beurteilung desselben Aufsatzes kann zunächst in den individuellen Bezugssystemen der einzelnen Beurteiler vermutet werden (Jörg 1964). Es wäre nun denkbar, daß solche Bezugssysteme von der Art und Länge der Erfahrung in der Schulpraxis abhängen und daß der einzelne Lehrer im Laufe der Zeit objektivere Maßstäbe für seine Beurteilung gewinnt. Nach den Ergebnissen von Kötter & Grau (1965, S. 288) scheint das jedoch nicht der Fall zu sein. Demgegenüber fanden diese Autoren, daß systematische Unterschiede in den Bewertungen durch das Geschlecht der Beurteiler und durch bestimmte Kriterien der Arbeiten zustande kommen, wobei die Länge der Aufsätze eine entscheidende Rolle spielt (S. 295 ff.). Die Tatsache, daß andererseits auch gleiche Beurteiler bei wiederholter Benotung desselben Aufsatzes zu unterschiedlichen Ergebnissen kommen, könnte einmal im Sinne einer geringeren Urteilszuverlässigkeit, zum anderen aber auch als Folge von gleitenden Bezugssystemen gedeutet werden. Möglicherweise bilden sich die Bezugssysteme immer erst während des Beurteilungsvorgangs heraus, vor allem dann, wenn es gilt, eine Serie von Aufsätzen zu bewerten. Dabei dürften in der Schulpraxis auch noch bestimmte Erwartungshaltungen die Urteilsbildung beeinflussen, insbesondere Leistungserwartungen aufgrund der Kenntnis und bisherigen Beurteilung des betreffenden Schülers (vgl. Erlemeier & Tismer 1973).

---

[1] Vgl. *Wieczerkowski, W., Nickel, H.* u. *Rosenberg, L.*, 1968.

In der ersten Untersuchung sollten zunächst die folgenden *Fragen* überprüft werden:

1. Wird die Beurteilung eines Aufsatzes davon beeinflußt, welche anderen Aufsätze gleichzeitig zu bewerten sind, d. h. in welche Serien von Vergleichsreizen er eingebettet ist?
2. Wie wirkt sich eine unterschiedliche Information der Beurteiler über die Aufgabenstellung, speziell über die Ausgangssituation der Schüler bei Niederschrift der Aufsätze, auf die Bewertung aus?
3. Lassen sich Zusammenhänge zwischen der Beurteilungshöhe und gewissen Sprachmerkmalen in den Aufsätzen feststellen, und gibt es dabei Merkmale, die besonders in die Bewertung eingehen?
4. Ist ein experimenteller Ansatz zur Überprüfung der Aufsatzbewertung ohne eine entsprechende Ernstsituation überhaupt möglich, oder bilden sich dabei zu unsichere Bezugssysteme heraus, so daß nur zufällige Einflüsse erfaßt werden?

5.3.2.2. Versuchsablauf

5.3.2.2.1. Das Beurteilungsmaterial

Es bestand aus 32 Aufsätzen von Kindern aus vier Klassen eines vierten Volksschuljahres über ein CAT-Bild (Nr. 3, Bellak & Bellak 1949, vgl. Abb. 1). Dieses war den Schülern in der ersten Versuchsphase zur Gewinnung des Beurteilungsmaterials einzeln dargeboten worden. Entsprechend der CAT-Instruktion wurden sie anschließend aufgefordert, dazu schriftlich eine Geschichte zu erzählen. Zunächst nahm der jeweilige Klassenlehrer eine Bewertung der Niederschriften vor; anschließend sollte er 15 zufällig ausgewählte Aufsätze einer Rangfolge von 1 (bester Aufsatz) bis 15 (schlechtester Aufsatz) zuordnen. Davon wurden für die weitere Untersuchung jeweils 8 Aufsätze ausgewählt, und zwar diejenigen mit den Rangplätzen 1 und 2 (obere Leistungsgruppe), 6 und 7 (obere Mittelgruppe), 9 und 10 (untere Mittelgruppe) sowie 14 und 15 (untere Leistungsgruppe).

Es verblieben somit für die eigentliche Untersuchung pro Klasse 8 Aufsätze, insgesamt also 32. Diese wurden in Maschinenschrift vervielfältigt, nachdem orthographische Fehler und grobe grammatische Verstöße verbessert worden waren.

5.3.2.2.2. Der Versuchsplan

Die eigentliche quasi-experimentelle Untersuchung erfolgte nach einem zweifaktoriellen varianzanalytischen Versuchsplan (vgl. Abb. 2). Experimentell variiert wurde einmal der Zusammenhang, in dem die einzelnen Auf-

Abb. 1: Bildvorlage für die Aufsätze der ersten und zweiten Untersuchung (vgl. BELLAK u. BELLAK 1949)

sätze zu beurteilen waren (Faktor S), zum anderen die Informationen über die Aufgabenstellung bzw. Ausgangssituation der Schüler bei der Niederschrift der Aufsätze (Faktor I). Acht Aufsätze wurden von allen Vpn beurteilt (Reizmaterial), dazu bildeten jeweils verschiedene Serien von ebenfalls acht Aufsätzen die Bezugsreize. Damit ergaben sich insgesamt drei Se-

Abb. 2: Zweifaktorieller Plan der ersten Untersuchung

Faktor S (Serie)	(1) Bildvorlage	(2) Beschreibung	(3) Bildvorlage u. Beschreibung	(4) keine Information
Serie 1				
Serie 2				
Serie 3				

rien von je 16 Aufsätzen. Es handelte sich bei dem Faktor „Serien" also um einen dreistufigen Faktor.

Der Faktor „Information" enthielt die folgenden vier Stufen:

(1) Den Beurteilern wurde das CAT-Bild im Diapositiv vorgeführt.
(2) Die Beschreibung des Bildes wurde vorgelesen (entsprechend dem Manual von BELLAK & BELLAK 1949).
(3) Das Diapositiv und die Beschreibung des Bildes wurden vorgegeben.
(4) Die Beurteiler erhielten keinerlei Information über das Bild und die Ausgangssituation der Schüler.

Unter den Versuchsbedingungen 1—3 wurde den Beurteilern mitgeteilt, daß die Schüler zu dem Bild eine Geschichte schreiben sollten, unter der Versuchsbedingung 4 dagegen nur, daß es sich um Schülerniederschriften handelt. Alle Beurteiler wurden über den Altersjahrgang der beteiligten Schüler informiert.

5.3.2.2.3. Beurteiler und Bewertungsvorgang

Als Beurteiler wurden 115 Studierende der Erziehungswissenschaft beiderlei Geschlechts eingesetzt, vorwiegend zukünftige Volks- und Realschullehrer. Da sie sich ungleich auf die 12 Zellen des Versuchsplans verteilten, wurden sie auf 72 zufällig ausgewählte Beurteiler reduziert, das entspricht einer fünffachen Planwiederholung. Ihre Aufgabe bestand zunächst darin, die ihnen vorgelegte Aufsatzserie in eine Rangordnung zu bringen. In einem zweiten Durchgang sollten sie dann Zensuren von 1 (bester) bis 9 (schlechtester Aufsatz) erteilen. Beide Bewertungen wurden getrennt voneinander vorgenommen. Nachdem die schriftlich fixierten Beurteilungen des ersten Durchgangs (Rangreihe) eingesammelt waren, erhielten die Vpn die Anweisung, die Aufsatzserie sorgfältig zu mischen. Erst dann wurden sie über den zweiten Teil des Beurteilungsvorgangs informiert. Unerwünschte Einflüsse dürften daher weitgehend auszuschließen sein. An zusätzlichen Daten wurden ferner noch erhoben: Alter, Geschlecht, didaktisches Wahlfach der Vpn und Lese- sowie Beurteilungszeit.

5.3.2.3. Ergebnisse

5.3.2.3.1. Beurteilungsübereinstimmung und Geschlechtsunterschiede

Der mittlere Rangkoeffizient zwischen der Beurteilung durch Ränge und Noten beträgt 0.892 (s = 0.399). Die Variationsweite der Koeffizienten über die 16 Aufsätze reicht für 72 Beurteiler von 0.44 bis 0.99; jedoch sind sie nur bei drei Beurteilern geringer als 0.70 (Signifikanz: 0.425*, 0.601 **).

Die intraindividuelle Übereinstimmung ist somit hinreichend hoch; man kann davon ausgehen, daß das kurzzeitige Bezugssystem der Beurteiler für beide Beurteilungsarten weitgehend gleich ist, d. h. daß sie sich tatsächlich bemühten, zu einem differenzierten Urteil zu kommen und nicht nach dem Zufall bewerteten.

Die interindividuelle Beurteilerübereinstimmung wurde mit dem Konkordanzverfahren von KENDALL über die Rangurteile der Beurteiler geprüft. Die Ergebnisse sind in Tabelle 1 aufgeführt.

Tabelle 1

Konkordanzkoeffizienten von 12 Beurteilergruppen

	Serie			
Information	1	2	3	$\overline{W}$
Bild	.393	.724	.244	.455
Beschreibung	.506	.624	.609	.594
Bild u. Beschr.	.387	.362	.564	.438
ohne Inform.	.729	.820	.489	.669
W	.500	.643	.475	.531
Signifikanz:	.244 n.s.	.362**	.506***	

Bei einer mittleren Übereinstimmung von W = 0.531 streut die interindividuelle Konkordanz in den Beurteilergruppen sehr weit, und zwar von sehr gering (0.244) bis hoch (0.820). Eine eindeutige systematische Beeinflussung läßt sich dabei nicht erkennen; jedoch tendieren die Informationsgruppen (2) und (4) zu einer relativ einheitlicheren Beurteilung als die beiden übrigen. Das ist insofern überraschend, als gerade die Vpn der vierten Gruppe (keine Information) erhebliche Einwände gegen das Verfahren geäußert hatten.

Der Einfluß des Geschlechts der Vpn auf die Strenge oder Milde des Urteils wurde mittels einer 2 x 3-Tafel überprüft. Dazu wurde die Bewer-

tung der Aufsätze für männliche und weibliche Beurteiler nach ihrer Strenge in drei Gruppen eingeteilt. Der Chi-Quadrat-Test ergab noch eine Zufallswahrscheinlichkeit von mehr als 30 % (Chi-Qu. = 1.88; 2 Fg; $p > .30$). Die Art der Beurteilung kann damit als unabhängig vom Geschlecht der Vpn bezeichnet werden.

5.3.2.3.2. Verteilungsform und Streuung der Beurteilungen

Die Verteilungen der Zensuren von 72 Beurteilern für die von ihnen bewerteten 16 Aufsätze wurden auf Schiefe und Exzeß geprüft (vgl. LIENERT 1961, S. 168). Bei keinem Beurteiler wichen sie signifikant von einer Normalverteilung ab. Ebenso verteilten sich die mittleren Urteile der 72 Vpn normal (geprüft mit dem KOLMOGOROFF-SMIRNOFF-Test: $D_{max} = 0.099$; $D_{.05} = 0.161$; $D_{.01} = 0.192$); die mittleren Zensuren variierten zwischen 3.06 und 6.38 ($M_x = 4.65$; $s_x = 0.65$). Hinsichtlich der Milde oder Strenge des Urteils lassen die 72 Vpn also recht unterschiedliche Bezugssysteme erkennen.

Zur Überprüfung der Frage, ob die mittleren Beurteilungen in Abhängigkeit von der Urteilshöhe unterschiedlich streuen, wurden Quartile über die Durchschnittszensuren gebildet und die Streuungswerte mit dem H-Test von KRUSKAL und WALLIS (vgl. SIEGEL 1956, S. 184 ff.) auf signifikante Abweichung geprüft. Diese ergab einen auf dem 10 %-Niveau der Tendenz nach noch bedeutsamen Wert ($H = 6.83$; $p < .10$). Es zeichnete sich dabei der Trend ab, daß die Noten bei den am besten bewerteten Aufsätzen weniger streuen als bei allen übrigen. Das konnte auch bei einer Gegenüberstellung jener Aufsätze mit allen übrigen statistisch hoch gesichert werden ($u = 2.65$; $p = .004$ bei einseitiger Fragestellung). Insgesamt bestand jedoch zwischen den im Durchschnitt erteilten Zensuren und der Streuung der Urteile kein bedeutsamer Zusammenhang ($r = 0.10$).

5.3.2.3.3. Der Einfluß von Informationsart und Vergleichsreizen auf die Urteile

Zunächst wurde für die acht in allen drei Serien enthaltenen Aufsätze (Reizmaterial) der Einfluß von Informationsart (Faktor I) und Vergleichsreizen (Faktor S) mittels einer zweifaktoriellen Varianzanalyse überprüft. Sodann wurden nach den vier Informationsarten über die drei Serien der Bezugsreize einfache Varianzanalysen durchgeführt. Die Ergebnisse sind in Tabelle 2 und 3 dargestellt. Insgesamt ergab sich nur bei vier von 32 Aufsätzen ein bedeutsamer Einfluß der Informationsart. Eine Änderung des Bezugssystems der Vpn durch die verschiedenen Vergleichsserien ließ sich in keinem Falle nachweisen; wohl aber trat bei einem Aufsatz eine signifikante Wechselwirkung zwischen Information und Vergleichsreiz auf.

Tabelle 2

Zweifaktorielle Varianzanalysen über acht von allen 72 Vpn beurteilte Aufsätze (Reizmaterial)

Aufsatz	Zensuren		I Information		S Vergleichsserie		Wechselwirkung	
	$\bar{X}$	s	F	p	F	p	F	p
1	3,14	2,11	2,52	ns	0,25	ns	1,07	ns
2	3,31	1,71	0,49	ns	1,58	ns	0,88	ns
3	3,91	1,68	0,28	ns	0,62	ns	1,49	ns
4	4,03	1,76	4,94	.01	0,82	ns	2,09	ns
5	4,38	2,05	7,10	.01	2,18	ns	1,35	ns
6	5,92	1,68	1,20	ns	0,19	ns	1,24	ns
7	6,82	1,90	1,22	ns	0,02	ns	4,07	.01
8	7,22	1,72	3,91	.05	0,24	ns	2,16	ns

Freiheitsgrade: im Faktor I: 3/60, im Faktor S: 2/60, I X S: 6/60; df total = 71.

Tabelle 3

Einfache Varianzanalysen (Faktor Information) über drei Serien von Bezugsreizen

Aufsatz	Serie 1				Serie 2				Serie 3			
	$\bar{X}$	s	F	p	$\bar{X}$	s	F	p	$\bar{X}$	s	F	p
1	2,75	1,66	1,16	ns	1,87	1,48	0,47	ns	2,16	1,57	0,53	ns
2	2,91	1,58	2,91	ns	2,45	1,15	0,72	ns	3,12	1,42	0,35	ns
3	3,41	1,85	0,42	ns	3,75	1,81	0,99	ns	3,41	0,86	0,33	ns
4	3,70	2,11	0,50	ns	3,83	1,62	2,19	ns	4,12	1,74	2,00	ns
5	3,83	2,13	2,22	ns	3,91	1,98	1,25	ns	4,41	1,58	0,96	ns
6	3,87	1,88	0,09	ns	5,00	1,66	1,53	ns	5,29	1,74	0,33	ns
7	5,87	2,22	0,26	ns	6,95	1,62	0,43	ns	6,25	1,69	0,36	ns
8	6,00	1,68	1,48	ns	7,12	1,81	3,51	.05	6,25	2,52	2,17	ns

Freiheitsgrade: jeweils 3/20.

Der Einfluß von Lesezeit, Beurteilungszeit, Alter und didaktischem Wahlfach der Vpn auf die Beurteilung wurde mittels einfacher Varianzanalysen überprüft. In keinem Falle konnte ein signifikanter Wert ermittelt werden. Diese Faktoren besaßen offensichtlich keine Bedeutung für die Bewertung der vorgelegten Aufsätze.

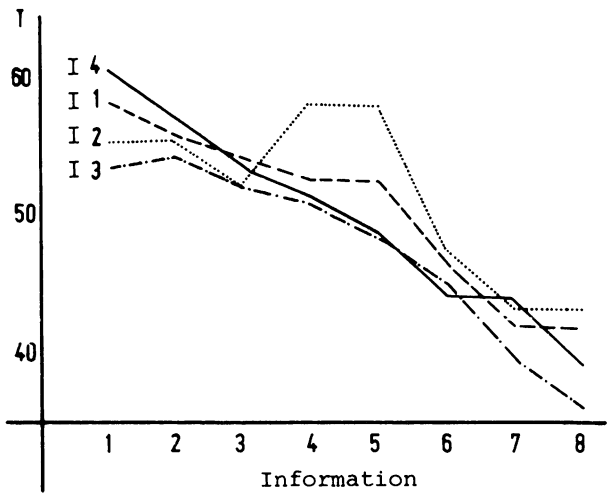
Die Beurteilungsprofile der acht von allen Vpn beurteilten Aufsätze (Reizmaterial) wurden nach einem von LIENERT (1963) vorgeschlagenen Verfahren einer Profilanalyse unterzogen. Die Ergebnisse sind in Tabelle 4 aufgeführt; die Beurteilungsprofile, getrennt nach Informationsart und Einbettung in Vergleichsreize, zeigt Abbildung 3.

