Ludwig-Maximilians-Universität München
Institut für Statistik
Herr Prof. Dr. Thomas Augustin

Master-Thesis
Cronbachs α im Kontext des Grundmodells der klassischen Testtheorie und darüber hinaus

Andreas Bauer
23. Februar 2015

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

In der Psychologie oder auch in der Soziologie wird die Reliabilität einer Skala oftmals mit dem Koeffizienten α abgeschätzt. Der Koeffizient α unter- bzw. überschätzt jedoch in verschiedenen Situationen die wahre Reliabilität einer Skala. Dies konnte im Rahmen der Arbeit insbesondere gezeigt werden für den Fall, dass eine Kovarianz zwischen dem wahren Wert (z.B. der Intelligenz) und dem in der klassischen Testtheorie angenommenen Fehler besteht. Der Koeffizient α ist jedoch des Weiteren bei einem Vorliegen einer derartigen Kovarianz eine untere Schranke für die Reliabilität. Um die Schätzungen der Reliabilität in diesem Fall zu verbessern, wurde der adjustierte Koeffizient α_ψ entwickelt. Zudem wurde ein α-Konfidenzintervall und ein entsprechender Test entwickelt. Hierdurch kann getestet werden, ob eine Skala eine bestimmte definierte Grenze für die Reliabilität im verbalen Sinn übersteigt oder nicht. Für die Grenze wird eine Adjustierung vorgeschlagen, die die Abhängigkeit des Koeffizienten von der Anzahl der verwendeten Items berücksichtigt. Des Weiteren wurde ein Gegenvorschlag, nämlich die Berechnung eines durchschnittlichen α-Wertes, zu der α-Maximierung entwickelt. Diese neu entwickelten Methoden wurden im Rahmen der Arbeit durch eine Simulation näher betrachtet. Durch diese neue Methodik können nach Ansicht des Autors differenziertere Angaben über die Reliabilität gemacht werden.

1 Einleitung

Diese Arbeit beschäftigt sich mit dem Reliabilitätsmaß Cronbachs α. Es handelt sich hierbei um ein Maß zur Bestimmung der Zuverlässigkeit einer Messung bzw. einer Skala. Insbesondere in der Psychologie ist es das Ziel, nicht beobachtbare theoretische Konstrukte, wie z.B. die Intelligenz, die Motivation oder die Extraversion zu messen. Aber auch in der Soziologie ist Cronbachs α ein sehr verbreitetes Maß zur Bestimmung der Reliabilität einer Skala. Als Beispiel können hier die Untersuchung des Wahlverhaltens von Personengruppen, oder Einstellungen, beispielsweise zu politischen Fragen, genannt werden. Die entscheidende Frage ist dabei, was verbirgt sich hinter diesen Begriffen (Intelligenz, Motivation, Extraversion, Rechtsextremismus). Wie ist beispielsweise der Begriff der Intelligenz substanzwissenschaftlich definiert? Hier gibt es natürlich unterschiedliche Vorstellungen der Wissenschaftler, was uns auch nicht beunruhigen sollte. Nassehi (2008: 182) beschreibt dies folgendermaßen: „Die Inanspruchnahme wissenschaftlichen Wissens produziert eher Unsicherheit, sie erzeugt mehr Nicht-Wissen als Wissen, denn diese Art wissenschaftlicher Erkenntnis schafft offensichtlich nicht die Fragen aus der Welt, sondern versieht alles Wissen mit dem Index der Kontingenz."

Die zu definierenden theoretischen Konstrukte lassen sich meist nicht direkt beobachten, denn sie sind latent. Es ist daher erforderlich, dass sie messbar gemacht werden. In der Psychologie, aber auch in anderen empirischen Wissenschaften, werden darum Indikatoren entwickelt, die das latente Konstrukt abbilden sollen. Es werden meist mehrere Indikatoren theoretisch entwickelt und zu einem Index zusammengefasst. Bei den Indikatoren handelt es sich regelmäßig um Fragen im Rahmen eines Fragebogens. Der entwickelte Index kann dann als psychologischer Test zur Beschreibung einer bestimmten Eigenschaft oder eines Merkmals einer Person aufgefasst werden. Beispielsweise kann ein psychologischer Test zur Diagnose verwendet werden, ob eine Person eine psychische Erkrankung hat oder nicht. Die Entwicklung eines Messverfahrens zur Abbildung eines theoretischen Konstrukts ist eine Aufgabe beispielsweise der Psychologen bzw. Soziologen.