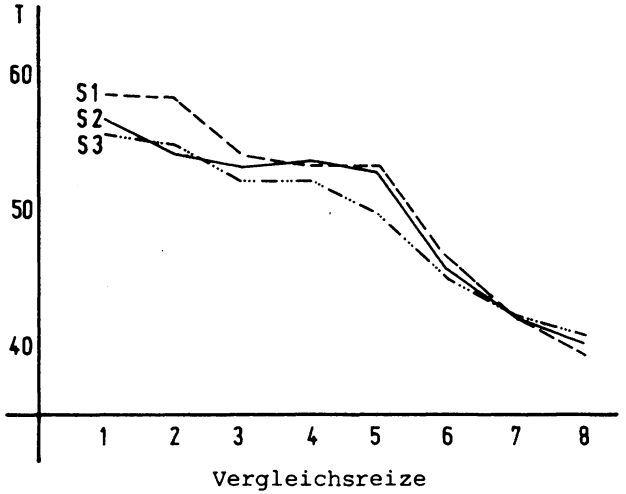
Tabelle 4

Profilanalyse über 8 von allen 72 Vpn beurteilten Aufsätze (Reizmaterial)

	QS	MQS	df	F	p
A Vergleichsserien	235,98	117,99	2	2,36	.10
B Information	887,94	295,38	3	5,91	.001
C Aufsätze	19875,69	2839,38	7	56,67	.001
A x B	1727,81	287,97	6	5,75	.001
A x C	389,41	27,81	14	0,56	ns
B x C	2475,05	117,86	21	2,35	.001
A x B x C	2438,91	58,07	42	1,16	ns
innerhalb	24047,83	50,10	480		
total	52078,62		575		

Abb. 3: Beurteilungsprofile für acht Aufsätze (Reizmaterial) unter den Bedingungen Information (I 1 bis I 4) und Vergleichsreize (S 1 bis S 3)



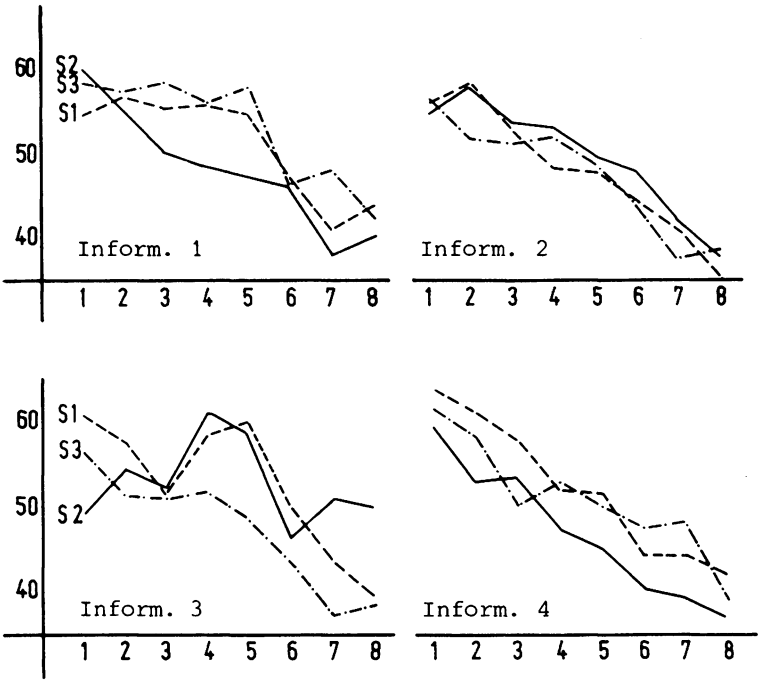


Die Profilhöhe ist signifikant unterschiedlich, die Aufsätze wurden also in bedeutsam abgestufter Weise beurteilt. Dabei hatten die Vergleichsserien auf die Profilgestaltung keinen bedeutsamen Einfluß (vgl. Wechselwirkung A x C). Anders verhält es sich mit der Informationsart (vgl. Wechselwirkung B x C); hier ergab sich ein sehr signifikanter F-Wert. Diese Ergebnisse waren aufgrund der oben angeführten Befunde der Varianzanalysen auch zu erwarten (vgl. Tabelle 2 und 3). Als hochsignifikant erwies sich auch die Wechselwirkung zwischen den Bedingungen Information und Bezugsreize (A x B). Die einzelnen Aufsätze der drei Vergleichsserien wurden also unter den vier Informationsarten unterschiedlich beurteilt, während die Serien allein keine signifikant verschiedene Profilgestalt erkennen ließen. Wie Abbildung 4 zeigt, scheint dabei für alle drei Serien unter den Informationsbedingungen „Beschreibung des Bildes" (2) und „keine Information" (4) eine Tendenz zu größerer Einheitlichkeit der Beurteilung zu bestehen als unter der Information „Diapositiv" (1) und „Diapositiv mit Beschreibung" (3).

5.3.2.3.4. Der Einfluß verschiedener Sprachmerkmale

Nach den bisher dargestellten Befunden ist zu vermuten, daß die Aufsätze selbst die stärksten differenzierenden Effekte auf die Beurteilung hervorgerufen haben. Es wurden daher einige Sprachvariablen ausgezählt und mit dem arithmetischen Mittel der Zensuren ($\overline{X}_Z$) sowie mit dessen Streuung (s_Z) in Beziehung gesetzt. Ferner wurde der Zusammenhang mit der

Abb. 4: Beurteilungsprofile für acht Aufsätze (Reizmaterial) in Abhängigkeit von drei Vergleichsserien (S 1 - S 3) und unter vier verschiedenen Bedingungen der Information



Zensur, die der Klassenlehrer erteilt hatte, und mit dessen Urteil über die Intelligenz der Kinder ermittelt (vgl. Tabelle 5).

Es ergab sich eine Reihe von Zusammenhängen, die im einzelnen der Tabelle entnommen werden können. Auffallend ist, daß zwischen den zwei Fehlervariablen (Kasusfehler und orthographische Fehler) und der Durchschnittszensur von 72 Beurteilern deutlich negative Beziehungen bestehen, obwohl die maschinengeschriebenen Aufsätze zuvor von solchen Fehlern gereinigt worden waren. Zwar ergab eine nachträgliche Durchsicht, daß bei dieser Verbesserung einige Versehen unterlaufen waren bzw. daß sich beim Abschreiben erneut Fehler eingestellt hatten, doch verteilten sich diese völlig unsystematisch über die Aufsätze. Da die Beurteiler ferner über die Fehlerberichtigung instruiert worden waren, darf man wohl annehmen, daß solche zufällig noch vorhandenen Fehler kaum als Ankerreize für den Aufbau eines Bezugssystems bei der Bewertung dienen konnten, wie das für

Tabelle 5

Korrelationen zwischen einigen Sprachmerkmalen in Schüleraufsätzen und verschiedenen Beurteilungsaspekten

	Beurteilung durch 72 Vpn (Zensuren)		Lehrerurteil	
	$\bar{x}_Z$	s_Z	Aufsatz	Intelligenz d. Schülers
Anzahl Sätze	.43	.17	-.03	-.02
Anzahl Wörter	.78	-.02	.22	.32
Wörter pro Satz	.47	-.24	.28	.40
Substantive	-.29	.06	-.41	-.26
zusammenges.Subst.	-.17	.13	.23	.12
Adjektive, prädik.	.27	-.17	.36	.13
Adjekt.,attributiv	.27	-.27	.42	.32
Verben	.05	.30	-.18	-.17
zusammenges. Verben	.39	.03	.55	.52
nicht wiederholte Verb.	.02	-.16	.34	.43
lange Wörter	.01	.52	.34	.20
direkte Rede	.21	.21	.11	.07
Da-,dann-Sätze	-.04	-.27	-.42	-.30
komplex Sätze	.05	.01	.12	.09
präposit. Gruppen	.08	-.28	-.10	.02
Kasusfehler	-.31	-.22	-.52	-.58
orthogr. Fehler	-.46	.14	-.68	-.74

Signifikanz: $.35^{*}$, $.45^{**}$

die Beurteilung durch den Klassenlehrer der Fall gewesen zu sein scheint. Es liegt vielmehr die Vermutung nahe, daß die mangelnde Beherrschung von Rechtschreibung und Grammatik mit anderen Schwächen der Darstellung kovariiert und daß darin die hauptsächliche Ursache für die negative Korrelation mit den Zensuren zu sehen ist. Eine gewisse Bestätigung dieser Annahme liefert die weitere Analyse.

Die ausgezählten Sprachmerkmale wurden nach dem Hauptachsenverfahren faktorisiert und einer Varimax-Rotation unterzogen. Die varimax-rotierten Faktoren wurden anschließend nach FISCHER-ROPPERT nochmals kriteriums-rotiert; die Ergebnisse zeigt Tabelle 6.

Für die sechs extrahierten Sprachfaktoren wurden proportional zu den Ladungshöhen grobgewichtete Merkmalsscores für die einzelnen Aufsätze errechnet (kritischer Ladungswert $\leqq |.40|$; Gewichte: .40 bis .69 = 1; .70 bis .89 = 2; .90 und mehr = 3). Diese Scores sind also vereinfachte Faktorenscores. Die Interkorrelationen zwischen den grobgewichteten Merkmalsscores sind aus Tabelle 7 zu ersehen.

Die grobgewichteten Merkmalsscores sind demnach insgesamt als relativ unabhängig zu bezeichnen. Gewisse Beziehungen bestehen jedoch zwischen

Tabelle 6

Übersicht über sechs extrahierte Sprachfaktoren

Faktor 1		Faktor 2	
Komplexe Sprachmuster		Länge der sprachlichen Produktion	
komplexe Sätze	.73		
Wörter pro Satz	.68	Anzahl Wörter	.96
prädik. Adjekt.	.66	Satzanzahl	.72
direkte Rede	.50	direkte Rede	.44
präpos. Gruppen	.38	nicht wiederholte Verben	-.36
Anzahl Sätze	-.45		
Anzahl Substant.	-.60		

Faktor 3		Faktor 4	
Fehlerhafte Darstellung (pos.: Sprachrichtigkeit)		Gebrauch differenzierender Wörter	
Orthogr. Fehler	.83	Lange Wörter	.81
Kasusfehler	.53	zusammengs. Substant.	.77
Wörter pro Satz	-.40	Kasusfehler	-.38
zusammenges. Verb.	-.49	präpos. Gruppen	-.67
nicht wiederholte Verben	-.72		

Faktor 5		Faktor 6	
Attributiver Stil		Verbbetonter Stil	
attributive Adj.	.91	Verben	.77
zusammenges. Verb.	.49	Da-,dann-Sätze	.53
Da-,dann-Sätze	-.53	direkte Rede	-.41
		präpos. Gruppen	-.43

Anteile der Faktoren an der totalen Varianz: 17,14,12,12, 11, 9 %

Insgesamt wurden 75 % der totalen Varianz extrahiert.

den Merkmalen „fehlerhafte Darstellung" einerseits und „komplexe Sprachmuster" sowie „attributiver Stil" andererseits, die in Richtung der oben geäußerten Vermutung laufen. Komplexe Sprachmuster und der attributive Gebrauch von Adjektiven bezeichnen wohl eine Darstellungsform, durch die Sachverhalte sprachlich differenzierter wiedergegeben werden. Dazu sind Schüler, die mit orthographischen und grammatischen Schwierigkeiten zu kämpfen haben, offensichtlich weniger in der Lage (vgl. dazu Faktor 3, Tabelle 6).

Tabelle 7

Interkorrelationen zwischen den grobgewichteten Merkmalsscores

	1	2	3			
1 Komplexe Sprachmuster	.-					
2 Länge der Produktion	-.11	—				
3 Fehlerhafte Darstellung	-.23	.15	—			
4 Differenzierende Wörter	-.10	.10	-.10	—		
5 Attributiver Stil	.13	.18	-.36	.11	—	
6 Verbbetonter Stil	-.38	.08	.24	.13	-.22	-

Signifikanz: .35*, .46**

Zwischen diesen gewichteten Merkmalsscores und den von 72 Vpn erteilten durchschnittlichen Zensuren sowie deren Streuung wurden Korrelationen berechnet; ferner wurden die Merkmalsscores zu der Bewertung durch den Klassenlehrer und seiner Schätzung der Intelligenz in Beziehung gesetzt. Die Ergebnisse sind in Tabelle 8 aufgeführt.

Tabelle 8

Korrelationen zwischen den grobgewichteten Merkmalsscores und der Bewertung durch 72 Vpn (Zensuren) sowie zwei Lehrerurteilen

	Beurteilung durch 72 Vpn (Zensuren)		Beurteilung durch den Klassenlehrer	
			Intell.-schätzung	Zensur
	$\bar{x}_Z$	s_Z		
1. Komplexe Sprachmuster	.20	-.10	.31	.23
2. Länge der Produktion	.68	.10	.14	.20
3. Sprachrichtigkeit	-.38	.08	-.71	-.78
4. Differenz. Wörter	-.10	.40	.35	.16
5. Attributiver Stil	.31	-.08	.57	.46
6. Verbaler Stil	.08	.13	-.35	-.38

Signifikanz: .35*, .45**

Eine systematische Differenzierung der Aufsatzbeurteilung durch die 73 Vpn erfolgte demnach zu einem großen Teil aufgrund der Aufsatzlänge. Das bedeutet, daß die Beurteiler ihr Bezugssystem hauptsächlich an einem

Kriterienkomplex verankerten, der — zumindest nach unseren Ergebnissen — relativ unabhängig von anderen Sprachkriterien ist (vgl. Tabelle 7). An zweiter Stelle — allerdings in geringerem Ausmaß — orientierte sich die Beurteilung an der fehlerhaften Darstellung, obwohl diese vornehmlich nur indirekt über Zusammenhänge mit anderen Kriterien (vgl. Tabelle 7) wirksam gewesen sein konnte.

Das Urteil der Klassenlehrer scheint sich noch in einem höheren Grad an der Fehlerhaftigkeit der Sprache zu orientieren und daneben in geringerem Umfang aber auch an Kriterien einer differenzierten Darstellung, insbesondere an einem attributiven Stil. Auch die Intelligenzschätzung durch den Lehrer bezieht sich hauptsächlich auf diese beiden Merkmale. Zwar lassen die Urteile der als Bewerter eingesetzten Vpn und die der Klassenlehrer auf eine unterschiedliche Verankerung der Bezugssysteme schließen, inwieweit das jedoch ein Ergebnis der speziellen Versuchsbedingungen ist, bei der die orthographischen und grammatischen Fehler eliminiert worden waren, muß offen bleiben. Andererseits kovariieren die Lehrerurteile und die mittleren Urteile der Bewerter auch jetzt noch relativ gut (0.63), während zwischen der Intelligenzschätzung der Lehrer und der durchschnittlichen Beurteilung durch die Vpn nur ein mittlerer Zusammenhang besteht (0.44).

Die Prüfung der Unterschiede in den Merkmalsscores von Niederschriften mit einem geringeren und von solchen mit einem größeren Beurteilungskonsensus erbrachte keine neuen Aufschlüsse. Nur in den Merkmalskriterien „Länge der Produktion" traten sehr signifikant abweichende Werte auf: Die sehr unterschiedlich bewerteten Schüleraufsätze waren signifikant kürzer als die übrigen.

5.3.3. Die zweite Untersuchung [2]

5.3.3.1. Fragestellung

Die anschließende zweite Untersuchung galt der Frage, inwieweit die Bewertung einer Serie von Schüleraufsätzen, die sich hinsichtlich ihrer Güte eindeutig in eine Rangreihe einordnen lassen, durch bestimmte induzierte Erwartungshaltungen bezüglich des Leistungsverhaltens der Aufsatzschreiber systematisch beeinflußt werden kann.

Zugleich wurde dabei die Hypothese geprüft, daß eine solche Beeinflussung in stärkerem Ausmaße auch abhängig ist von den in den Aufsätzen selbst enthaltenen differenzierenden Sprachfaktoren. Diese Annahme war durch die Ergebnisse der ersten Untersuchung nahegelegt worden. Ferner sollte nochmals überprüft werden, inwieweit die Bewertung von Schüler-

[2] Vgl. *Wieczerkowski, W.* u. *Keßler, G.* 1970.

aufsätzen durch das Geschlecht der Beurteiler bzw. der Aufsatzschreiber sowie durch einen unterschiedlichen Erfahrungshintergrund der Beurteiler eine systematische Differenzierung erfährt.

5.3.3.2. Versuchsablauf

5.3.3.2.1. Das Beurteilungsmaterial

Das Beurteilungsmaterial bestand wiederum aus den Aufsätzen zu einem CAT-Bild (vgl. Abb. 1). Sie wurden derselben Serie von 32 Niederschriften entnommen, die bereits der ersten Untersuchung zugrundelagen. Auch das Auswählverfahren war dasselbe, jedoch setzte sich jede Serie jetzt nur noch aus sechs Aufsätzen zusammen, bei denen die erteilten Noten die geringsten Streuungen aufwiesen, und zwar aus je zwei guten, mittleren und schlechten Niederschriften. Durch die vorgenommene Selektion sollte der Einfluß unterschiedlicher individueller Bezugssysteme bei der weiteren Beurteilung im Rahmen unseres Experiments möglichst reduziert werden.

Die ausgewählten Aufsätze wurden wiederum von orthographischen Fehlern und von gröberen grammatischen Verstößen gereinigt und in Maschinenschrift vervielfältigt.

5.3.3.2.2. Die Beurteiler und ihre experimentelle Beeinflussung

Die Beurteilerstichprobe setzte sich aus je 80 (40 männlichen und 40 weiblichen) Schülern der Obersekunda, Lehrerstudenten, Psychologiestudenten und Junglehrern vor der zweiten Staatsprüfung zusammen. Die Beurteilung der Aufsätze erfolgte nach der in Schulen üblichen Notenskala von 1 (= sehr gut) bis 6 (= ungenügend), wobei auch Zwischenzensuren erteilt werden konnten (z. B. 1—2, 2—3 usw.).

Die experimentelle Beeinflussung der Beurteiler wurde durch eine systematisch variierte Information über das angebliche Leistungsverhalten der Aufsatzschreiber vorgenommen; sie erfolgte in schriftlicher Form. Die Beurteiler wurden veranlaßt, sich jeweils zuerst mit dieser dem Aufsatz zugeordneten Leistungsmitteilung über den Schüler vertraut zu machen und erst dann die Bewertung des Aufsatzes vorzunehmen.

Die Informationen bestanden aus zwei Serien zu je drei Stufen: Serie 1 enthielt in sich widerspruchsfreie Informationen über das Leistungsverhalten mit den Stufen (a) gute, (b) befriedigende, (c) mangelhafte Leistungen des Schülers, Serie 2 in sich widersprüchliche Informationen mit den Stufen (w—a) gute, (w—b) befriedigende, (w—c) mangelhafte Leistungen des Schülers. Die Verteilung der Leistungshinweise über die Aufsätze geht aus Tabelle 9 hervor.

Tabelle 9

Verteilung der Leistungsinformationen über die Aufsätze

Leistungsinformation	A gute Aufsätze (Nr. 1 + 2)	B durchschnittliche Aufsätze (Nr. 3 + 4)	C schlechte Aufsätze (Nr. 5 + 6)
1	a - gut	b - befriedigend	c - mangelhaft
2	b - befriedigend	c - mangelhaft	a - gut
3	c - mangelhaft	a - gut	b - befriedigend
4	wa - gut widersprüchlich	wb - befriedigend widersprüchlich	wc - mangelhaft widersprüchlich
5	keine Informationen über das Leistungsverhalten		

5.3.3.2.3. Der Versuchsplan

Es handelt sich um einen vier-faktoriellen Versuchsplan (vgl. Abb. 5). Experimentell variiert wurden:

Faktor S (= Stichprobenzugehörigkeit): je 80 Schüler, Lehrerstudenten, Psychologiestudenten, Junglehrer;

Faktor G_B (= Geschlecht der Beurteiler): jeweils 40 männliche und 40 weibliche Beurteiler in 4 Stichproben;

Faktor I (= Information über das angebliche Leistungsverhalten der Aufsatzschreiber): 5 Informationsstufen, variiert über jeden Aufsatz (vgl. Tab. 9);

Faktor G_A (= Geschlecht der Aufsatzschreiber): Jeder Aufsatz wurde über 5 Informationsstufen jeweils einem Jungen und einem Mädchen zugeschrieben.