Die statistische Analyse beginnt nach der Festlegung der zu testenden Skala. Es kann eine Überprüfung der Zuverlässigkeit der entwickelten Skala im Rahmen der klassischen Testtheorie (KTT) erfolgen. Die Überprüfung der Reliabilität kann unter anderem mit dem Koeffizienten α nach Cronbach (1951) erfolgen. Dieser Koeffizient kann aus der klassischen Testtheorie entwickelt werden. Die soll im Rahmen dieser Arbeit dargestellt werden.

Die Arbeit beginnt mit der Darstellung von Grundbegriffen (Kapitel 1). Dabei handelt es sich um das Überbrückungsproblem von der Theorie zur Empirie, die Defini-

tion von psychometrischen Tests und Items und um die theoretische Definition des Konzeptes des wahren Wertes und des Messfehlers. Es erfolgt darauf eine kurze Einführung zu den Gütekriterien Objektivität, Reliabilität und Validität (Kapitel 2). Anschließend erfolgt eine detaillierte Darstellung der klassischen Testtheorie. Hierbei handelt es sich um die Grundlage für den Koeffizienten α von Cronbach (Kapitel 3). Danach geht es darum, einen Zusammenhang zwischen der klassischen Testtheorie und der Reliabilität einer Skala herzustellen (Kapitel 4). Es wird sich zeigen, dass der Koeffizient α aus der Grundannahme der klassischen Testtheorie und den Folgerungen hieraus hergeleitet werden kann. Falls die Grundannahmen der klassischen Testtheorie eingehalten sind, dann ist der Koeffizient eine untere Schranke für die wahre aber unbekannte Reliabilität. Liegen zudem essentiell τ-äquivalente Messungen vor, dann entspricht der Koeffizient der wahren Reliabilität.
Im Anschluss werden die Vorteile des Koeffizienten dargestellt (Kapitel 5). Der überwiegende Teil der Arbeit befasst sich mit den Schwierigkeiten des Koeffizienten. Es wird im Kapitel 6 hierzu zuerst die Abhängigkeit des Koeffizienten von der Anzahl der verwendeten Items dargestellt. Ferner stellt sich die Frage, ab wann eine Skala eine zufriedenstellende Reliabilität im verbalen Sinn aufweist. Hierzu wurden in der Literatur allgemeine Grenzen definiert. Durch ihre pauschale Anwendung kann es jedoch zu Problemen kommen. Es wird hierzu ein Gegenvorschlag unterbreitet. Weiter wird dargestellt, welche Auswirkungen sich ergeben, falls keine essentiell τ-äquivalenten Messungen vorliegen. Der Fokus dieser Arbeit liegt darin herauszufinden, wie sich die Verletzung von Folgerungen aus der Grundannahme der klassischen Testtheorie auf den Koeffizienten α auswirken. Es erfolgt eine analytische Betrachtung für den Fall, dass keine unkorrelierten Fehler vorliegen und die Darstellung von Simulationsergebnissen. Entsprechend wird vorgegangen, falls der wahre Wert und der Fehler unkorreliert sind.
Im Kapitel 7 erfolgt ein Gegenvorschlag zu der oft durchgeführten α-Maximierung. Im folgenden Kapitel 8 wird betrachtet, wie ein Konfidenzintervall für den Koeffizienten α konstruiert werden kann und wie ein entsprechender Test formuliert werden kann. Es können hierdurch asymptotische Aussagen hinsichtlich der Items getroffen werden. Die Konstruktion des Konfidenzintervalls erfolgt mit Hilfe der Bootstrap-Methode.
Die Vorschläge aus den Kapiteln 7 und 8 werden im Kapitel 9 durch eine Simulation und durch die Auswertung von empirischen Daten näher betrachtet. Abschließend wird ein Resümee über die Arbeit gezogen.

2 Grundbegriffe

In diesem Kapitel werden die Grundbegriffe geklärt, um einen ersten Überblick über das Thema zu erhalten. Hierbei werden insbesondere die Begriffe Überbrückungsproblem, Indexbildung, Messfehler und Messfehlertheorie dargestellt.

2.1 Überbrückungsproblem

In der Psychologie oder Soziologie werden bestimmte Theorien entwickelt. Es wird versucht, diese mit Messungen (vorerst) zu bestätigen bzw. zu verwerfen. Es geht um eine Verknüpfung der Theorie mit der Empirie. Diese Verbindung wird als Operationalisierung bezeichnet bzw. auch als Überbrückungsproblem (Steyer 1993: 3). Die folgenden Ausführungen erfolgen, damit ein Eindruck gewonnen werden kann, wie die für die statistische Analyse relevanten Skalen entstehen.
Es gibt eine Vielzahl von Begriffen, die bestimmte Sachverhalte abbilden. Hier sind als Beispiele die Begriffe Intelligenz, Depression oder Sozialkapital zu nennen. Es handelt sich hierbei um nicht direkt beobachtbare und somit latente Begriffe. Es ist nun wichtig auch Anweisungen theoretisch zu fundieren, wie die abstrakten und latenten Begriffe messbar gemacht werden können. Es handelt sich hierbei um die Operationalisierung (Schnell 1995: 117 ff.). Operationalisierung wird dadurch folgendermaßen definiert: „Als Operationalisierung einer Variable definieren wir eine Menge hinreichend genauer Anweisungen, nach denen Untersuchungseinheiten den Kategorien einer Variablen zugewiesen werden.“ (Diekmann 2010: 239). Hiermit verbunden ist, dass ein theoretischer Begriff mindestens eine Dimension haben muss. Die beobachteten Sachverhalte können durch die konstruierten Zuordnungsvorschriften den Dimensionen der abhängigen Variablen zugeordnet werden.
Als Beispiel kann hier die Frage verwendet werden, welche Kriterien Personen zu ihrer subjektiven Abgrenzung „ihrer“ Ethnie gegenüber anderen Ethnien verwenden. Für den Begriff der Ethnie bzw. für die Abgrenzung von Ethnien gibt es nach Weber z.B. die Dimensionen: „die in die Augen fallenden Unterschiede in der Lebensführung des Alltags“, die Sprachgemeinschaft, „die Unterschiede der typischen Kleidung, der typischen Wohn- und Ernährungsweise, die übliche Art der Arbeitsteilung zwischen den Geschlechtern.“ (1980: 238-239). Sicherlich gibt es hierzu auch andere Ansätze für eine geeignete Dimensionierung.
Es stellt sich nun die Frage, wie beispielsweise die übliche Art der Arbeitsteilung zwischen den Geschlechtern messbar gemacht werden kann. Dies könnte man mit der Vollzeit- bzw. Teilzeitquote oder auch mit Daten zur Haushalts- und Familienarbeit nach Geschlechtern und Altersgruppen in Deutschland für die Tätigkeiten

Zubereitung von Mahlzeiten, Instandhalten von Haus und Wohnung, Kinderbetreuung, Wäschepflegen, Gartenarbeit und Einkaufen (in Minuten pro Tag) versuchen (Statistisches Bundesamt 2004: 62). Neben den Daten des statistischen Bundesamtes könnten auch Befragungen der Personen durchgeführt werden. Die Operationalisierung erfolgt hier durch die Entwicklung von Testfragen.