Für jeden Aufsatz ergab sich somit zunächst ein 4 x 2 x 5 x 2 = 80zelliger Versuchsplan mit einer Zellenbesetzung von je 4 Beurteilern.

5.3.3.3. Ergebnisse

5.3.3.3.1. Globalanalyse und Reduktion des Versuchsplans

Der Einfluß der vier Faktoren wurde für jeden Aufsatz einzeln geprüft. Die Homogenität der Zellenvarianzen wurde mit Hilfe des F_{max}-Tests von HARTLEY (s. WINER 1962, S. 93) kontrolliert. Kein F-Wert erreichte dabei das 1 %-Niveau. Für alle Aufsätze ergaben sich signifikante Einflüsse des Informationsfaktors, für die meisten auch solche des Stichprobenfaktors. Die Faktoren G_B und G_A waren dagegen für die Beurteilung der Schüleraufsätze bedeutungslos. Systematische Wechselwirkungen zwischen den vier Faktoren traten nicht auf. Von insgesamt 66 Wechselwirkungen erreichten nur 10 das 5 %- bzw. das 1 %-Niveau, ohne daß sich ein systematischer Trend beobachten ließ.

Abb. 5: Vierfaktorieller Plan der zweiten Untersuchung

Faktor S		Schüler		Lehrer-studenten		Psychologie-studenten		Junglehrer	
Faktor G_B		m	w	m	w	m	w	m	w
Faktor I / Faktor G_A									
Stufe 1	m								
	W								
Stufe 2	m								
	w								
Stufe 3	m								
	w								
Stufe 4	m								
	w								
Stufe 5	m								
	W								

Da eine systematische experimentelle Beeinflussung der Aufsatzbewertung durch die Faktoren I (Informationsart) und S (Stichprobenzugehörigkeit) erfolgte, nicht aber durch das Geschlecht der Beurteiler (G_B) bzw. der Aufsatzschreiber (G_A), wurde der vierfaktorielle Plan auf einen zweifaktoriellen reduziert. Eine eindrucksmäßige Analyse der Daten ließ es außerdem ratsam erscheinen, den ursprünglich komplexen Versuchsplan aufzuspalten, um die Wirkungen der Faktoren S und I eindeutiger abgrenzen zu können.

Die Daten wurden daher erneut in zwei getrennten zweifaktoriellen Varianzanalysen geprüft:

Versuchsplan 1

Faktor S: Stichprobenzugehörigkeit (4 Stufen)
Faktor I_1: Informationsart (3 Stufen in sich widerspruchsfreier Information: (1), (2), (3) — vgl. Abb. 5)

Versuchsplan 2

Faktor S: Stichprobenzugehörigkeit (4 Stufen)

Faktor I_2: Informationsart (3 Stufen: (1) adäquate Information, (4) adäquat-widersprüchliche Information, (5) keine Information — vgl. Abb. 5).

5.3.3.3.2. Effekte der Informationsart auf die Aufsatzbewertung

Die in sich widerspruchsfreien Informationen über das Leistungsverhalten der Aufsatzschreiber beeinflußten die Urteile über die Schüleraufsätze in bedeutsamer Weise (vgl. Tab. 10, 11 und Abb. 6). Analog zu der in der Information vorgegebenen Leistungsqualifizierung wurde ein und derselbe Aufsatz mit unterschiedlicher Strenge bewertet.

Tabelle 10

F-Werte aus zweifaktoriellen Varianzanalysen

(a) 3 Stufen in sich widerspruchsfreier Information (gut, befriedigend, mangelhaft)

Aufsatz:	1	2	3	4	5	6	df
Faktor I_1	3,82*	9,03	29,76	6,05	5,14	6,93	2;180
Faktor S	0,94	3,09	1,83	0,78	9,48	7,07	3;180
I x S	0,51	1,64	1,02	3,05	2,55	6,16	6;180

(b) 3 Informationsstufen: adäquate, adäquat-widersprüchliche, keine Information

Aufsatz:	1	2	3	4	5	6	df
Faktor I_2	0,10	3,93*	2,80	0,62	2,57	3,97	2;180
Faktor S	2,59	21,29	0,00	19,99	18,81	6,17	3;180
I x S	0,21	0,89	1,34	5,20	0,22	0,11	6;180

Signifikanz: * P <.05; ** P <.01; *** P <.001

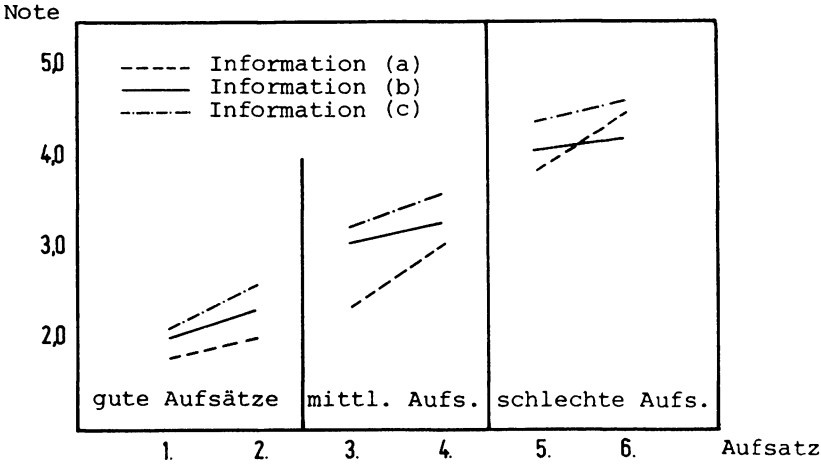
Der unter dem Einfluß der jeweiligen Information zu beobachtende monotone Urteilstrend (vgl. Tab. 11 u. Abb. 6) wurde nur für den schwächsten Aufsatz (Nr. 6) aufgehoben: Die Beurteiler benoteten diese Niederschrift schärfer unter dem Einfluß der Information „guter Schüler" (a) als bei Vorgabe der Information „durchschnittlicher Schüler" (b).

Tabelle 11

Mittelwerte der erteilten Aufsatznoten in Abhängigkeit von der jeweiligen Information

Information / Aufsätze		gut	befriedigend	mangelhaft	adäquat	adäquat/ widersprüchl.	ohne Information
Nr. 1	= gut	1,78	2,01	2,09	1,78	1,79	1,77
Nr. 2		2,03	2,31	2,60	2,03	2,40	2,25
Nr. 3	= mittel	2,35	3,06	3,23	3,06	2,85	2,77
Nr. 4		3,05	3,27	3,62	3,27	3,16	3,30
Nr. 5	= schlecht	3,87	4,10	4,40	4,40	4,28	4,01
Nr. 6		4,52	4,25	4,66	4,66	4,55	4,22

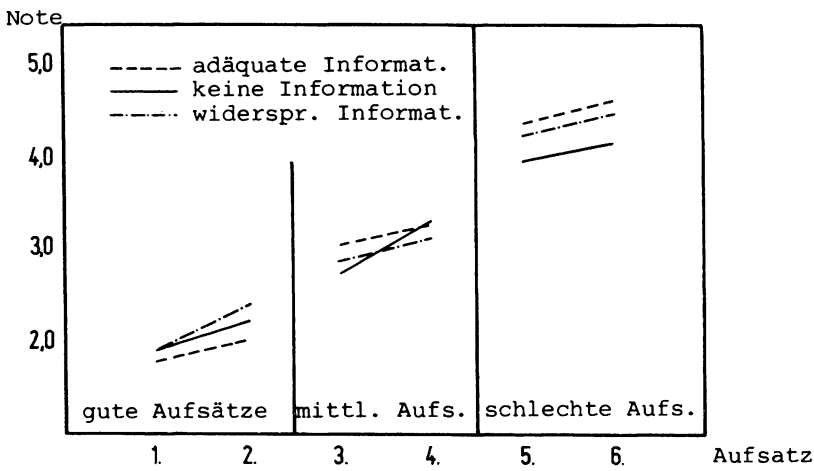
Abb. 6: Mittelwerte der erteilten Aufsatznoten in Abhängigkeit von der jeweiligen Information (Versuchsplan 1)



Die durch die Selektion des Beurteilungsmaterials vorgegebene qualitative Abstufung der Aufsätze erwies sich dabei den Informationseinflüssen gegenüber als relativ invariant. Die Beurteiler modifizierten ihre Urteile lediglich innerhalb bestimmter, vermutlich durch die Güte der Aufsätze festgelegter Toleranzgrenzen. Nur in einem Aufsatz (Nr. 3) trat unter dem Einfluß der Information „guter Schüler" (a) eine Überschneidung auf.

Ein signifikanter Einfluß des Informationsfaktors bei Vorgabe adäquater, adäquat-widersprüchlicher und keiner Information war zwar nur bei zwei Aufsätzen nachzuweisen; der Tendenz nach konnte bei widersprüchlichen Informationen jedoch eine Urteilsbeeinflussung in der beabsichtigten Richtung beobachtet werden. Bei den guten Aufsätzen sank die mittlere Note leicht ab, bei den mittleren und schlechten Aufsätzen stieg sie etwas an (vgl. Abb. 7). Ebenso trat tendenziell eine Urteilsverschärfung bei adäquater Leistungsinformation auf.

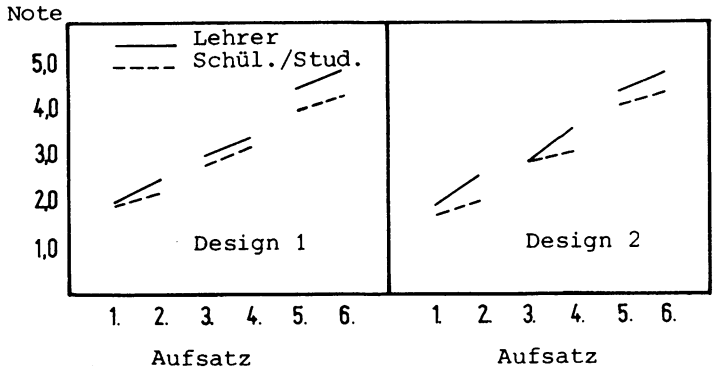
Abb. 7: Mittelwerte der erteilten Aufsatznoten in Abhängigkeit von der jeweiligen Information (Versuchsplan 2)



5.3.3.3.3. Einflüsse der Stichprobenzugehörigkeit

Die Prüfung mit dem NEWMAN-KEULS-Test (WINER 1962) ergab, daß sich insbesondere die erfahreneren Beurteiler (Junglehrer) von den unerfahrenen (Schülern, Lehrerstudenten, Psychologiestudenten) durch eine Neigung zu strengerer Benotung unterschieden. Die Schätzung der durch den Erfahrungshintergrund determinierten Varianzanteile mittels Omega ^{2} (HAYS 1963) zeigte, daß sich dieser im ersten Versuchsplan (drei in sich widerspruchsfreie Informationen) im geringeren Ausmaße auswirkte als im zweiten mit den Informationsstufen adäquate, adäquat-widersprüchliche, keine Informationen (vgl. Abb. 8).

Abb. 8: Mittelwerte der von Lehrern und von Schülern/Studenten erteilten Aufsatznoten



5.3.3.3.4. Wechselwirkungen von Informationsart und Erfahrungshintergrund der Beurteiler

Signifikante Wechselwirkungen traten bei Vorgabe abgestufter, in sich widerspruchsfreier Information (Plan 1) bei den drei schwächeren Aufsätzen (Nr. 4, 5, 6) hervor. Bei der Vorgabe adäquater, adäquat-widersprüchlicher bzw. keiner Information (Plan 2) dagegen waren sie nur in einem Aufsatz (Nr. 4) nachzuweisen.

Abb. 9: Mittelwerte der erteilten Aufsatznoten in Abhängigkeit von der jeweiligen Information bei unerfahrenen und erfahrenen Beurteilern (Versuchsplan 1)

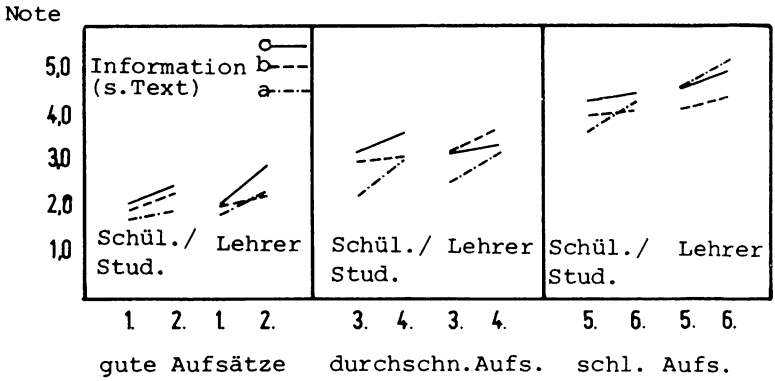
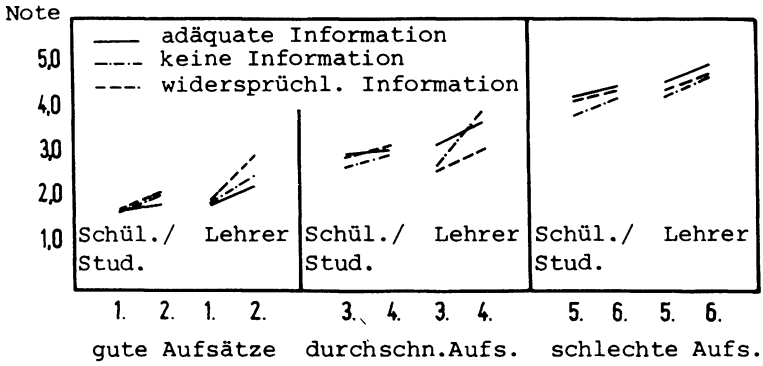


Abb. 10: Mittelwerte der erteilten Aufsatznoten in Abhängigkeit von der jeweiligen Information bei unerfahrenen und erfahreneren Beurteilern (Versuchsplan 2)



Wie die Abbildungen 9 und 10 zeigen, ließen sich vor allem die unerfahrenen Beurteiler durch entsprechende verbale Hinweise in ihrem Urteil beeinflussen, während die Junglehrer in einigen Aufsätzen differenzierter reagierten. Sie tendierten eher dazu, die durchschnittlichen Aufsätze bei einer adäquaten Leistungsinformation strenger zu beurteilen und bei den schwächsten Aufsätzen ebenso zu verfahren, wenn über den Aufsatzschreiber ein guter Hinweis (Information a) erteilt wurde.

5.3.4. Die dritte Untersuchung [3]

5.3.4.1. Fragestellung

In einer weiterführenden dritten Untersuchung wurden schließlich noch zwei Fragen in einer differenzierteren Form überprüft, die sich aus den Befunden der vorangegangenen Arbeiten ergeben hatten. Es handelt sich einmal um den Einfluß der bisherigen Lehr- bzw. Beurteilungserfahrung der Bewerter, zum anderen um die Bedeutung verschiedener Sprachkriterien für die Art der Beurteilung. Damit standen bei dieser Untersuchung neben anderen hier weniger bedeutsamen die beiden folgenden Fragestellungen im Mittelpunkt:

---

[3] Die Ergebnisse wurden auszugsweise einem unveröffentlichten Manuskript von *Wieczerkowski, W.* u. *Schiebel, W.* entnommen: „Über den Einfluß von Leistungserwartungen auf die Benotung von Schüleraufsätzen durch Lehrer".

a) Wird eine Aufsatzserie von Beurteilern mit verschieden langer Schulpraxis in unterschiedlicher Weise bewertet?
b) Besteht ein Zusammenhang zwischen den Aufsatzzensuren und einigen inhaltlichen und stilistischen Sprachkriterien?

5.3.4.2. Versuchsablauf

5.3.4.2.1. Das Beurteilungsmaterial

Die zu beurteilenden Aufsätze wurden wiederum einer Stichprobe von Niederschriften zu einer CAT-Vorlage entnommen (BELLAK & BELLAK 1949), jedoch handelte es sich diesmal um das Bild Nr. 9 (vgl. Abb. 11). Um jedoch die unterschiedliche Länge der Aufsätze, die nach den Ergebnissen der ersten Untersuchung die Beurteilung in entscheidender Weise beeinflußt, als Variable auszuschalten, wurden diesmal nur solche Niederschriften ausgewählt, die annähernd gleich lang waren (82 bis 100 Wörter). Diese wurden in einer Voruntersuchung 46 Studierenden der Erziehungswissenschaft zur Beurteilung vorgelegt. Für die Hauptuntersuchung erfolgte dann eine weitere Auswahl derart, daß von den 18 zu beurteilenden Aufsätzen fünf herangezogen wurden, die die vorgegebene Notenskala von „eins" bis „sechs" bestmöglichst umspannten und diese zugleich in annähernd gleiche Abschnitte teilten. Dabei hatten sich für die fünf endgültig ausgewählten Niederschriften folgende Mittelwerte und Streuungen ergeben:

Aufsatz Nr.:	1	2	3	4	5
Mittelwert (M)	2.08	2.69	3.51	4.21	5.04
Standardabweichung (s)	.67	.91	.90	.89	.80

5.3.4.2.2. Beurteilungsstichprobe und Bewertungsablauf

Als Beurteiler wurden 100 Referendare für das Lehramt an Volks- und Realschulen und 100 Lehrer dieser Schularten mit längerer Berufserfahrung herangezogen, und zwar jeweils die gleiche Anzahl Frauen und Männer.

5.3.4.3. Ergebnisse

5.3.4.3.1. Der Einfluß der schulpraktischen Erfahrung

Bei allen fünf Aufsätzen erwies sich die Zugehörigkeit zur Gruppe der Referendare oder der Lehrer als sehr bedeutsam für die Urteilshöhe. Dies wird bereits bei der Gegenüberstellung der Notenmittelwerte beider Beur-

Abb. 11: Bildvorlage für die Aufsätze der dritten Untersuchung (vgl. BELLAK u. BELLAK 1949)

teilungsgruppen deutlich (vgl. Tab. 12) und auch durch die statistische Analyse (t-Tests) bestätigt. Die entsprechenden t-Werte sind sämtlich negativ (vgl. Tab. 13), das bedeutet, daß die Lehrer mit Berufserfahrung die vorgegebenen Aufsätze durchgehend strenger beurteilten als die Referendare.