2.2 Psychometrische Tests

Es erfolgt eine Definition von psychometrischen Tests, da diese vor allem die Grundlage für die Berechnung des Koeffizienten α nach Cronbach bilden. Ziel von Untersuchungen in der Psychologie ist es unter anderem, die wahren Unterschiede zwischen Personen abzubilden (Danner 2015: 1). Durch die erhobenen Daten, z.B. im Rahmen eines Intelligenztestes, soll möglichst genau ein Unterschied zwischen Personen oder auch Personengruppen herausgearbeitet werden. Um die wahren Unterschiede zwischen den Personen herausfinden zu können, ist insbesondere eine hohe Reliabilität der unterschiedlichen Items von Nöten. Der Begriff der Reliabilität wird im Kapitel 4 eingeführt.
Psychometrische Tests können dabei folgendermaßen definiert werden: „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 1998: 18). Hierfür sollte natürlich eine Theorie zur Beschreibung von Personen vorliegen (Bühner 2004: 17). In dieser Definition findet sich das unter Gliederungspunkt 2.1 beschriebene Überbrückungsproblem wieder. Es wird eine latente Variable, also die abhängige Variable, mit Routineverfahren messbar gemacht, um Aussagen über das Vorliegen bestimmter Ausprägungen bei den Probanden machen zu können. Es erfolgt also eine Verknüpfung von der Empirie mit der Theorie. Die Routineverfahren zur Erhebung der Daten können beispielsweise ein Fragebogen, ein (standardisiertes) Interview oder eine (standardisierte) Beobachtung sein. Ziel eines Tests soll es sein, vergleichende Aussagen über die an der Studie teilnehmenden Personen herauszufinden (Rost 1996: 19).

Mit Hilfe einer Theorie soll erreicht werden, dass ein empirisches Phänomen beschrieben werden kann, oder auch eine Verhaltensweise einer Person prognostiziert werden kann. Ein Testmodell soll die systematischen Zusammenhänge zwischen den gestellten Fragen und den gegebenen Antworten durch latente, nicht beobachtbare Variablen erklären (Mortensen 2009: 15). Diese ergeben sich aus der vorgegebenen Theorie. Da die latenten Variablen nicht direkt beobachtbar sind, muss ein Weg ge-

funden werden, diese messbar zu machen. Die Messung geschieht durch verschiedene Items. Unter dem Begriff Items werden in der Testtheorie Aufgaben oder Fragen im Rahmen von Tests zusammengefasst. Sie sind beobachtbare Indikatoren von latenten Variablen.

Die Antwortmöglichkeiten können dichotom oder polynom kodiert sein (Mortensen 2009: 16). Falls eine dichotome Kodierung vorliegt, wird von den Probanden eine Antwort beispielsweise zu einer Richtig-oder-Falsch-Frage gefordert. Bei einer polynomen Kodierung wird die Einordnung der Antwort in geordnete oder ungeordnete Kategorien gefordert.

2.3 Indexbildung

Items werden zur Beschreibung eines theoretischen Konstrukts definiert. Durch die Analyse des Antwortverhaltens der Probanden zu den Items können Rückschlüsse auf das zu messende theoretische Konstrukt getroffen werden. Die Vorschriften hierzu werden als Messmodell bezeichnet (Steyer 1993: 3 ff.).
Eine Vorschrift eines Messmodells kann darin bestehen, dass ein Index gebildet wird. Ein Index ist eine Zusammenfassung von mehreren Items zu einer neuen Variable (Schnell 1995: 157). Es stellt sich hierbei die Frage, weshalb man überhaupt einen Index bilden sollte. Zum Einem kann dies erforderlich sein, um die gesamte Dimension des theoretischen Konstruktes abbilden zu können. Um die Intelligenz einer Person zu erheben, wird sicherlich nicht eine Frage ausreichen. Zum Anderen kann hierdurch auch erreicht werden, dass die Genauigkeit der Messung durch mehrere unabhängige Messungen zunimmt, da zufällig hohe bzw. tiefe Testergebnisse durch mehrere Items ggf. ausgeglichen werden können (Schnell 1995: 125). Indizes können dabei additiv, multiplikativ oder auch gewichtet additiv sein (Schnell 1995: 162). Vor allem einfache Summierungen oder die Bildung eines Durchschnittes über die einzelnen Items bilden den empirischer Ausgangspunkt für die klassische Testtheorie (Steyer 1993: 100, Lord 1974: 82).