Tabelle 12

Notenmittelwerte und Mittelwertsdifferenzen der beiden Stichproben

	Referendare/Lehrer		
Aufsatz Nr.	Referendare	Lehrer	D_M
1	1.86	2.14	.28
2	2.34	2.77	.43
3	2.84	3.12	.28
4	3.69	4.11	.42
5	4.15	4.62	.47

Tabelle 13

t-Werte und Signifikanzniveaus beim Vergleich Referendare/Lehrer (df = 198)

Aufsatz	1	2	3	4	5
t	-2.80	-5.03	-2.76	-4.28	-5.46
p <	.01	.01	.01	.01	.01

5.3.4.3.2. Der Einfluß verschiedener Sprachkriterien

Die quantitativ meßbaren Variablen Aufsatzlänge, Rechtschreibung, Grammatik und Schriftgüte wurden durch Auswahl, Berichtigung und Darbietung in Maschinenschrift bei den fünf Aufsätzen weitgehend konstant gehalten. Notendifferenzen konnten daher — abgesehen von individuellen Bezugssystemen und induzierten Leistungserwartungen — nur aufgrund von Güteunterschieden in qualitativen Sprachvariablen auftreten. Im folgenden wurde daher der Zusammenhang zwischen den Zensuren und drei zunächst als wesentlich angenommenen inhaltlichen bzw. stilistischen Kriterien untersucht, und zwar

a) Originalität der Einfälle (O)
b) Differenziertheit des sprachlichen Ausdrucks (Stil = S)
c) Flüssigkeit und Abgeschlossenheit des Handlungsablaufes (H)

Zwischen den Gesamtnoten (N) und den Bewertungen in den drei Sprachkriterien (a, b, c) ergaben sich folgende Korrelationen:

$$r_{NO} = .75; r_{NS} = .76; r_{NH} = .79$$

Die drei Variablen hatten also bei der Zensurenbildung annähernd die gleiche Bedeutung.

Die Kriterien untereinander korrelierten mit:

$$r_{OS} = .61;\ r_{OH} = .63;\ r_{SH} = .70$$

Die Berechnung der Partialkorrelationen (vgl. HOFSTÄTTER & WENDT, 1966) ergab folgende Werte:

$$r_{OS.H} = .31;\ r_{OH.S} = .35;\ r_{HS.O} = .52$$

Der unterschiedliche Zusammenhang zwischen den drei Sprachkriterien ist verständlich, wenn man sich deutlich macht, daß Differenziertheit des sprachlichen Ausdrucks und Flüssigkeit sowie Abgeschlossenheit des Handlungsablaufs sich auf die sprachliche Gewandtheit des Aufsatzschreibers beziehen ($r_{HS.O} = .52$), während die Originalität der Einfälle ein Maß der gedanklichen Kreativität darstellen ($r_{OS.H} = .31$; $r_{OH.S} = .35$).

Mit Hilfe der Berechnung der Beta-Gewichte über das DOOLITTLE-Verfahren (vgl. LIENERT 1961, S. 396 ff.) wurde die multiple Korrelation zwischen den erteilten Noten und den Beurteilungen bezüglich der drei Kriterien ermittelt, sie beträgt $R_{N(OSH)} = .88$. Dies entspricht einem Schätzungseffekt von etwa 52 % (vgl. HOFSTÄTTER & WENDT 1966), d. h. eine Voraussage der Aufsatzzensuren ist schon allein aufgrund der Erfüllung dieser drei qualitativen Sprachkriterien in recht hohem Maße möglich. Das gilt zumindest für Niederschriften jener Art, wie sie dieser Untersuchung zugrundelagen.

5.3.5. Diskussion der Ergebnisse

5.3.5.1. Allgemeine Befunde zur experimentellen Situation

In den drei Untersuchungen wurden ganz verschiedene Gruppen von Vpn als Beurteiler eingesetzt, und zwar Schüler (Unters. 2), Studierende der Erziehungswissenschaft und Psychologie (Unters. 1 u. 2), Referendare (Unters. 3) und Lehrer (Unters. 2 u. 3). Die Beurteilung selbst erfolgte in allen Fällen in einer weitgehend standardisierten Situation unter quasi-experimentellen Bedingungen mit isolierender Variation einzelner Variablen. Auch die als Reizmaterial zur Beurteilung vorgegebenen Niederschriften waren unter einheitlichen Bedingungen, ausgelöst durch einen Standardreiz (CAT-Bild), gewonnen worden. Vor einer Erörterung und Interpretation der einzelnen Untersuchungsergebnisse stellt sich daher die Frage, ob und inwieweit ein solches Verfahren sich bewährt hat und als geeignet erscheinen kann, die hier zur Überprüfung vorgelegten Fragen in angemessener Weise zu beantworten; insbesondere auch im Hinblick darauf, daß damit tatsächlich Bedingungsvariablen ermittelt werden, die bei der Beurteilung

von Aufsätzen wirksam sind und in der konkreten Situation eine bedeutsame Rolle spielen.

Zur Verwendung der einzelnen Beurteilergruppen ist zunächst festzustellen, daß sie bei einem unterschiedlichen Bewertungsmodus eine recht gute intraindividuelle Übereinstimmung zeigten. Sie entwickelten also jeweils ein Bezugssystem, das zumindest über die kurze Zeitspanne des Experiments relativ stabil war und ihnen eine differenzierende Bewertung der Aufsätze gestattete. Diese ließ zugleich auch eine gute Abstufung derart erkennen, daß sich bei allen Vpn die Zensuren angenähert normal verteilten (vgl. Unters. 1).

Andererseits ergab sich zwischen den Vpn eine weite Streuung der Urteile, die in den einzelnen Gruppen unterschiedlich groß war und von außerordentlich gering bis hoch reichte, ohne daß sich eine systematische Beeinflussung erkennen ließ. Das bestätigt erneut die bisherigen Befunde über die unzureichende Übereinstimmung verschiedener Beurteiler bei der Bewertung desselben Aufsatzes (vgl. JÖRG 1964, KÖTTER & GRAU 1965). Dabei zeigte sich bei allen beurteilten Aufsätzen keine bedeutsame Beziehung zwischen der Güte der Urteile und ihrer Streuung, lediglich die am besten zensierten Aufsätze wurden signifikant einheitlicher bewertet (vgl. Unters. 1). Alter und didaktisches Wahlfach der Vpn sowie Lese- und Beurteilungszeit blieben dagegen ohne nachweisbaren Einfluß auf die Bewertung. Auch durch das Geschlecht der Beurteiler erfuhr sie im Unterschied zu den Befunden von KÖTTER & GRAU (1965, S. 292 u. 295 ff.) keine systematische Differenzierung (übereinstimmende Ergebnisse aller drei Untersuchungen).

Zur Frage der Bewährung des experimentellen Ansatzes für eine Analyse der Bedingungsvariablen von Aufsatzbeurteilungen läßt sich feststellen, daß sich der experimentelle Ansatz auch ohne „Ernstsituation" für die Beurteiler als hinreichend praktikabel erwies. Durch die experimentellen Bedingungen konnten ferner die Urteile systematisch beeinflußt werden, indem bestimmte Erwartungshaltungen kurzzeitig induziert wurden. Vermutlich liegen ähnliche Einflüsse auch bei der Bewertung von Aufsätzen in der schulischen Situation vor, insbesondere da in ihr mit stabileren Erwartungshaltungen zu rechnen ist. Zur systematischen Aufdeckung solcher Einflüsse erscheint ein experimenteller Ansatz wie der vorliegende besonders geeignet. Allerdings sollten die in den vorliegenden Befunden sich andeutenden differentiellen Einflüsse von Stärke und Richtung jener Diskrepanzen zwischen erwarteter und tatsächlicher Leistung noch weiter überprüft werden.

Die wichtigsten Ergebnisse der drei Untersuchungen sollen nun im folgenden zunächst nacheinander im Hinblick auf die jeweiligen Fragestellungen diskutiert werden. Dabei wird jedoch gleichzeitig auf die sich ergebenden Zusammenhänge sowie auf einander ergänzende Befunde Bezug genommen.

5.3.5.2. Diskussion der ersten Untersuchung

5.3.5.2.1. Zur Beeinflussung der Beurteilung durch Vergleichsserien

Die Einbettung in verschiedene Bezugsreize erwies sich für die Beurteilung der Aufsätze nicht als bedeutsam; sowohl die Varianz- wie Profilanalysen ließen keinen signifikanten Einfluß der drei Vergleichsserien auf die Bewertung erkennen. Für die Ausbildung individueller Bezugssysteme bei der Beurteilung scheint es daher kaum von Bedeutung zu sein, welche anderen Niederschriften gleichzeitig zu zensieren sind, zumindest solange bei diesen jeweils eine ungefähr gleiche Variationsbreite und Abstufung besteht, wie das im vorliegenden Experiment der Fall war. Eine andere Frage ist, ob sich etwa dann ein Einfluß der Vergleichsreize nachweisen ließe, wenn die einzelnen Serien Aufsätze recht unterschiedlicher Güte erhielten. Das wäre nach den bisherigen Ergebnissen über die Relativität von Zensuren durchaus zu erwarten und sollte in weiteren experimentellen Variationen überprüft werden.

5.3.5.2.2. Zum Einfluß unterschiedlicher Information über die Ausgangssituation

Die Art der den Vpn erteilten Information über die Ausgangssituation der Schüler bei Niederschrift der Aufsätze wirkte im Unterschied zur Einbettung in verschiedene Vergleichsreize in bedeutsamer Weise als ein Faktor, der die Bewertung systematisch beeinflußte. Darüber hinaus ergaben sich auch signifikante Wechselwirkungen mit den Vergleichsserien. Das bedeutet, daß die Aufsätze dann, wenn sie in eine bestimmte Serie eingebettet waren und die Vpn zusätzlich eine bestimmte Art von Information erhielten, teilweise anders beurteilt wurden, als es unter der Wirkung eines dieser beiden Faktoren allein der Fall gewesen wäre. Insgesamt zeichnete sich der Trend ab, daß beim Fehlen jeglicher Information (4) oder bei bloßer Beschreibung des Bildes (2) die Urteile der Vpn für alle Serien größere Einheitlichkeit aufwiesen als bei Darbietung des Bildes im Diapositiv allein (1) oder zusätzlich zur Beschreibung (3). Eine Erklärung dafür könnte die Annahme liefern, daß durch die Darbietung des Bildes bei den Beurteilern recht verschiedene Eindruckserlebnisse hervorgerufen wurden. Infolgedessen hatten sich bei ihnen offenbar auch wesentlich unterschiedlichere Erwartungshaltungen über das gebildet, was die Aufsätze beinhalten sollten, als bei jenen Beurteilern, die lediglich eine kurze verbale Beschreibung des Bildes oder überhaupt keine Information bekommen hatten.

Gerade die vierte Gruppe (ohne Information) meinte zunächst, die Aufsätze nicht beurteilen zu können, wenn sie keine weiteren Angaben über die

Ausgangssituation der Schüler erhielte. Demgegenüber scheint es aber so zu sein, daß eine sehr detaillierte Kenntnis dieser Situation die Uneinheitlichkeit der Beurteilung eher systematisch erhöht. Inwieweit diese Befunde auch für andere Aufsatzgattungen als die vorliegenden Niederschriften im Anschluß an eine Bilddarbietung zutreffen, muß offen bleiben. Weitere Untersuchungen dazu wären erforderlich.

In der Schulwirklichkeit ist dem zensierenden Lehrer die besondere Aufgabenstellung und Ausgangssituation immer voll bekannt. Das könnte eine systematische Beeinflussung seines Bezugssystems im Sinne ganz bestimmter Erwartungshaltungen zur Folge haben, die möglicherweise erheblich von dem eines anderen Lehrers abweichen. Es wäre daher zu überlegen, ob nicht bei der Beurteilung von Aufsätzen, zumindest dieser Gattung, auch in der Schulpraxis eine Beurteilung durch einen unbeteiligten Lehrer entsprechend den Informationsgruppen 2 oder 4 eine objektivere Bewertung ermöglicht. Diese Frage sollte einer systematischen Überprüfung wert sein.

5.3.5.2.3. Zum Einfluß verschiedener Sprachvariablen

Eine Überprüfung der Beziehungen zwischen der Beurteilungshöhe und verschiedenen Sprachvariablen ließ einige bedeutsame Zusammenhänge erkennen. Am höchsten korrelierten dabei jene Merkmale mit der Güte der Beurteilung, die sich auf die Länge der Darstellung bezogen (Anzahl der Wörter, Wörter pro Satz, Anzahl der Sätze). Dieselbe Beziehung hatten KÖTTER & GRAU (1965, S. 295 ff.) bei der Beurteilung von Nacherzählungen gefunden. Mit der Zahl der orthographischen Fehler ergab sich eine hochsignifikante negative Korrelation. Eine Faktorenanalyse der Sprachmerkmale und eine Überprüfung des Zusammenhangs zwischen den grobgewichteten Merkmalsscores und der Bewertung der Aufsätze bestätigten diese Befunde. Von insgesamt sechs extrahierten Sprachfaktoren erwies sich der Faktor, der die Länge der sprachlichen Produktion repräsentierte, am weitaus bedeutsamsten für die Güte der Beurteilung. Mit erheblichem Abstand folgten der Faktor „fehlerhafte Darstellung" und der Faktor „attributiver Stil", der vor allem durch den Gebrauch von Adjektiven in attributiver Form sowie auch von zusammengesetzten Verben konstituiert wurde.

Für die Beurteilung durch den jeweiligen Klassenlehrer besaß der Faktor „fehlerhafte Darstellung" das größte Gewicht; daß er bei den Vpn erst an zweiter Stelle für die Bewertung in Betracht kam, ist sicher auf die vorgenommene Fehlerkorrektur zurückzuführen. Gerade deshalb überrascht es besonders, daß die Fehlerzahl überhaupt in diesem Ausmaß in die Beurteilung eingehen konnte. Das läßt sich sicher auch nicht damit hinreichend erklären, daß bei der Berichtigung einige Versehen unterliefen, zumal die Vpn entsprechend instruiert worden waren. Es drängt sich vielmehr die Vermu-

tung auf, daß die fehlerhafte Darstellung mit anderen Merkmalen der Niederschrift kovariiert. Durch die Interkorrelation der grobgewichteten Merkmalsscores konnte diese Annahme bestätigt werden. Insbesondere ergab sich eine signifikante negative Beziehung zwischen dem Faktor „fehlerhafte Darstellung" und dem Faktor „attributiver Stil", der sich ebenfalls für eine Differenzierung des Güteurteils sowohl der Vpn wie der Klassenlehrer als bedeutsam erwiesen hatte.

Die Aufsätze von Schülern, die Schwierigkeiten mit der Rechtschreibung und dem richtigen Kasusgebrauch haben, scheinen also auch nach anderen Aspekten der sprachlichen Darstellung von geringerer Qualität zu sein. Dadurch wird einerseits die Annahme von Schulpraktikern bestätigt, daß eine Mitbewertung der Rechtschreibung bei der Aufsatzbeurteilung auch deshalb gerechtfertigt sei, weil sie in enger Beziehung zum schriftlichen Ausdruck allgemein stehe (vgl. AHRENS 1964). Zum anderen ergibt sich allerdings die Gefahr einer Überbewertung der orthographischen und grammatischen Fehler, wie sie auch in dem sehr hohen Gewicht zum Ausdruck kommt, das dieser Faktor für die Beurteilung der Aufsätze durch den Klassenlehrer besaß. Hier scheint die Warnung vor einer Doppelbewertung der Rechtschreibung bei der Aufsatzbeurteilung, wie sie von einigen Autoren geäußert wurde (vgl. AHRENS 1964, SCHRÖTER 1965) durchaus gerechtfertigt.

5.3.5.3. Diskussion der zweiten Untersuchung

5.3.5.3.1. Zum Einfluß unterschiedlicher Informationen über die Leistungen der Schüler

Die Benotung einer Serie von je zwei guten, mittleren und schlechten Aufsätzen, deren Güte in einer Vorerhebung ermittelt worden war, konnte durch Hinweise auf das allgemeine Leistungsverhalten der Aufsatzschreiber in bedeutsamer Weise beeinflußt werden.

Bei einer in sich widerspruchsfreien positiven Information über den Schüler wurde dessen Aufsatz besser beurteilt als bei einem negativen Hinweis. Hier zeigte sich also ein in der Tendenz ähnlicher Einfluß wie bei der Information über die Ausgangssituation bei der Niederschrift der Aufsätze in der ersten Untersuchung (vgl. Abschn. 5.3.5.2.2.). In beiden Fällen wurden durch die Information offensichtlich in den Beurteilern bestimmte Erwartungshaltungen ausgelöst.

Die in der zweiten Untersuchung vorgegebene quantitative Abstufung des Reizmaterials bewirkte dabei allerdings, daß die Beurteiler unter den verschiedenen Informationsbedingungen ihr Urteil nur jeweils innerhalb relativ fester Toleranzgrenzen modifizierten.

Die Vorgabe unterschiedlicher Information über die Aufsatzschreiber führte nicht zu durchgehend signifikanten systematischen Differenzierungen des Urteils. In der Tendenz bestand jedoch bei widersprüchlicher Information eine Beeinflussung derart, daß bei den zwei guten Aufsätzen das Urteil strenger, bei den zwei mittleren und schlechten milder ausfiel.

5.3.5.3.2. Die Beurteilung durch Schüler, Studenten und Junglehrer

Systematische Einflüsse der Stichprobenzugehörigkeit wirkten sich insbesondere darin aus, daß Junglehrer mit einer gewissen Beurteilungserfahrung eher dazu neigten, die Schüleraufsätze insgesamt strenger zu beurteilen als die unerfahrenen Beurteiler (Studenten und Schüler). Junglehrer reagierten ferner auf die Leistungshinweise bei 3 von 6 Schüleraufsätzen signifikant abweichend vom oben dargestellten Haupteffekt: So beurteilten sie durchschnittliche Aufsätze bei Vorgabe der Information „Schüler mit mangelhaften Leistungen" milder als bei einem adäquaten Hinweis und die schlechten Aufsätze besser bei einem „befriedigenden" Hinweis als bei adäquater oder guter Leistungsinformation. Unter dem Eindruck der Information „guter Schüler" war ihre Benotung der durchschnittlichen Aufsätze besser, die der schlechten Aufsätze niedriger.