Ein Index kann auch als zusammengesetzter Test betrachtet werden. Es ergibt sich eine zusammengesetzte Messung aus den verschiedenen Komponenten des Tests. Die Länge eines Tests ergibt sich im einfachsten Fall aus der Anzahl der verwendeten Items. Andererseits könnte sich die Länge eines Tests auch z.B. aus der Anzahl der Wörter bei einem Lesetest ergeben, der die Lesegeschwindigkeit abbilden soll. Es könnte auch die Zeit genommen werden, bis eine bestimmte Anzahl von Wörtern gelesen wurde (Lord 1974: 82 ff., Fischer 1974: 52 ff.).

2.4 Wahrer Wert und Messfehler

Im Rahmen dieses Abschnitts werden die Messergebnisse näher betrachtet. Es wird in der klassischen Testtheorie angenommen, dass sich ein Messwert aus einem (nicht beobachtbaren) wahren Wert τ und zusätzlich aus einem Messfehler ϵ ergibt (Novick 1966: 56, Bühner 2004: 22).

Thorndike würde gegen das Konzept des wahren Wertes argumentieren, dass der wahre Wert nicht beobachtbar ist und dadurch die Idee der klassische Testtheorie lediglich mythisch ist und der wahre Wert keine theoretische Bedeutung hat (1964). Ferner argumentiert Loevinger gegen das Konzept des wahren theoretischen Wertes, dass einzig allein der beobachtete Wert Relevanz hat und jede Frage, die mit dem wahren Wert beantwortet werden soll, keine Bedeutung hat (1957).

Als Messfehler wird eine zufällige Variable des Fehlerwertes definiert (Guliksen 1967:6-7; Lord 1974: 38-37, Mortensen 2009: 58). Es handelt sich um eine Störgröße, die sich aufgrund von verschiedenen nicht kontrollierbaren Einflüssen auf die Messung ergibt. Im Rahmen der klassischen Testtheorie werden lediglich unsystematische Fehler betrachtet. Systematische Fehler werden hingegen nicht betrachtet, da diese in den wahren Wert absorbiert werden können. Bei den unsystematischen Fehlern kann zwischen unsystematischen äußeren Einflüssen, wie z.B. schwankende Lichtverhältnisse oder Regen und unsystematischen inneren Einflüssen wie z.B. Müdigkeit oder Motivation differenziert werden. Die unterschiedlichen Fehlerquellen akkumulieren sich zumeist.

2.5 Gütekriterien von Messungen

Die Messungen sollen im Rahmen einer Untersuchung möglichst objektiv, zuverlässig und valide sein (z.B. Diekmann 2010: 247 ff., Bühner 2004: 115 ff., Danner 2015, Fischer 1974: 36 ff., Rost 1996: 31 ff., Schnell 1995: 139 ff., Mortensen 2009: 13). Der Grad der Objektivität ergibt sich daraus, inwiefern die Testergebnisse unabhängig von den Forscherinnen und Forschern sind. Durchführungsobjektivität bedeutet, dass die Forscher A und B jeweils zu dem gleichen Ergebnis kommen müssen. Der Grad der Übereinstimmung kann z.B. anhand eines Korrelationskoeffizienten berechnet werden. Auch sollten die Ergebnisse unabhängig von den gewählten Testitems sein. Die Auswertung der Ergebnisse sollte sich bei den Forschern entsprechen (Auswertungsobjektivität). Die Reliabilität (Zuverlässigkeit) ist die Genauigkeit eines Tests. Die Maßzahlen für die Reliabilität (z.B. Cronbachs α) geben an, wie genau

der Test das misst, was er tatsächlich misst. Eine ausführliche Definition des Begriffes der Reliabilität erfolgt im Kapitel 4. Dabei müssen objektive und reliable Messungen nicht valide sein. Validität bedeutet, dass der Test auch tatsächlich das misst, was er messen soll.

3 Klassische Testtheorie

Die Antworten der Versuchspersonen zu den verwendeten Items werden im Rahmen eines Testmodells ausgewertet. Es wird im Folgenden als Testmodell die klassische Testtheorie verwendet. Das allgemeine Ziel der Testtheorie ist es, von manifesten auf latente Variablen schließen zu können (Mortensen 2009: 19).
Die klassische Testtheorie war die erste Testtheorie, die für die Konstruktion von psychologischen Tests eingesetzt wurde. Deshalb wird sie auch als klassische Testtheorie bezeichnet (Bühner 2004: 21). Ihr großer Vorteil ist die einfache Anwendbarkeit. Diekmann (2010: 261) definiert die klassische Testtheorie wie folgt: „Die sogenannte klassische Testtheorie behandelt den Zusammenhang zwischen Messinstrumenten und theoretischen Konstrukten im Rahmen eines mathematisch-statistischen Modells.“ Die klassische Testtheorie interessiert sich dafür, aus welchen Komponenten sich die Messwerte ergeben, also aus dem wahren Wert und aus dem Messfehler (Fischer 1974: 124). Es handelt sich um ein Modell mit relativ schwachen Annahmen und hieraus ergibt sich eine breite Anwendbarkeit. Ein Modell mit strengeren Annahmen ist beispielsweise das sogenannte Poisson-Process-Modell (Lord 1974: 480 ff.). Die klassische Testtheorie wird im Folgenden detailliert dargestellt, da durch diese eine Definition des Begriffs der Reliabilität möglich ist. Des Weiteren kann aus der klassischen Testtheorie das Reliabilitätsmaß α nach Cronbach hergeleitet werden.
Die Ausführungen zur klassischen Testtheorie beziehen sich auf Mortensen (2009: 56 ff.), Steyer (1993: 101 ff.), Lord (1974: 13 ff.), Gulliksen (1950: 4ff.) und Diekmann (2010: 261 ff.).

Für die Erhebung der Daten wird ein Zufallsexperiment angenommen und aus diesem Grund ist der wahre Wert auch stets zufällig. Im Rahmen eines Zufallsexperiments erfolgt eine zufällige Auswahl von Personen, die einen zufälligen Beobachtungswert aufweisen. Sei a eine Person aus einer Population P ($a \in P$). Die Beobachtungswerte einer Person haben eine Wahrscheinlichkeitsverteilung bzw. eine Wahrscheinlichkeitsdichte. Sei des Weiteren ein Test (oder ein Testitem) mit dem Buchstaben g definiert, der aus einem Set von Tests bzw. Testitems G stammt ($g \in$

G). Der Testwert X_{ga} gibt nun die Gesamtzahl von Punkten, oder auch die Punktzahl zu einer Aufgabe an, die die zufällig ausgewählte Person a im Test g erhalten hat. Die Gesamtpunktzahl einer Person ist dabei, wie bereits erwähnt, zufällig veränderlich und lässt sich hinsichtlich der einzelnen Items in die Teilbereiche eines wahren Wertes und eines zufälligen Fehlers diskriminieren. X_{ga} ist also eine Zufallsvariable, die für die Population P definiert ist. Die Variable ist zum einem zufällig, da man die Fähigkeiten einer zufällig gezogenen Person nicht kennt und zum anderen fluktuieren auch die Fähigkeiten und Ansichten innerhalb einer Person von Zeitpunkt zu Zeitpunkt (Mortensen 2009: 16). Die Beobachtungswerte für die unterschiedlichen Personen der Population P werden mit x_{ga} definiert. Wir gehen davon aus, dass sich das Testergebnis aus zwei Teilen ergibt. Der erste Teil entspricht den konstanten Fähigkeiten einer Person und der zweite Teil entspricht einer zufälligen Störgröße bzw. einem zufälligen (unsystematischen) Fehler. Eine Differenzierung zwischen systematischen und unsystematischen Fehlern erfolgte bereits unter Punkt 2.4.

Bei der klassischen Testtheorie wird der erwartete beobachtete Wert als wahrer Wert τ angenommen. Der Erwartungswert ergibt sich aus der Wahrscheinlichkeitsverteilung einer Person hinsichtlich eines Tests bzw. Testitems g:

$$\tau_{ga} = E(X_{ga}). \tag{1}$$

Der Erwartungswert einer zufälligen veränderlichen Variable X_{ga} entspricht somit dem wahren Wert bei Person a hinsichtlich der Messung g. Es ist der Mittelwert über unendlich viele beobachtete Testergebnisse einer Person.