Hier stellt sich zugleich auch die Frage nach der praktischen Relevanz dieser Ergebnisse. Nach den Befunden früherer Arbeiten legen Lehrer bei der Bewertung von Schüleraufsätzen interindividuell unterschiedliche Bezugssysteme an (Kötter & Grau 1965). Das wurde auch in der ersten Untersuchung deutlich bestätigt. Vermutlich bilden sich solche Bezugssysteme aufgrund bestimmter Erwartungshaltungen aus. So konnte Weiss (1965) den Einfluß von Leistungserwartungen aufgrund der Zugehörigkeit der Aufsatzschreiber zu unterschiedlichen soziokulturellen Schichten auf die Benotung von Aufsätzen kontrollieren. Danach werden die weniger privilegierten Schüler bei der Bewertung benachteiligt. Auch aufgrund unserer Ergebnisse muß man mit Einflüssen der Leistungserwartung auf das Aufsatzurteil rechnen. Indessen scheinen sich in den von uns erfaßten Urteilen komplexere Beurteilungsvorgänge abzuzeichnen.

Vermutlich spielten bei der Bewertung der Aufsätze durch die erfahrenen Beurteiler (Junglehrer) im Gegensatz zu den unerfahrenen (Schüler und Studenten) das Ausmaß und die Richtung der Abweichung zwischen der durch die Information aufgebauten Leistungserwartungen und der objektiv erfaßten Leistung in den Aufsätzen eine Rolle und ließen die zu beobachtenden differentiellen Urteilstendenzen zustande kommen: Eine stärkere negative Diskrepanz (guter Schüler/schlechter Aufsatz) bewirkte eine Verschärfung des Urteils, eine mittlere (befriedigender Schüler/schlechter Aufsatz) dagegen eine Milderung. Starke positive Abweichungen (schlechter

Schüler, guter Aufsatz) verschärften ebenfalls das Urteil, wogegen die mittlere Abweichung bei den durchschnittlichen Aufsätzen das Lehrerurteil milderte, eine mittlere negative Diskrepanz (guter Schüler/durchschnittlicher Aufsatz) jedoch stärker als die entsprechende negative Abweichung (schlechter Schüler/durchschnittlicher Aufsatz).

Vermutlich liegen solchen Urteilstendenzen sowohl bewußte pädagogische Maßnahmen zugrunde, etwa wenn ein schlechter Aufsatz bei einem guten Schüler weniger toleriert wird als bei einem Schüler mit befriedigenden Leistungen, als auch nichtbewußte Einflüsse, so vielleicht dann, wenn der gute Aufsatz eines sonst schlechten Schülers eine geringere Chance hat, ebenso günstig beurteilt zu werden wie der gleichwertige Aufsatz eines guten Schülers.

5.3.5.3.3. Geschlechtsspezifische Einflüsse

Das Geschlecht der Beurteiler war auch für die Ergebnisse dieses zweiten Experiments ohne Einfluß. Das bestätigt die gleichsinnigen Befunde der ersten Untersuchung. Auch das Geschlecht der Aufsatzschreiber trug nicht zu einer systematischen Differenzierung der Urteile bei.

5.3.5.4. Diskussion der dritten Untersuchung

5.3.5.4.1. Zum Einfluß unterschiedlicher Information über das Leistungsverhalten der Schüler

Wie schon in der vorangegangenen Untersuchung, so konnten auch in diesem Experiment durch unterschiedliche Informationen über das allgemeine Leistungsverhalten der Schüler bei den Vpn gewisse Erwartungshaltungen hervorgerufen werden, die die Beurteilung der Aufsätze bedeutsam beeinflußten. In der Tendenz traten solche Einflüsse bei Referendaren in stärkerem Maße auf als bei Lehrern mit längerer Schulpraxis. Damit wurden die entsprechenden Ergebnisse der zweiten Untersuchung vollauf bestätigt.

5.3.5.4.2. Zur unterschiedlichen Beurteilung durch Referendare und Lehrer

Bei der Benotung einer Serie von fünf Aufsätzen wurden von unterrichtserfahrenen Lehrern generell strengere Maßstäbe angelegt als von Referendaren. Damit bestätigten sich die in der zweiten Untersuchung gewonnenen Ergebnisse für Schüler/Studenten einerseits und Junglehrer mit Beurteilungspraxis andererseits (vgl. Abschn. 5.3.5.3.2.). Im Laufe ihrer Schulpraxis haben Lehrer sicherlich größere Erfahrungen sammeln können bezüglich dessen, was 10jährige Schüler zu leisten in der Lage sind.

5.3.5.4.3. Zum Einfluß verschiedener Sprachkriterien

Die Einschätzung der Aufsätze nach den Sprachkriterien „Originalität der Einfälle", „Differenziertheit des sprachlichen Ausdrucks (Stil)" und „Flüssigkeit sowie Abgeschlossenheit des Handlungsablaufes" war für die Benotung in annähernd gleichem Maße bedeutsam. Diese drei Variablen scheinen neben den schon in der ersten Untersuchung analysierten Merkmalen wesentlich zur Entstehung einer Aufsatzzensur beizutragen. Das dürfte zumindest für Aufsätze mit der hier vorliegenden Aufgabenstellung zutreffen. Dabei kommt den genannten Kriterien besonders dann ein hoher Prädiktorwert zu, wenn andere Differenzierungsmerkmale, wie z. B. die Länge des Aufsatzes, ausfallen.

5.3.6. Zusammenfassung und Schlußfolgerungen

Die vorliegende experimentelle Untersuchungsreihe über die Wirkung verschiedener Einflüsse auf die Beurteilung von Schüleraufsätzen konnte mehrere bedeutsame Bedingungsvariablen isolieren, die vermutlich auch in der schulischen Beurteilungspraxis eine wichtige Rolle spielen und wesentlich zu der weithin beklagten Diskrepanz in der Bewertung von Aufsätzen durch verschiedene Lehrer beitragen dürften. Die Kenntnis dieser Einflüsse und ihres Ausmaßes stellt aber eine notwendige Voraussetzung dafür dar, Möglichkeiten für eine einheitlichere und damit gerechtere Aufsatzbewertung zu erarbeiten.

Keinerlei Einfluß auf die Art der Bewertung konnte für zwei Faktorengruppen nachgewiesen werden, für das Geschlecht und für die Einbettung in Vergleichsserien. Für die Bewertung eines Aufsatzes ist es ohne Bedeutung, ob er von einem Mädchen oder von einem Jungen geschrieben wurde, und ebenso spielt es keine Rolle, ob ihn ein weiblicher oder männlicher Beurteiler zensiert. Angesichts einer auch heute noch durch die meisten Erziehungssituationen unterstützten geschlechtsspezifischen Rollendifferenzierung gerade im Schulkindalter (vgl. Mussen u. a. 1969, Nickel 1974) erscheint das keineswegs als banal und selbstverständlich.

Ebenso dürfte es nach den vorliegenden Ergebnissen für die Ausbildung individueller Bezugssysteme bei der Beurteilung von Aufsätzen ohne wesentliche Bedeutung sein, welche anderen Arbeiten dieser Art gleichzeitig beurteilt werden; das gilt zumindest so lange, als diese Arbeiten eine ähnliche Variationsbreite aufweisen.

Als eine erste wichtige Gruppe von Bedingungsvariablen, die zu unterschiedlicher Bewertung führten, konnten verschiedene Informationen sowohl über die Ausgangssituation der Schüler bei der Niederschrift als insbesondere auch über das allgemeine Leistungsverhalten im Unterricht ermittelt

werden. In allen drei Untersuchungen wurden damit Erwartungshaltungen bei den Beurteilern ausgelöst, die auch sonst bei der Schülerbeurteilung einen gewichtigen Fehlerfaktor darstellen (vgl. ROSENTHAL & JACOBSON 1971, ERLEMEIER & TISMER 1973).

Eine weitere Gruppe von Bedingungsvariablen betrifft das Ausmaß an Lehr- und Beurteilungserfahrungen der Bewerter. Generell zeigt sich dabei die Tendenz, daß sich die Strenge des Urteils mit zunehmender schulpraktischer Erfahrung erhöht. So urteilen Junglehrer strenger als Schüler (Obersekundaner) und Studierende (Erziehungswissenschaft) und schulerfahrene Lehrer wiederum strenger als Referendare.

Während es sich bei diesen beiden Variablengruppen um Bedingungsfaktoren handelt, die wohl allgemein bei jeder schulischen Leistungsbeurteilung wirksam sein dürften, können die im weiteren ermittelten Sprachkriterien speziell für die Aufsatzbewertung als charakteristisch gelten.

In der ersten Untersuchung konnten sechs Sprachfaktoren ermittelt werden, die insgesamt 75 % der Beurteilungsvarianz erklären. Inhaltlich ergaben sich dabei in verschiedenen Variablengruppen deutliche Übereinstimmungen mit den von DIEDERICH u. a. (1961) an 53 Beurteilern verschiedener beruflicher Stellung ermittelten fünf Faktoren der Aufsatzbewertung. Das gilt insbesondere für den Faktor „Sprachrichtigkeit" bzw. „fehlerhafte Darstellung". Zugleich erweiterten und differenzierten sie jedoch auch jene Befunde. Interessant ist ferner, daß die Lehrer unter den Beurteilern von DIEDERICH sich ebenso wie in unserer Untersuchung bei der Bewertung eines Aufsatzes in erster Linie an dessen sprachlicher Richtigkeit bzw. Fehlerhaftigkeit orientierten, während Autoren und Verleger ihre Beurteilung vor allem auf den Faktor Lebendigkeit (Originalität, Anschaulichkeit) stützten. Die in der vorliegenden ersten Untersuchung eingesetzten Studierenden der Erziehungswissenschaft berücksichtigten bei ihrem Urteil dagegen in erster Linie jene Variablen, die für die Länge der sprachlichen Produktion kennzeichnend sind.

Eine Analyse des Einflusses verschiedener Sprachkriterien bei annähernder Konstanthaltung des Faktors „Länge" und „Sprachrichtigkeit" ergab, daß die drei Variablen „Originalität der Einfälle", „Differenziertheit des sprachlichen Ausdrucks" und „Flüssigkeit bzw. Abgeschlossenheit des Handlungsablaufs" allein für die Gesamtzensur bereits einen Schätzungseffekt von 52 % besitzen. Wenn also die Richtigkeit und die Länge der Darstellung als Grundlage für eine Beurteilungsdifferenzierung entfallen, läßt sich die Aufsatzzensur in beträchtlichem Umfang bereits allein aufgrund dieser drei Sprachkriterien voraussagen.

Diese Ergebnisse dürften insgesamt die Bemühungen um die Erarbeitung einheitlicher und zuverlässiger Bewertungskriterien mit dem Ziel einer größeren Objektivität und Reliabilität der Aufsatzbeurteilung wesentlich unter-

stützen. Während solche Kriterien sich bisher fast nur auf intuitiv erarbeitete Faktoren stützten und daher unter mangelnder Eindeutigkeit und Einheitlichkeit litten (vgl. WEBER 1969), wurden in dieser Versuchsreihe erstmals Ansätze für eine empirische Erarbeitung solcher Kriterien unternommen.

Abschließend bleibt noch die Frage zu klären, ob die Ergebnisse der vorliegenden Untersuchungen denn auf den Schüleraufsatz schlechthin generalisiert werden können, da es sich doch bei den verwendeten Arbeiten um eine spezielle Form der Niederschrift handelte. Skeptisch könnte in dieser Hinsicht etwa auch die Feststellung von GOSLING (1966) stimmen, daß sich bei verschiedenen Aufsatztypen unterschiedlich hohe Übereinstimmungen der Beurteiler ergaben. Allerdings überprüfte GOSLING allein die Beurteilungsobjektivität ohne Analyse der verschiedenen wirksamen Bedingungsvariablen. Gerade bezüglich dieser Variablen ergaben sich jedoch in den vorliegenden Experimenten verschiedene Hinweise darauf, daß es sich weniger um eng auf die besondere Aufsatzform bezogene Merkmale als eher um allgemeinere Kriterien handelt.

Das wird besonders deutlich bei dem Einfluß einer unterschiedlichen Information der Beurteiler über die Ausgangssituation und das Leistungsverhalten der Schüler und die dadurch ausgelösten Erwartungshaltungen, die auch in anderen Beurteilungssituationen nachweisbar sind. Eine weitgehende Generalisierbarkeit dürfte aber auch für die Sprachkriterien zutreffen, wie nicht zuletzt eine gewisse Übereinstimmung mit den von DIEDERICH u. a. (1961) sogar in einem anderen Sprachraum ermittelten Sprachfaktoren nahelegt.

Selbstverständlich können jedoch die drei hier dargestellten Untersuchungen keinen Anspruch auf Vollständigkeit erheben. Es ging hier ja vorwiegend darum, erste Schritte in Richtung auf eine experimentelle Aufklärung der bei einer Aufsatzbeurteilung wirksamen verschiedenen Bedingungsvariablen zu unternehmen. Daß der beschrittene Weg sich als gangbar und durchaus erfolgreich erwiesen hat, darf wohl als wichtigstes Ergebnis festgehalten werden. Weitere Untersuchungen dieser Art sind aber zweifellos dringend erforderlich und sollten unbedingt folgen.

5.3.7. Literaturverzeichnis

Ahrens, G.: Wie zensieren wir? Lebendige Schule, 1964, *19*, 368.

Bellak, L. u. *Bellak, S. S.:* Der Kinder-Apperzeptions-Test. Göttingen 1949.

Bobertag, O.: Leistungsschätzung und Leistungsmessung in der Volksschule. Ein Beitrag zur Frage: Was leistet unsere Schule? Ztschr. Päd. Psychol., 1931, *32*.

Diederich, P. B., French, J. W. u. *Carlton, S. T.:* Factors in Judgements of Writing Ability. Princeton 1961.

Erlemeier, N. u. *Tismer, K. G.:* Einstellungen und Erwartungen bei Lehrern und ihre Auswirkungen auf die Beurteilung und das Verhalten von Schülern. In: *Nickel, H.* u. *Langhorst, E.* (Hrsg.), 1973.

Falk, R.: Zur Psychologie der schulischen Leistungsbeurteilung durch Zensierung. Wiss. Ztschr. d. Univ. Halle, Ges.-Sprachwiss. Reihe, 1962, *11*, 1015—1032.

Glatz, G.: Kann man dem Schulzeugnis Glauben schenken? Lebendige Schule, 1967, *22*, 25—31.

Gosling, G. W. H.: Marking English Compositions. Austr. Council Educ. Research, 1966.

Hays, W. L.: Statistics for psychologists. New York 1963.

Hofstätter, P. u. *Wendt, D.:* Quantitative Methoden der Psychologie. Barth, München 1966.

Ingenkamp, K. H.: Die Fragwürdigkeit der Zensurengebung. Beltz Studienbuch, 3. Aufl., Beltz, Weinheim 1972.

Jörg, H.: Zum Problem der Schülerbeurteilung und der Zeugnisnoten. Lebendige Schule, 1964, *19*, 383—394.

Karnick, R.: Zeugnisse! — Zeugnisse? Lebendige Schule, 1959, *14*, 113—121.

Kötter, L. u. *Grau, U.:* Zur Bedingtheit der uneinheitlichen Benotung von Schüleraufsätzen (Nacherzählung). Ztschr. exp. u. angew. Psychol., 1965, *12*, 278—301.

Landsheere, de, G.: Ein Wendepunkt in der Geschichte der Bewertung. Vortrag auf dem Internationalen Zeugnis-Kongreß d. Ztschr. Schule, Düsseldorf 1973.

Laurien, H.-R.: Welche Verbesserungen am Notensystem sind kurzfristig möglich? Vortrag auf dem Internationalen Zeugniskongreß der Ztschr. Schule, Düsseldorf 1973.

Lienert, G. A.: Testaufbau und Testanalyse. Beltz, Weinheim 1961.

Lienert, G. A.: Eine varianzanalytische Methode zum Nachweis der Gruppenspezifität von Testprofilen. Ztschr. exp. u. angew. Psychol., 1963, *10*, 333—345.

Mussen, P. H., Conger, J. J. u. *Kagan, J.:* Child Development and Personality. 3. Aufl., Harper u. Row, New York u. London 1969.

Nickel, H.: Entwicklungspsychologie des Kindes- und Jugendalters. Bd. 2, Huber, Bern 1974 (im Druck).

Nickel, H. u. *Langhorst, E.* (Hrsg.): Brennpunkte der Pädagogischen Psychologie. Huber u. Klett, Bern u. Stuttgart 1973.

Rosenthal, R. u. *Jacobson, L.:* Pygmalion im Unterricht. Beltz, Weinheim 1971.

Schröter, G.: Zensuren- und Zeugniserteilung als Problem für den Junglehrer. Lebendige Schule, 1965, *20*, 89—97.

Schröter, G.: Die ungerechte Aufsatzzensur. Kamps pädag. Taschenbücher Nr. 48, Kamp, Bochum 1971.

Schröter, G.: Neue Ergebnisse der Zensurenforschung. Vortrag auf dem Internationalen Zeugnis-Kongreß der Ztschr. Schule, Düsseldorf 1973.

Siegel, S.: Nonparametric Statistics for the Behavioral Sciences. McGraw Hill, New York, Toronto, London 1956.

Steinkamp, G.: Die Rolle des Volksschullehrers im schulischen Selektionsprozeß. Ergebnisse einer empirisch-soziologischen Untersuchung. Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik, 1967, (12), 302—324.

Ulshöfer, R.: Zur Beurteilung von Reifeprüfungsaufsätzen. Der Deutschunterricht, 1948/49, *1*, 84—102.

Valentine, C. W.: Psychology and its beating on education. London 1950.

Weber, A.: Das Problem der Aufsatzbeurteilung. Ludwig Auer, Donauwörth 1969.

Weiss, R.: Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. Schule und Psychologie, 1965, *12*, (9), 257—269.
Wieczerkowski, W. u. *Keßler, G.:* Über den Einfluß der Leistungserwartung auf die Bewertung von Schüleraufsätzen. Schule und Psychologie, 1970, *17*, 240—250.
Wieczerkowski, W., Nickel, H. u. *Rosenberg, L.:* Einige Bedingungen der unterschiedlichen Bewertung von Schüleraufsätzen. Psychol. Rundschau, 1968, *11*, 280—295.
Winer, B. J.: Statistical principles in experimental design. New York 1962.