Die folgenden Ausführungen beziehen sich stets auf einen Test g $\in$ G und auf eine zufällig ausgewählte Person a $\in$ P, so dass zur Vereinfachung die Indizes g und a weggelassen werden können. ($X_{ga} = X$, a $\in$ P zufällig). Da X eine Zufallsvariable ist, kommt es zu Abweichungen zwischen dem beobachteten Messwert und dem wahren Wert τ. Die Abweichung wird wie folgt definiert: $\epsilon = X - \tau$. Der Ausdruck ϵ wird als Fehler bezeichnet, spezifischer als unsystematischer Fehler. Ein systematischer und bekannter Fehler, ein Bias, stellt für das stochastische Modell keine Schwierigkeiten dar, da der Bias in den wahren Wert τ integriert werden kann. Schwierigkeiten ergeben sich erst, wenn der systematische Fehler unbekannt ist und es dadurch zu verzerrten Ergebnissen kommen kann.
Der zufällige Testwert (Score) für eine zufällige Person ergibt sich somit, wie bereits dargestellt, aus dem wahren Wert und einem unsystematischen zufälligen Fehler.

Dies kann wie folgt formuliert werden:

$$X = \tau + \epsilon \qquad \textit{über} \qquad P. \tag{2}$$

Diese Grundannahme gilt dabei für eine zufällige Person, so dass alle drei Größen der Gleichung stochastisch sind. Es sind also X, τ und ϵ stochastische Variablen in einen Wahrscheinlichkeitsraum. Wird bei einem Zufallsexperiment die Person a $\in$ P gezogen, so nimmt τ den Wert τ_a und ist somit eine Konstante für die Person a.

Aus der formulierten Grundannahme ergeben sich verschiedene Folgerungen. Sie werden auch als Axiome der klassischen Testtheorie (KTT) bezeichnet, da sie von Gulliksen (1950) ursprünglich als Axiome eingeführt wurden. Tatsächlich ergeben sich die folgenden Folgerungen aus der Grundannahme und sind dadurch keine Axiome. Es wird angenommen, dass zwei Tests mit den Scores X und Y vorliegen. Die Kovarianzen zwischen X, τ und ϵ sowie ϵ_x und ϵ_y werden mit $Cov(\tau, \epsilon), Cov(\epsilon_x, \tau_y), Cov(\epsilon_x, \epsilon_y)$ bezeichnet. Die Variablen ϵ_x und ϵ_y sind dabei die zufälligen Fehler für zwei unterschiedliche Messungen. Die Folgerungen lauten wie folgt:

$$E(\epsilon) = 0 \tag{3}$$

$$Cov(\tau, \epsilon) = 0 \tag{4}$$

$$Cov(\epsilon_x, \tau_y) = 0 \tag{5}$$

$$Cov(\epsilon_x, \epsilon_y) = 0 \tag{6}$$

Diese Folgerungen gelten für jede von null verschiedenen Subpopulationen von P. Es wird angenommen, dass
$0 < \sigma_X^2 < \infty, \quad 0 < \sigma_\tau^2 < \infty, \quad 0 < \sigma_\epsilon^2 < \infty$. Dabei bezeichnet σ_X^2 die Varianz des beobachten Wertes, σ_τ^2 die Varianz des wahren Wertes und σ_ϵ^2 die Varianz der Störvariable. Für die Formel 6 wird angenommen, dass zwischen den Scores X und Y Unabhängigkeit besteht. Generell werden die Formeln 3 bis 6 auch als Messfehlertheorie bezeichnet.

Gegeben seien also zwei Tests mit den Scores X und Y, mit $X = \tau_x + \epsilon_x$ und $Y = \tau_y + \epsilon_y$ mit $E(X) = E(\tau_x)$ und $E(Y) = E(\tau_y)$. Es gilt der Zusammenhang

$$Cov(X, Y) = Cov(\tau_x, \tau_y) \tag{7}$$

Dies folgt aus der Grundannahme der klassischen Testtheorie und aus den Folgerungen hieraus:
$Cov(X,Y) = E(XY) - E(X)E(Y)$
$= E[(\tau_x + \epsilon_x)(\tau_y + \epsilon_y)] - E(\tau_x)E(\tau_y)$
$= E(\tau_x\tau_y) + E(\tau_x\epsilon_y) + E(\tau_y\epsilon_x) + E(\epsilon_x\epsilon_y) - E(\tau_x)E(\tau_y)$
$= E(\tau_x\tau_y) - E(\tau_x)E(\tau_y)$
$= Cov(\tau_x, \tau_y)$, denn es gilt: $E(\tau_x\epsilon_y) + E(\tau_y\epsilon_x) + E(\epsilon_x\epsilon_y) = 0$.
Dieser Zusammenhang wird bei der Herleitung von Cronbachs α benötigt (vgl. Kapitel 4.4).