5.4. Bemühungen um Vereinheitlichung der Aufsatzbeurteilung

Jürgen Wendeler

Daß Schüleraufsätze uneinheitlich beurteilt werden, ist oft nachgewiesen worden und inzwischen sicherlich weithin bekannt. Das Ausmaß der Uneinheitlichkeit ist tatsächlich oft erschreckend groß. Manche Experten haben deshalb vorgeschlagen, auf die Aufsatzbeurteilung, insbesondere in Form einer Zensur, überhaupt zu verzichten. Andere haben versucht, den Aufsatz durch objektiver auswertbare Prüfverfahren zu ersetzen. Wieder andere haben sich bemüht, durch Kriteriensysteme mit Auswertungsvorschriften eine einheitlichere Aufsatzbeurteilung zu erreichen. Über diese Bemühungen soll im folgenden berichtet werden.

5.4.1. Die mangelnde Übereinstimmung von Aufsatzbeurteilungen

Zur Verdeutlichung der Problematik seien zunächst einige der Untersuchungsbefunde referiert, die das Ausmaß der Uneinheitlichkeit aufzeigen. Auch in Deutschland lag hierzu bereits in den 20er und 30er Jahren eine Reihe von Veröffentlichungen vor (z. B. LIETZMANN 1927; KIESSLING 1929). Besonders bekannt wurde dann ULSHÖFERS Untersuchung aus dem Jahr 1949. In der Zeitschrift „Der Deutschunterricht" (Heft 6, S. 95 ff.) druckte er einen Reifeprüfungsaufsatz ab und forderte Deutschlehrer auf, ein Gutachten mit Zeugnisvorschlag einzusenden. Er erhielt 42 verwertbare Gutachten, d. h. solche mit eindeutig festgelegtem Zeugnisvorschlag von Lehrern mit Oberstufenerfahrung. Die Zensuren verteilten sich wie folgt:

sehr gut	1 mal
gut	5 mal
befriedigend	13 mal
ausreichend	10 mal
mangelhaft	11 mal
ungenügend	2 mal

„Dieses Ergebnis gibt zu denken" schreibt ULSHÖFER, und er weist nachdrücklich darauf hin, daß es sich in diesem Fall ja um keinen beliebigen Schulaufsatz handelt, sondern um einen Prüfungsaufsatz, dessen Beurteilung für den Schüler größte persönliche Konsequenzen hat. „Während die einen Begutachter den Aufsatzschreiber als wohl oder gar hervorragend geeignet zum Studium empfehlen, halten ihn andere, und zwar ebensoviele Fachleute für ungeeignet oder sogar völlig unfähig" (ULSHÖFER 1949 b, S. 85).

Das Faktum der uneinheitlichen Benotung von Deutschaufsätzen wurde in späteren Untersuchungen wiederholt bestätigt (z. B. LEHMANN 1951; KÖTTER & GRAU 1965; WIECZERKOWSKI et al. 1968) und mit Schrecken zur Kenntnis genommen (LEHMANN: „Es ist erschreckend. Man kann es nicht anders bezeichnen"). Kürzlich hat SCHRÖTER (1971) sehr umfangreiches und lehrreiches Material zu diesem Problem veröffentlicht.

SCHRÖTER hatte in einer eigenen Vorerhebung Schüleraufsätze aus allen Klassenstufen zu verschiedenen Themen gesammelt und 617 davon an Lehrer mit der Bitte um Beurteilung und Zensierung geschickt. Jeder Aufsatz wurde 10 bis 20 Lehrern vorgelegt. Insgesamt wurden 11 153 Urteile abgegeben; durchschnittlich waren es 18 Urteile je Aufsatz. Die Zensurenbreite war wie folgt (S. 111):

Zensurenbreite	Prozentuale Häufigkeit
Dieselbe Zensur	0 %
Zwei verschiedene Zensuren	4,9 %
Drei verschiedene Zensuren	40,4 %
Vier verschiedene Zensuren	43,4 %
Fünf verschiedene Zensuren	10,3 %
Sechs verschiedene Zensuren	1,0 %

Es kam also nie vor, daß alle Lehrer dieselbe Zensur gaben, dagegen streuten in etwa 50 % der Fälle die erteilten Noten über vier oder fünf Notenstufen. Noch deutlicher zeigt sich die Problematik, wenn die Zensuren „gut" und „sehr gut" als Zensuren „besonderen Lobs", „mangelhaft" und „ungenügend" als Zensuren „besonderen Tadels" zusammengefaßt werden. Mehr als ein Drittel aller Aufsätze erhielt von manchen Lehrern dieses Lob und gleichzeitig von anderen Lehrern diesen Tadel.

Die Uneinheitlichkeit der Aufsatzbewertung ist natürlich weder eine spezielle Eigenart deutscher Lehrer noch ein spezielles Problem von Aufsätzen im muttersprachlichen Unterricht. Einige der wichtigsten englischen und amerikanischen Arbeiten zu diesem Problem findet man in INGENKAMPS Sammelband „Die Fragwürdigkeit der Zensurengebung" (1971), dessen

drittes Kapitel „Subjektive Fehlerquellen der Zensurengebung" allen empfohlen sei, die sich für Fragen der Aufsatzbeurteilung interessieren. Dieses Kapitel enthält auch Auszüge aus dem in England sehr bekannt gewordenen Buch von HARTOG & RHODES „An Examination of Examinations" (1935), das der Öffentlichkeit die überaus unterschiedliche Beurteilung von traditionellen Prüfungsaufsätzen anhand umfangreicher Untersuchungen nachdrücklich vor Augen führte. Bei Prüfungszensuren in Geschichte zeigte sich eine ähnlich hohe Abweichung der Zensuren wie bei Zensuren im muttersprachlichen Fach.

Besonders aufschlußreich und für Deutschlehrer im Kollegenstreit eine gute Waffe ist eine Untersuchung von STARCH & ELLIOT, die ebenfalls in INGENKAMPS Sammelband enthalten ist. Sie beschäftigt sich mit der Verläßlichkeit von Zensuren bei Mathematikarbeiten. Obwohl bereits 1913 erschienen, ist das methodische Vorgehen ebenso wie bei neueren Untersuchungen, so daß die Ergebnisse als durchaus aktuell anzusehen sind. Nachdem die Autoren in einer früheren Erhebung die Ungleichheit der Bewertung von Englischarbeiten nachgewiesen hatten, wollten sie prüfen, ob sich die Situation bei einer exakten Wissenschaft wie der Mathematik nicht ganz anders darstelle. Denn bei der Mathematik, so wird auch heute oft behauptet, müßten die vielen subjektiven Faktoren entfallen, die in die Beurteilung sprachlicher Arbeiten eingehen. 128 Lehrer beurteilten eine Geometriearbeit, die ein Schüler als Abschlußarbeit geschrieben und bei der er sich von zehn Problemen acht zur Bearbeitung hatte auswählen dürfen. Es konnten maximal 100 Punkte gegeben werden. Die tatsächlich erteilten Punkte lagen zwischen 25 und 89 und verteilten sich breit über den gesamten Zwischenbereich. Die Streuung der Beurteilungen ist also außerordentlich groß und entgegen der üblichen Meinung keineswegs geringer als in sprachlichen Fächern.

Das Problem der Aufsatzbeurteilung ist kein Spezialproblem des Sprachunterrichts, sondern das Problem dieser Prüfungsform als solcher. Nun läßt sich die Aufsatzform allerdings in vielen Bereichen durch andersartige, objektiv auswertbare Formen ersetzen. Auch im muttersprachlichen Unterricht kann man zweifellos von solchen Verfahren weitaus mehr Gebrauch machen als es zur Zeit geschieht, nicht nur beim Anfangslesen, Rechtschreiben und in der Grammatik, sondern auch im Literaturunterricht und bei stilistischen Übungen. Wenn man jedoch daran festhält, daß die Schüler die freie schriftliche Darstellung üben und lernen sollen, und wenn dies sogar ein wesentliches Ziel gerade des Deutschunterrichts ist, dann stellt sich die Frage nach der „gerechten Aufsatzzensur" für den Deutschlehrer besonders dringend. So sind wohl auch die meisten Versuche zur Verbesserung der Aufsatzbeurteilung im Rahmen des muttersprachlichen Unterrichts durchgeführt worden.

Die Problematik der Aufsatzbeurteilung wird noch dadurch verschärft, daß die einzelnen Beurteiler in ihrer Einstufung der Leistung schwanken und zu verschiedenen Zeiten derselben Leistung verschiedene Noten geben. LEHMANN (1951, S. 36) weist auf eine frühe deutsche Untersuchung von DÖRING (1925) hin, die solche Schwankungen aufgezeigt hat. Die Arbeiten von HARTOG & RHODES (in: INGENKAMP 1971, S. 80) und von EELLS (in: INGENKAMP 1971, Seite 117 ff.) haben dasselbe ergeben. EELLS ließ aufsatzähnliches Material (Geographie- und Geschichtsfragen) von 61 Lehrern anhand einer 20- bzw. 10-Punkte-Skala einstufen und legte denselben Lehrern dasselbe Material nach 11 Wochen zur nochmaligen Beurteilung vor. Die Korrelationen zwischen den Beurteilungen lagen zwischen 0,25 und 0,51. Dazu schreibt EELLS: „Es ist unnötig festzustellen, daß so niedrige Zuverlässigkeitskoeffizienten wie diese kaum besser sind als bloßes Raten ... Die Fehlbarkeit des menschlichen Urteils — selbst wenn dieselbe Person dasselbe Material wiederholt beurteilt — wird eindrucksvoll demonstriert" (loc. cit., S. 122).

Derartige Schwankungen sind für die betroffenen Schüler unerträglich, sie sind es aber auch für die urteilenden Lehrer, die vermutlich selbst nicht selten spüren, wie unsicher ihre Urteile sind. Mit der Aufgabe, leistungsgerechte Urteile abzugeben, sind sie völlig überfordert, ganz abgesehen davon, daß sie zwischen einer solchen Art des Urteilens und ihrem pädagogischen Auftrag meist einen unaufhebbaren Widerspruch sehen. Gelänge es, adäquatere Beurteilungsverfahren zu entwickeln, die nicht nur „technisch" besser wären, sondern auch die prinzipiellen Unsicherheiten, Unzulänglichkeiten und Widersprüche besser erkennen ließen, so wäre dies sicher auch ein wichtiger Beitrag zur Psychohygiene des Lehrerberufs.

5.4.2. Leistungsermittlung und Zensierung

Die Unterschiedlichkeit der Aufsatzbeurteilung hat verschiedenartige Ursachen, von denen einige bei allen Prüfungsformen wirksam sind, nicht nur beim Aufsatz. Manche Probleme, die man in Verbindung mit der „ungerechten Aufsatzzensur" diskutiert hat, sind eigentlich allgemeine Zensierungsprobleme, die bei allen, auch den „objektivsten" Verfahren, zu unterschiedlichen Zensuren führen können. Diese allgemeinen Probleme sollen hier nicht erörtert werden, und so ist es zunächst wichtig, sie von den speziellen Fragen der Aufsatzbewertung zu trennen.

Die Zensurengebung kann man sich als einen zweiphasigen Vorgang vorstellen. Der erste Schritt ist die *Leistungsermittlung,* bei der man die Gesamtleistung oder bestimmte Teilleistungen nach einem Bewertungsschlüssel als „richtig", „falsch", „gut" oder „schlecht" beurteilt, meist mit Punktwerten versieht, und bei der man daraufhin einen Gesamtpunktwert be-

stimmt. So stellt man z. B. in einem Diktat die Fehler fest, gewichtet sie möglicherweise nach dem Schweregrad und errechnet die Gesamtzahl der Fehler. Der zweite Schritt ist die *Zensierung*, d. h. die Zuordnung einer Zensur zu der errechneten Punktzahl. Dabei kann man in sehr unterschiedlicher Weise vorgehen, sodaß verschiedene Lehrer für dieselbe Punktzahl völlig unterschiedliche Zensuren erteilen können. Offensichtlich handelt es sich hierbei um kein Problem, das spezifisch für den Aufsatz ist, vielmehr können solche Unterschiede auch bei völlig objektiv auswertbaren Prüfmethoden auftreten. So braucht an dieser Stelle auf das Zensierungsproblem nicht näher eingegangen zu werden. Nur an einem Beispiel sei verdeutlicht, welche Bedeutung es für die unterschiedliche Aufsatzbeurteilung haben kann.

Ulshöfer (1949) hat in seinem bereits erwähnten Bericht auch die Kommentare der urteilenden Lehrer wiedergegeben, und hierunter ist die Begründung des einzigen Lehrers, der mit „sehr gut" geurteilt hat und damit vom Durchschnittsurteil am stärksten abgewichen ist, besonders aufschlußreich:

„Ich habe das Aufsatzthema in meiner UI nebst drei anderen Themen gestellt ... Das Ergebnis war bescheiden ... Dann habe ich den von Ihnen wiedergegebenen Abiturientenaufsatz genau überprüft. Er steht weit über dem Niveau meiner Klasse."

Dieser Lehrer, der vielleicht die Schwächen des fraglichen Aufsatzes ebenso gesehen hat wie die anderen Lehrer, urteilt vorwiegend gruppenbezogen und kommt so zu einer guten Note. Die übrigen Lehrer urteilen vorwiegend lernzielbezogen und beurteilen den Aufsatz dementsprechend meist viel ungünstiger.

Das spezifische Problem der Aufsatzbewertung liegt aber, wie gesagt, nicht im Bereich der Zensierung, sondern in dem der Leistungsermittlung. Wie bei allen Leistungsmessungen ergeben sich dabei zwei *Hauptfragen:* (1) Nach welchen Kriterien soll die Leistung erfaßt werden? (2) Wie lassen sich diese Leistungskriterien möglichst objektiv ermitteln? Diese beiden Probleme sollen mit Blick auf den Aufsatz im folgenden besprochen werden.

5.4.3. Kriterien der Aufsatzbeurteilung

Schüleraufsätze, das ist allgemein bekannt, können nach den verschiedenartigsten Gesichtspunkten beurteilt werden, und die Fachleute sind sich nicht einig, welche Kriterien verwendet werden sollen und wie diese zu gewichten sind. Daß es keine allgemeinen Kriterien für alle Aufsatzarten geben kann, liegt auf der Hand. Aber auch beim gleichen Aufsatzthema werden die Schülerleistungen oft nach sehr unterschiedlichen Kriterien beurteilt. Sicher-

lich ist dies eine der entscheidenden Ursachen für die Abweichungen in der Leistungsbewertung, und doch ist diese Uneinheitlichkeit kaum verwunderlich, sondern eigentlich eine Selbstverständlichkeit.

Die Festlegung, nach welchen Kriterien eine Leistung beurteilt werden soll, ist immer subjektiv, es lassen sich höchstens einige zeitlich und regional begrenzte Übereinkünfte erzielen. Bei objektiv auswertbaren Tests, z. B. Tests mit Antwort-Auswahl-Aufgaben, liegen diese Kriterien aber spätestens zum Zeitpunkt der Aufgabenformulierung fest. Stellt man z. B. einen Geographietest mit derartigen Aufgaben zusammen, so entscheidet man die Frage, ob es mehr auf Fachwissen oder mehr auf das Verständnis von Zusammenhängen ankommen soll, durch die Formulierung der Aufgaben. Die Auswertung ist danach wieder völlig objektiv. Läßt man dagegen einen Aufsatz zu demselben Themenkreis schreiben, so verlagert sich die Subjektivität der Kriterienfestlegung in den Auswertungsvorgang selbst: Der erste Schritt der Aufsatzbeurteilung ist, ob bewußt oder unbewußt getan, die Festlegung, nach welchen Gesichtspunkten man überhaupt urteilen will.

Sicherlich gibt es mehr oder weniger vernünftige Leistungskriterien, und sicherlich wird nicht selten nach unvernünftigen geurteilt. Da die Lehrer sich zu einer Beurteilung genötigt sehen, eine sichere Tradition von Kriterien aber fehlt, ist es nicht verwunderlich, wenn sie oft nach recht speziellen Vorlieben vorgehen. Die vorhandenen Kriteriensysteme (vgl. WEBER 1969, S. 53 ff.) sind so uneinheitlich und meist so vage formuliert, daß man sich fragt, ob sie für die Praxis eine Hilfe sind. Zweifellos sind wesentlich genauere und umfangreichere Kriteriensysteme erforderlich, als es sie heute gibt. Diese Systeme müßten auf den jeweiligen Aufsatztyp und das jeweilige Stilalter abgestellt und von möglichst präzise definierten Lernzielen abgeleitet sein. Die „Flensburger Norm für die Aufsatzbeurteilung" (HAHN 1966) und BOHUSCHS „Neue Kriterien für die Aufsatzbewertung" (1972) sind Ansätze in dieser Richtung.

Eine interessante und aufschlußreiche Arbeit über Kriterien der Aufsatzbeurteilung haben DIEDERICH, FRENCH & CARLTON (1961) veröffentlicht. Sie wurde im Rahmen eines langjährigen Forschungsprogramms des Educational Testing Service in den USA durchgeführt. Eine der Hauptaufgaben dieses Instituts ist die Entwicklung von Testmethoden, die bei Aufnahmeprüfungen für die Colleges verwendet werden können, und in diesem Rahmen hatte man sich um Methoden zur Erfassung der sprachlichen Ausdrucksfähigkeit bemüht. Die Untersuchung von DIEDERICH, FRENCH & CARLTON sollte aufzeigen, nach welchen Kriterien Aufsätze beurteilt werden, und damit eine Grundlage für eine größere Vereinheitlichung der Beurteilungen und vielleicht einen Ansatzpunkt für spezielle Testentwicklungen bilden.

Die Autoren haben also nicht deduktiv ein Kategoriensystem entworfen, sondern induktiv tatsächlich verwendete Urteilsdimensionen zu finden ver-

sucht. Zu diesem Zweck haben sie 53 Lesern in verschiedener beruflicher Stellung, also nicht nur Lehrern, sondern auch Juristen, Verlegern usw. je 300 Aufsätze zur Beurteilung nach einem 9-Stufen-System übergeben, wobei den Beurteilern freigestellt war, nach welchen Gesichtspunkten sie dabei vorgehen wollten. Die verwendeten Aufsätze, die man bei uns als „Besinnungsaufsätze" bezeichnen würde, waren als Hausarbeit von Studenten im ersten Collegejahr geschrieben worden. Wie in anderen Untersuchungen zeigte sich auch hier die große Uneinheitlichkeit der Beurteilung, die allerdings nicht als bedauernswerter Fehler, sondern als notwendige Folge unterschiedlicher Urteilskriterien aufgefaßt wurde. Um diese Kriterien zu finden, wurde mit Hilfe von Korrelationskoeffizienten und einer darauf aufbauenden Faktorenanalyse versucht, Gruppen von relativ einheitlich urteilenden Lesern zu finden, in der Erwartung, damit zugleich eine Gruppierung nach verschiedenen Urteilsgrundlagen aufzufinden.

Die Faktorenanalyse ergab fünf Faktoren, d. h. fünf unterscheidbare Gruppen von Urteilern. Aufgrund der Kommentare, die zu den Beurteilungen gegeben waren, interpretierten die Autoren diese Faktoren wie folgt:

I *Ideen:* Relevanz, Klarheit, Umfang, Überzeugungskraft,
II *Innere Form:* Aufbau, Gliederung, logische Struktur,
III *Lebendigkeit:* Originalität, Anschaulichkeit,
IV *Sprachrichtigkeit:* Korrektheit von Rechtschreibung, Grammatik usw.,
V *Wortwahl:* Flüssigkeit und Treffsicherheit des Ausdrucks.

Wer deutsche Kriterienkataloge kennt, wird hierin natürlich manches Vertraute wiedererkennen. Die Faktoren I, II und IV finden sich mit unterschiedlichen Bezeichnungen in nahezu jeder Aufstellung. Vergleicht man z. B. mit der „Flensburger Norm", so entspräche dem Faktor I der Aspekt „Inhalt", dem Faktor II „Gliederung und Gedankenführung", dem Faktor IV „Sprachrichtigkeit", und auch für den Faktor V gäbe es im Aspekt „Gebrauch der lexischen Ausdrucksmittel" eine Entsprechung. Interessant ist, daß in der „Flensburger Norm" ebenso wie in den meisten anderen Katalogen der Faktor III als eine Hauptkategorie fehlt. Dieser Faktor wurde vor allem von Autoren und Verlegern der Beurteilung zugrunde gelegt, die darauf sahen, ob der Text für möglichst viele Leser interessant und lesenswert sein könnte. Schulaufsätze haben zuallermeist nur einen Leser, den Lehrer, der sie von vorn bis hinten liest, wie langweilig sie auch immer sind — und so mag es nicht verwunderlich sein, daß in den Kriterienkatalogen der Aspekt „Lebendigkeit" sonst höchstens an recht nebengeordneter Stelle auftritt.

Interessant ist auch der Befund, daß unter den Lesern, die vor allem nach dem Gesichtspunkt der „Sprachrichtigkeit" urteilen, besonders viele Lehrer waren. Vielleicht verleitet der Lehrerberuf zu dieser Perspektive, während

Autoren und Verleger gewohnt sind, solche Äußerlichkeiten, die in ihrer Branche von Korrektoren in Ordnung gebracht werden, bei der Beurteilung der Qualität eines Textes außer acht zu lassen.

Der Faktor IV, Sprachrichtigkeit, aber auch der Faktor V, Wortwahl, lassen sich im Prinzip durch objektive Tests erfassen, wofür es schon seit langem eine Reihe von Beispielen gibt (siehe u. a. Kap. 4.3. in diesem Band). Für die anderen Faktoren stehen aber derartige Tests nicht zur Verfügung, nicht einmal als Entwürfe, und es ist auch schwer vorstellbar, wie ein entsprechender Test beschaffen sein müßte. So dürfte die eigentliche Aufgabe und das besondere Problem der Aufsatzbeurteilung darin liegen, die Faktoren „Ideen", „Innere Form" und „Lebendigkeit" möglichst genau zu erfassen (vgl. Kap. 5.3. oben).

5.4.4. Erfassung der Urteilsobjektivität

Nach welchen Kriterien man auch immer eine Beurteilung vornimmt, und wie auch immer man diese Kriterien definiert — in den allermeisten Fällen wird ein subjektiver Ermessenspielraum übrigbleiben, wird die Leistungsermittlung also nicht völlig objektiv sein. Das ist anders als bei der Auswertung von Diktaten, Rechenarbeiten oder Tests, bei denen — sofern klare Auswertungsrichtlinien vorliegen — die Leistungsfeststellung prinzipiell 100%ig objektiv ist und nur durch grobe Flüchtigkeitsfehler gemindert wird. Es ist aber ähnlich wie bei vielen anderen Beurteilungen psychologischer, pädagogischer, sozialer und medizinischer Art, bei denen ebenfalls derselbe Sachverhalt von verschiedenen Beurteilern unterschiedlich eingestuft werden kann.

Mit „Objektivität" ist hier die „Auswertungsobjektivität" gemeint, wie sie z. B. Lienert (1967, S. 13 f.) für Tests im allgemeinen definiert hat. Sie betrifft „die numerische oder kategoriale Auswertung des registrierten Testverhaltens nach vorgegebenen Regeln". Diese Definition erfaßt sowohl die einfache Feststellung von Einzelkriterien (z. B. „Thema verfehlt", „Handlung hat einen Höhepunkt" usw.) als auch die numerische Einstufung von Texteigenschaften (z. B. „Menge und Brauchbarkeit der stofflichen Inhalte", „Folgerichtigkeit der Gedankenführung" usw.) und ist deshalb für alle Probleme der Aufsatzauswertung zutreffend. Die „Auswertungsobjektivität" erfaßt man im allgemeinen aufgrund der Übereinstimmung mehrerer, unabhängig voneinander arbeitender Auswerter: Je größer die Übereinstimmung ihrer Ergebnisse, desto höher die Objektivität der Auswertung.

Es gibt mehrere Verfahren, den Grad der Urteilsübereinstimmung numerisch zu erfassen, korrelationsstatistische und varianzanalytische. Am häufigsten wurden wohl Korrelationsmethoden verwendet: Man korreliert die Urteile jeden Auswerters mit denen jedes anderen und berechnet vielleicht

noch den Durchschnitt dieser Korrelationen. GOSLING (1966) hat an diesem Vorgehen eine gewisse Kritik geübt und eine Verbesserung vorgeschlagen, nach der es möglich ist, die Korrelation der Urteile eines Auswerters mit den „wahren Werten" (den Durchschnittsurteilen) zweckmäßiger zu schätzen. Hinsichtlich der Einzelheiten sei auf die Originalarbeit verwiesen.

Eine andere Kritik betrifft die Korrelationsmethode überhaupt. Bestimmte Unterschiede zwischen den Auswertern gehen nämlich bei der Berechnung von Korrelationskoeffizienten verloren. Wenn z. B. ein Urteiler stets einen Punkt weniger gibt als ein anderer, so hat die Korrelation trotzdem den Maximalwert 1, ebenso, als wenn beide Urteiler immer genau denselben Wert gegeben hätten. Deshalb scheint das varianzanalytische Modell vorteilhafter zu sein, das korrelationsstatistische nur dann gleichwertig, wenn die Urteilsverteilungen der Auswerter in Mittelwert und Streuung gleich sind. Das varianzanalytische Modell liefert eine Schätzung des „Objektivitätskoeffizienten": Dieser gibt an, wie hoch der Anteil der Variabilität zwischen den „tatsächlichen" Punktwerten der Leistungen an der gesamten Variabilität aller abgegebenen Punkturteile ist. Wenn alle Urteile völlig übereinstimmen, ist die Varianz der „tatsächlichen" Werte gleich der Gesamtvarianz, der Objektivitätskoeffizient also 1. Je größer die Abweichungen der Urteiler voneinander, desto niedriger wird der Koeffizient. Die Einzelheiten der Methode und des Rechenverfahrens sind sehr übersichtlich bei ISELER (1967) dargestellt, und wer selbst Untersuchungen zur Objektivität von Aufsatzbeurteilungen durchführen will, dem sei diese Darstellung (S. 135 ff.) als Arbeitsgrundlage empfohlen.

Zu beachten ist, daß es bei diesen Berechnungen nur um die Ermittlung der Auswertungsobjektivität geht, nicht um das, was üblicherweise die „Zuverlässigkeit" einer Prüfmethode genannt wird. Unzuverlässig wird eine Prüfung dadurch, daß der Schüler bei verschiedenen Aufgaben desselben Charakters unterschiedliche Leistungen erbringt, wenn er z. B. heute einen wesentlich besseren Aufsatz schreibt als vor 14 Tagen. Würde man solche Wiederholungsprüfungen durchführen und im Sinne freier Einstufungen beurteilen lassen, so wären die Unterschiede zwischen den Beurteilungen auf zwei Komponenten zurückzuführen: auf die mangelnde Auswertungsobjektivität *und* auf die Schwankungen der tatsächlichen Leistungen. Hier wird nur das erste Problem behandelt.

5.4.5. Versuche zur Erhöhung der Urteilsobjektivität

Mit verschiedenartigen Mitteln wurde versucht, die Objektivität von Aufsatzbeurteilungen zu erhöhen: natürlich zunächst durch Definition von Urteilskriterien, durch Festlegung des Gewichts von Einzelkriterien oder

durch allgemeine Verfahrensregeln für die Beurteilung. Zu diesen Regeln kann auch die Vorgabe einer Punktverteilung gehören, an der sich die Beurteiler streng oder ungefähr orientieren sollen. Man hat weiterhin Beispielaufsätze und deren Beurteilung vorgegeben, so daß die Beurteiler Anhaltspunkte für ihre eigenen Einstufungen hatten. Schließlich ist vorgeschlagen worden, die Beurteiler zu schulen und vielleicht nur bestimmte unter ihnen für die Aufgabe auszuwählen; denn es hatte sich gezeigt, daß es im Grad der Übereinstimmung mit dem Durchschnittsurteil recht große Unterschiede gibt, und manche Beurteiler ständig stark abweichende Einstufungen vornehmen. Beispiele und Vorschläge für diese Vorgehensweisen findet man in der deutschen ebenso wie in der ausländischen Literatur. Systematische Untersuchungen, mit denen geprüft wurde, inwieweit dadurch tatsächlich das Ziel einer möglichst hohen Urteilsobjektivität erreicht wurde, sind vor allem in den angelsächsischen Ländern durchgeführt worden, und darüber soll hier berichtet werden.

Einer der ersten Versuche war die „Willing Scale" (WILLING 1926), die ebenso wie andere „Skalen" dieser Art tatsächliche Aufsätze und deren Punktbewertung angibt, an denen der Auswerter seine Einstufung orientieren soll. Aufsätze zu zehn verschiedenen Themen sollten mit dieser Skala eingestuft werden. Dabei wurde nach Inhalt („story value") und „Form" („form value") getrennt geurteilt; der Punktwert für „Form" ergab sich aufgrund einer einfachen Fehlerzählung. Eine kritische Untersuchung von RUCH & STODDARD (1929) ergab Objektivitätswerte von .0,4 für Inhalt und 0,92 für Form. Offensichtlich ist dieser Objektivitätsgrad für den eigentlich interessanten Aspekt, den Inhalt, viel zu niedrig; die einfache Vorlage von Beispielaufsätzen genügt also keineswegs, eine hinreichende Objektivität zu erreichen, insbesondere nicht, wenn auf derselben Skala Aufsätze zu verschiedenartigen Themen eingestuft werden sollen und wenn klare Auswertungsgesichtspunkte fehlen.

Als wesentlich objektiver erwies sich das Beurteilungssystem von HUDELSON (1923), das ebenfalls Beispielaufsätze zur Orientierung vorgab, das sich allerdings nur auf ein einziges Thema, eine Nacherzählung, bezog. Die Auswerter mußten eine Vorübung absolvieren, indem sie 30 Aufsätze beurteilten, und dann kontrollierten, inwieweit sie dabei die „richtige" Einstufung vorgenommen hatten. Die Objektivität der Beurteilungen lag bei 0,8 und darüber, war also recht zufriedenstellend. Natürlich ist eine Nacherzählung relativ leicht zu bewerten. Trotzdem zeigt das Beispiel, was bei sorgfältiger Vorbereitung der Beurteilungen an Objektivität erreicht werden kann.

Obwohl es noch weitere Beispiele für solche „Aufsatzskalen" gibt, ist es leider bisher zu einer systematischen Weiterentwicklung dieses Ansatzes nicht gekommen. Die Forschung ist andere Wege gegangen, wohl haupt-

sächlich deshalb, weil sie sich auf das besonders dringliche Problem der Beurteilung von Prüfungsaufsätzen konzentriert hat, und weil in diesem Rahmen wegen der großen Zahl der Aufsätze die Frage der Verankerung der Urteile weniger wichtig ist als für den einzelnen Lehrer, der nur die Arbeiten einer Schulklasse vor sich hat und der deshalb nach einer äußeren Orientierung suchen könnte.

Daß es möglich ist, bei Prüfungsaufsätzen eine hohe Übereinstimmung zu erzielen, hatte schon STALNAKER zu Beginn der 30er Jahre bewiesen. Seine Erfahrungen beziehen sich auf Aufsätze, die bei der Aufnahmeprüfung für die Universität Chicago geschrieben wurden. Nachdem die Beurteilungsobjektivität zunächst 0,42 betragen hatte, konnte sie nach zweijähriger Arbeit durch sorgfältige Themenstellung, genaue Auswertungsrichtlinien und durch Auswahl geeigneter Beurteiler auf 0,92 gesteigert werden.

Je größer aber die Zahl der Prüflinge, desto zeitraubender wird die Aufsatzbeurteilung, insbesondere dann, wenn ein „analytisches" Vorgehen gefordert wird, bei dem sorgfältig anhand verschiedenartiger Urteilskriterien vorgegangen werden soll. Bei der Prüfung sehr vieler Schüler verbietet sich deshalb ein solches Verfahren, und wenn man auf die Aufsatzbeurteilung nicht völlig verzichten will, so muß man nach anderen Methoden suchen. Mit diesem Problem hat man sich vor allem in Großbritannien beschäftigt, weil hier bei den 11+ Ausleseuntersuchungen alljährlich sehr viele Prüfungen durchzuführen waren. Man hat deshalb untersucht, ob eine schnelle Einstufung nach dem „allgemeinen Eindruck" nicht ebenfalls zu brauchbaren Resultaten führen könne.

Über den ersten recht erfolgreichen Versuch dieser Art hat WISEMAN im Jahr 1949 berichtet. Beurteilt wurden Aufsätze, die bei der 11+ Prüfung (die etwa unseren Prüfungen für weiterführende Schulen entspricht) geschrieben worden waren. Jeder Aufsatz wurde von vier Beurteilern nach dem Gesamteindruck punktmäßig bewertet, wobei ausdrücklich zu einem raschen, zügigen Arbeiten aufgefordert wurde. Der Durchschnitt dieser vier unabhängigen Urteile galt als Maßzahl für die Qualität des Aufsatzes. Drei Monate nach der ursprünglichen Einstufung wurde die Bewertung für ein Zehntel der Aufsätze wiederholt. Für jeden Beurteiler wurde die Übereinstimmung seiner Einstufungen festgestellt. Wie auch in anderen Untersuchungen festgestellt (z. B. HARTOG 1941), gibt es zwischen den Beurteilern beträchtliche Unterschiede in der Konsistenz ihrer Urteile. Alle Beurteiler, deren Einstufungen eine Korrelation unter 0,70 aufwiesen, wurden durch andere ersetzt. Die „Objektivität" der Beurteilung wurde ermittelt, indem die Durchschnittsurteile bei der ersten Einstufung mit den Durchschnittsurteilen bei der Wiederholung korreliert wurden. Es ergaben sich Korrelationen zwischen 0,91 und 0,94, was sicherlich sehr zufriedenstellende Werte sind. NISBET (1955) hat mit demselben Verfahren einen

ebenso hohen Wert (0,96) erzielt. Daß allerdings nicht die Methode allein für diesen Erfolg verantwortlich ist, ergibt sich aus einer Vergleichsuntersuchung von FINLAYSON (1951). Hier betrug die Korrelation zwischen Erst- und Wiederholungsurteil für das Durchschnittsurteil von 6 Beurteilern nur 0,86. Anders als bei WISEMAN war die Einstufung aber nicht von „erfahrenen Beurteilern" vorgenommen worden, sondern von Lehrern, die sich freiwillig für dieses Experiment zur Verfügung gestellt hatten. Trotzdem ist auch der Wert 0,86 noch beachtlich hoch, so daß das Verfahren der Urteilsmittelung insgesamt eine Bestätigung erfährt. So ist die Methode auch später mit recht gutem Erfolg weiterverwendet worden, z. B. bei Ausleseprüfungen in Australien, über die GOSLING (1966) berichtet.

GOSLINGS Vorgehen nimmt viele verschiedene Ansätze auf und scheint in mancher Hinsicht vorbildlich zu sein. Deshalb soll es etwas ausführlicher beschrieben werden. Von der Annahme ausgehend, daß die verschiedenen Aufsatztypen unterschiedliche sprachliche Fähigkeiten verlangen, wurden den Schülern drei Aufgaben gestellt: (1) Eine sachliche Darstellung (zum Thema „Zeitungen"), (2) Eine Bildbeschreibung (über ein Zirkusbild), (3) Eine Phantasiegeschichte (mit einer kurzen Zeitungsannonce als Anknüpfungspunkt). Den Schülern wurde jeweils genau gesagt, was erwartet und nach welchen Gesichtspunkten ihr Aufsatz bewertet würde. Die Beurteiler gaben ihr Urteil auf einer 15stufigen Skala ab, wobei sie sich je Aufsatztyp an 10 bereits eingestuften Aufsätzen orientieren sollten. Sie wurden in einer Sitzung auf ihre Aufgabe vorbereitet, erhielten genaue Instruktionen, welche Kriterien der Beurteilung zugrunde gelegt werden sollten und übten die Einstufung an mehreren Beispielen. Die drei Aufsätze eines Schülers wurden grundsätzlich von verschiedenen Beurteilern eingestuft. Dieses Urteil, das an den vorgegebenen Kriterien orientiert sein mußte, war ein Gesamturteil, eine separate Einstufung nach den Einzelkriterien erfolgte nicht.

Eine Untersuchung der Auswertungsobjektivität, bei der jeder Aufsatz von drei Beurteilern eingestuft wurde, ergab folgende Ergebnisse: Nahm man als Gesamtwert eines Schülers seine durchschnittliche Einstufung in allen drei Aufsätzen, so korrelierten diese Werte bei der Erst- und Wiederholungsbeurteilung durch dieselben Beurteiler zwischen 0,82 und 0,88. Die Korrelation zwischen den Durchschnittsurteilen von zwei verschiedenen Dreier-Gruppen von Beurteilern lag zwischen 0,65 und 0,75, die Korrelation der Durchschnittsurteile dieser Dreier-Gruppen mit der Einstufung durch den Untersuchungsleiter zwischen 0,82 und 0,90. Im darauf folgenden Jahr ergab sich zwischen dem Urteil der Dreier-Gruppe und dem des Untersuchungsleiters sogar eine Korrelation von 0,95. Die Übereinstimmung der Beurteilung war beim Phantasieaufsatz immer am höchsten, bei der sachlichen Darstellung am niedrigsten, was GOSLING mit der größeren Länge und damit besseren Urteilsgrundlage bei den Phantasieaufsätzen erklärt.

Diese Ergebnisse geben wohl ein realistisches Bild davon, was bei sorgfältiger Vorbereitung und kontrollierter Anwendung zur Zeit an Auswertungsobjektivität erreicht werden kann. Sicherlich ist der Objektivitätsgrad niedriger als bei den meisten Tests. Aber es läßt sich doch eine beachtliche Übereinstimmung erzielen, die weit größer ist, als man nach den beunruhigenden Resultaten, die einleitend beschrieben wurden, erwarten konnte.

Leider gibt es nahezu keine Untersuchungen über die Objektivität in der Feststellung einzelner Beurteilungskriterien. In den beschriebenen Arbeiten, die sich, wie gesagt, auf Prüfungsaufsätze bezogen, war man an dieser Frage weniger interessiert. Für die Aufsätze, die im üblichen Schulalltag geschrieben werden, ist dieses Problem aber wichtiger, besonders dann, wenn man, wie INGENKAMP (in: Der Deutschunterricht, 1971) vorschlägt, bei solchen Aufsätzen keine Zensur gibt, sondern sich mit einzelnen Korrekturen und Anmerkungen begnügt. SCHRÖTER (1971) hat in seinem sehr lesenswerten Untersuchungsbericht einige der neuralgischen Punkte herausgearbeitet, bei denen es anscheinend oft zu Meinungsverschiedenheiten kommt. Allerdings scheinen die abweichenden Ansichten oft durch unterschiedliche Bewertung von Texteigenschaften bedingt, weniger durch Abweichungen in der reinen Feststellung solcher Merkmale. Besonders deutlich ist dies z. B. bei der Frage „lyrisch, dichterisch oder kitschig, klischeehaft?" (SCHRÖTER, S. 52 ff.). Auch wenn man die von ULSHÖFER (1949) veröffentlichten Kommentare zu dem umstrittenen Reifeprüfungsaufsatz liest, fällt auf, daß ungeachtet der großen Unterschiede in der Zensierung bestimmte Aufsatzmerkmale recht einheitlich festgestellt werden.

So werden die notwendigen Bemühungen um eine bessere Aufsatzbewertung zu einem entscheidenden Teil darin bestehen müssen, angemessene Beurteilungskriterien zu erarbeiten. Sicherlich werden vielfach recht fragwürdige Kriterien verwendet: Äußerliche Merkmale wie Rechtschreibung, Schrift, grammatische Fehler, weil sich dies eben leicht feststellen läßt; moralische Gesichtspunkte, weil der Lehrer als Leser nicht völlig vom Inhalt absehen kann und die moralische Reaktion auf Texte eigentlich völlig natürlich ist; schließlich lange tradierte, in Schulbüchern weitergegebene oder höchst subjektive Stilideale. Aber es gibt auch manche Versuche zur rationalen Erarbeitung von Kriterien, z. B. die Untersuchungen über „Altersmundart" und „Stilalter" oder die Ableitung von Kriterien aus Vorstellungen über die Wirkung von Texten auf mögliche Hörer oder Leser, etwa nach den Gesichtspunkten „Interesse", „Verständlichkeit" usw. Freilich — das ist inzwischen weithin bekannt — nutzt das beste Kriterium nichts, wenn es sich nicht empirisch erfassen läßt, und so sind Objektivitätsprüfungen der beschriebenen Art ein unerläßlicher Bestandteil der Kriteriendefinition. Daß tatsächlich ein befriedigender Objektivitätsgrad erreichbar ist, geht aus den beschriebenen Untersuchungen hervor.

5.4.6. Literaturverzeichnis

Bohusch, O.: Neue Kriterien für die Aufsatzbewertung. Don Bosco, München 1972.

Cast, B. M.: The efficiency of different methods of marking English compositions. Brit. J. Educ. Psych., 1939, *9*, 257—269.

Diederich, P. B., French, J. W. u. *Carlton, S. T.:* Factors in Judgements of Writing Ability. Res. Bull. Series RB — 61 — 15. Princeton, E. T. S. 1961.

Eells, W. C.: Die Zuverlässigkeit wiederholter Benotung von aufsatzähnlichen Prüfungsarbeiten. In: *Ingenkamp, K.* (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim 1971.

Finlayson, D. S.: The Reliability of the Marking of Essays. Brit. J. Educ. Psych., 1951, *21*, 216—234. — Dt. Übers. in: *Ingenkamp, K.* (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim 1971.

Gosling, G. W. H.: Marking English Compositions. Australian Council for Educational Research, 1966.

Hahn, H.: Flensburger Norm für die Aufsatzbeurteilung. Die pädagogische Provinz, 1966, *20*, 189—194.

Hartog, P. u. *Rhodes, E. C.:* An examination of examinations. McMillan, London 1936. — Auszugsweise in dt. Übers. in: *Ingenkamp, K.* (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim 1971.

Hudelson, E., The Hudelson Typical Composition Ability Scale. Public School Publishing Co. Bloomfield, Ill. 1923.

Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Beltz, Weinheim, Berlin, Basel 1971.

Ingenkamp, K.: Probleme der schulischen Leistungsbeurteilung unter besonderer Berücksichtigung des Deutschunterrichts. Der Deutschunterricht, 1971, *23*, 54—76.

Iseler, A.: Zur varianzanalytischen Schätzung der Auswertungsobjektivität von psychologischen Tests. Diagnostica, 1967, *13*, 135—148.

Kießling, A.: Leistungsbeurteilung und Leistungsmessung. Z. f. Päd. Psych., 1929, *30*.

Kötter, L. u. *Grau, U.:* Zur Bedingtheit der uneinheitlichen Benotung von Schüleraufsätzen (Nacherzählungen). Z. exp. angew. Psych., 1965, *12*, 278—301.

Lehmann, E.: Das gerechte Zeugnis im Aufsatz. Schulwarte, 1951, *4*, 32—43.

Lietzmann, W.: Über die Beurteilung von Leistungen in der Schule. Leipzig 1927.

Nisbet, J. D.: English Composition in Secondary School Selection. Brit. J. Educ. Psych., 1965, *35*, 51—54.

Ruch, G. M. u. *Stoddard, G. D.:* Tests and Measurement in High School Instruction. World Book Co., New York 1929 (112—129, 592—603).

Schröter, G.: Die ungerechte Aufsatzzensur. F. Kamp, Bochum 1971.

Schröter, G.: Aufsätze, Zensuren und Moral. Westermanns Päd. Beitr., 1968, *20*, 26—27.

Stalnaker, J. M.: The Construction and Results of a 12 Hour Test in English Composition. School a. Society, 1934, *39*, 218—224.

Starch, D. u. *Elliot, E. C.:* Die Verläßlichkeit der Zensuren von Mathematikarbeiten. In: *Ingenkamp, K.* (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim 1971.

Ulshöfer, R.: Wie beurteilen Sie diesen Reifeprüfungsaufsatz? Der Deutschunterricht, 1948/49, *1*, 95—98.

Weber, A.: Das Problem der Aufsatzbeurteilung. Ludwig Auer, Donauwörth 1969.

Wieczerkowski, W., Nickel, H. u. *Rosenberg, L.*: Einige Bedingungen der unterschiedlichen Bewertung von Schüleraufsätzen. Psychol. Rundsch., 1968, *19*, 280 bis 295.

Willing, M. H.: Scale for Measuring Written Composition. Public School Publishing Co., Bloomington, Ill. „Individual Diagnosis in Written Composition". Journal of Educational Research, 1926, 77—89.

Wiseman, S.: The Marking of English Composition in Grammer School Selection. Brit. J. Educ. Psych., 1949, *19*, 200—209.

6. Autorenverzeichnis

Peter Büscher, Jahrgang 1936, studierte Psychologie und Erziehungswissenschaft an den Universitäten Heidelberg und Mannheim (Diplom-Prüfung 1970). Der Autor ist zur Zeit als Sonderschullehrer an der Gehörlosenschule Heidelberg tätig und war von 1971—73 Leiter der Bildungsberatungsstelle Mannheim. Außerdem ist er Lehrbeauftragter für pädagogische Diagnostik an der Universität Mannheim (Lehrstuhl Prof. Dr. L. Michel).
Arbeitsschwerpunkt: Psychologische Diagnostik im Bildungswesen.
Wichtigste Veröffentlichungen: Probleme der Differenzierung. Hörgeschädigtenpädagogik, 25 (1971), 1—12; Mannheimer Intelligenztest (zus. mit W. Conrad u. a.). Weinheim (Beltz) 1971; Operationalisierte Lernziele und pädagogische Taxonomien. Hörgeschädigtenpädagogik, 27 (1973), 31—42.
Anschrift: 6806 Viernheim, Eichenstraße 14.

Walter Fingerhut, Jahrgang 1946, studierte Psychologie an der Universität Marburg (Diplom-Prüfung 1971). Der Autor ist zur Zeit Wissenschaftlicher Angestellter im Fachbereich Psychologie der Universität Marburg.
Arbeitsschwerpunkt: Pädagogische Psychologie.
Veröffentlichungen: Der Beitrag biographischer Daten von Schülern und Lehrern zur Vorhersage von Schulnoten (zus. mit H.-P. Langfeldt). Vortrag auf dem 28. Kongr. Dt. Ges. Psychol. 1972 in Saarbrücken, Göttingen (Hogrefe), Kongreßbericht im Druck; Erfahrungen mit dem Allgemeinen Schulleistungstest AST 4 (zus. mit H.-P. Langfeldt). Psychol. in Erz. u. Unterr., 20 (1973), 249—257; Entwicklung eines Lehrer-Einstellungs-Fragebogens LEF-3 (zus. mit H.-P. Langfeldt). Z. f. Entwicklungspsychol. u. Päd. Psychol., 6 (1974), im Druck.
Anschrift: 3551 Cappel, Goethestraße 40.

Anne-Katrin Gaedike, Jahrgang 1949, studierte Psychologie an den Universitäten Braunschweig und Hamburg (Diplom-Prüfung 1972) sowie Pädagogik an der Pädagogischen Hochschule Rheinland, Abteilung Bonn. Die Autorin ist zur Zeit Wissenschaftliche Assistentin am Psychologischen Seminar der Abteilung Bonn der Päd. Hochschule Rheinland.
Arbeitsschwerpunkte: Experimentelle und Angewandte Psychologie, Psychologische Diagnostik und Methodenlehre.
Veröffentlichung: Planung und Auswertung empirischer Untersuchungen — Einführung für Pädagogen, Psychologen und Soziologen (zus. mit K. Heller und B. Rosemann). Stuttgart (Klett) 1974.
Anschrift: 53 Bonn, Rittershausstraße 23.

Kurt Heller, Jahrgang 1931, studierte Psychologie und Erziehungswissenschaft an den Universitäten Freiburg/Br. und Heidelberg (Diplom-Prüfung 1964, Promotion 1968). Der Autor ist ordentl. Professor für Psychologie an der Pädagogischen Hochschule Rheinland, Abteilung Bonn, und Lehrbeauftragter an der Universität Heidelberg im Fachbereich Erziehungswissenschaft.
Arbeitsschwerpunkte: Begabungs- und Bildungsforschung, Psychologische Diagnostik und Beratung im Bildungswesen, Sonderpädagogische Psychologie.
Wichtigste Veröffentlichungen: Modell eines Guidance-Systems für Abiturienten und Studenten (zus. mit E. Demel u. G. Schorre). In: Bildung in neuer Sicht, Bd. 20, Villingen (Neckarverlag) 1969; Aktivierung der Bildungsreserven. Bern u. Stuttgart (Huber/Klett) 1970; Heidelberger Hörprüf-Bild-Test HHBT für Schulanfänger (zus. mit A. Löwe). Villingen (Neckarverlag) 1972; Intelligenzmessung — Zur Theorie und Praxis der Begabungsdiagnostik in Schule und Son-

derpädagogik. Villingen (Neckarverlag) 1973; Wortschatztests für Sehbehinderte WST(Sb) 4—9 (zus. mit B. Schirmer). Weinheim (Beltz) 1973; Wortschatztests für Blinde WST(Bl) 4—8 (zus. mit B. Köhn). Weinheim (Beltz) 1973; Planung und Auswertung empirischer Untersuchungen (zus. mit B. Rosemann u. A.-K. Gaedike). Stuttgart (Klett) 1974; Kognitiver Fähigkeits-Test KFT 4—13 (zus. mit A.-K. Gaedike u. H. Weinläder). Weinheim (Beltz) 1974. Mitherausgeber der „Schriftenreihe zur Bildung und Rehabilitation Sehgeschädigter" (Schindele) und „Sehgeschädigte" — Intern. Wiss. Arch. (Schindele).
Anschrift: 6903 Neckargemünd, In den Wingert 6.

Ralf Horn, Jahrgang 1942, studierte Psychologie an der Universität Freiburg/Schweiz und Frankfurt/Main (Diplom-Prüfung 1967). Der Autor ist wiss. Leiter der Beltz Test Gesellschaft in Weinheim/Bergstr. und Lehrbeauftragter an der Erziehungswissenschaftlichen Hochschule Rheinland-Pfalz, Abteilung Worms.
Arbeitsschwerpunkte: Testentwicklung, Forschungsmethoden.
Wichtigste Veröffentlichungen: Lernziele und Schülerleistung. Weinheim (Beltz) 1972, 1973 (3. Aufl.); Lernziel-Test Mathematik 5. Schuljahr (zus. mit B. Andelfinger). Weinheim (Beltz) 1973; Sozialschicht und Intelligenzleistung. Psychol. in Erz. u. Unterr., 21 (1974).
Anschrift: 684 Lampertheim, Daimlerstraße 49.

Hans-Peter Langfeldt, Jahrgang 1943, studierte Psychologie und Erziehungswissenschaft an den Universitäten Tübingen und Marburg (Diplom-Prüfung 1971). Der Autor ist zur Zeit Wissenschaftlicher Assistent an der Pädagogischen Hochschule Heidelberg, Fachbereich Sonderpädagogik.
Arbeitsschwerpunkte: Pädagogische Psychologie und Diagnostik.
Veröffentlichungen: Der Beitrag biographischer Daten von Schülern und Lehrern zur Vorhersage von Schulnoten (zus. mit W. Fingerhut). Vortrag auf dem 28. Kongreß Dt. Ges. Psychol. 1972 in Saarbrücken, Göttingen (Hogrefe), Kongreßbericht im Druck; Erfahrungen mit dem Allgemeinen Schulleistungstest AST 4 (zus. mit W. Fingerhut). Psychol. in Erz. u. Unterr., 20 (1973), 249—257; Entwicklung eines Lehrer-Einstellungs-Fragebogens LEF-3 (zus. mit W. Fingerhut). Z. f. Entwicklungspsychol. u. Päd. Psychol., 6 (1974); Entscheidungstheoretische Aspekte der Umschulungsdiagnostik in die Sonderschule für Lernbehinderte. Z. f. Heilpäd., 25 (1974).
Anschrift: 6903 Neckargemünd, Waldstraße 14.

Erich Langhorst, Jahrgang 1925, studierte nach der 1. und 2. Lehrerprüfung Psychologie an der Universität Bonn (Diplom-Prüfung 1962, Promotion 1966). Der Autor ist Akademischer Oberrat für Psychologie an der Pädagogischen Hochschule Rheinland, Abteilung Bonn.
Arbeitsschwerpunkte: Entwicklung der kognitiven Funktionen und allgemeinen Leistungsfähigkeit, Psychologie des Lese- und Rechtschreibeunterrichts, Lese-Rechtschreib-Schwäche.
Wichtigste Veröffentlichungen: Märchenbilder im Urteil von Kindern der Vorkriegszeit und Gegenwart — Ein Beitrag zur Psychologie des Bilderlebens der Sechs- bis Vierzehnjährigen. Bonn (Bouvier) 1967; Brennpunkte der pädagogischen Psychologie (Hrsg. zus. mit H. Nickel). Bern u. Stuttgart (Huber/Klett) 1973.
Anschrift: 53 Bonn, Luisenstraße 2.

Horst Nickel, Jahrgang 1929, studierte Psychologie und Erziehungswissenschaft an den Universitäten Marburg (Diplom-Prüfung 1961) und Erlangen-Nürnberg (Promotion 1965). Der Autor ist ordentl. Professor für Psychologie (Entwicklungs-

Psychologie und Schulpädagogik

Heidelberger Autorengruppe: Steckbrief der Psychologie. Hrsg. von Klaus Eckart Rogge. 2., durchgesehene Auflage, 290 Seiten, DM 12,80. UTB 37

Hans-Jürgen Pfistner (Hrsg.): Aspekte der Pädagogischen Psychologie. Psychologie im Studium der Lehrer. 236 Seiten, DM 14,—

Hans Maier / Hans-Jürgen Pfistner: Die Grundlagen der Unterrichtstheorie und der Unterrichtspraxis. Ein Beitrag zur Phänomenologie des Unterrichts. 248 Seiten, DM 37,—

Wolfgang Fischer (Hrsg.): Schule und kritische Pädagogik. Fünf Studien zu einer pädagogischen Theorie der Schule. 176 Seiten, DM 19,—

Jörg Ruhloff: Ein Schulkonflikt wird durchgespielt. Beschreibung und Analyse. Mit einem Vorwort von Wolfgang Fischer. 112 Seiten, DM 12,—

Christa Berg: Die Okkupation der Schule. Eine Studie zur Aufhellung gegenwärtiger Schulprobleme an der Volksschule Preußens (1872—1900). 272 Seiten, DM 28,—

Bruno Krapf: Unterrichtsverlauf und programmierte Lernhilfen. 144 Seiten, DM 21,—

Georg E. Becker: Optimierung schulischer Gruppenprozesse durch situatives Lehrtraining. Mit Studienmaterialien und Trainingsunterlagen. 180 Seiten, DM 16,50 (Gruppenpädagogik — Gruppendynamik, Bd. 3)

David Warwick: Team Teaching. Grundlegung und Modelle. Aus dem Englischen übersetzt und bearbeitet von Rainer Winkel. 148 Seiten, DM 19,— (Gruppenpädagogik — Gruppendynamik, Bd. 5)

Ernst Meyer und Börje Forsberg (Hrsg.): Einführung in die Praxis der schulischen Gruppenarbeit. Materialien für Lehrer, Schüler und Eltern. 180 Seiten, DM 16,— (Gruppenpädagogik — Gruppendynamik, Bd. 8)

Quelle & Meyer Heidelberg