Inhaltliche Interpretation der Folgerungen aus der Grundannahme:

1. Formel 3 folgt direkt aus der Grundannahme, da $E(\epsilon) = E(X) - E(\tau) = \tau - \tau = 0$. Der Ausdruck $E(\epsilon)$ entspricht dem durchschnittlichen (unsystematischen) Messfehler. Es ist der Durchschnitt aller Messfehler der Personen aus P. Der Messfehler hat bei (hypothetisch) genügend oft wiederholten Messungen im Durchschnitt den Wert 0. Aus dieser Aussage kann man nicht schließen, dass kein Messfehler vorliegt. Es kann sein, dass unter bestimmten Mess- bzw. Testbedingungen ein systematischer Fehler (Bias) vorliegt. Dies ist unproblematisch, soweit dieser systematische Fehler konstant bei allen Personen vorliegt. In diesem Fall sei τ_0 der konstante Bias. Es gilt dann $E(X_a) = \tau_a + \tau_0$. Wenn man nun τ_a in $\tau_a + \tau_0$ umbenennt, dann „verschwindet“ der Fehler in τ_a. Der Fehler der Messungen des klassischen Modells ist also per Definition unverzerrt. Liegen z.B. Bedingungen wie schlechtes Licht oder Lärm vor, dann wird dieser systematische Fehler im wahren Wert korrigiert und der Fehler bleibt unverzerrt. Die Schwierigkeit besteht darin, dass ein systematischer Fehler nur aus inhaltlichen Überlegungen erkannt werden kann. Anderweitig „schlummert“ dieser Fehler zusätzlich im unsystematischen Fehler, was zu verzerrten Ergebnissen führen kann. Wird beispielsweise ein Test zur Reaktionsschnelligkeit durchgeführt und einem Teil der Stichprobe zuvor vorgeschrieben einen Liter Bier zu trinken, dann wird der Test bei dieser Personengruppe um einen bestimmen Betrag $\tau_0 < 0$ schlechter ausfallen, als bei den Personen, die nicht unter Alkoholeinfluss stehen.

2. Die Folgerung nach der Formel 4 gilt, da $Cov(\tau\epsilon) = E(\epsilon\tau) - E(\epsilon)E(\tau) = E(\epsilon\tau)$, da $E(\epsilon) = 0$. Zudem folgt aus der Grundannahme, dass $E(\epsilon\tau) = E[(X - \tau)\tau] = E(X\tau) - \tau^2 = \tau E(X) - \tau^2 = \tau^2 - \tau^2 = 0$.
Es ergibt sich also, dass die Korrelation zwischen dem wahren Wert und dem Fehler null ist. Diese Aussage bringt mit sich, dass über den gesamten Wertebereich des

wahren Werts die Fehler gleich sind, so dass keine Kovarianz zwischen dem wahren Wert und dem Fehler vorliegt. Beispielsweise müsste also der Messfehler bei höherer und niedriger Intelligenz immer gleich sein. Tatsächlich könnte es jedoch sein, dass eine höhere Intelligenz mit sich bringt, dass man in seinen Leistungen nicht so stark schwankt und dadurch der Messfehler kleiner ist als bei Personen mit geringerer Intelligenz. Bestätigt sich dies auch empirisch, dann gilt $Cov(\tau, \epsilon) = 0$ nicht mehr.

3. Neben des Tests X wird davon ausgegangen, dass noch ein zweiter Test Y durchgeführt wird. Die Aussage $Cov(\epsilon_x, \tau_y) = 0$ zeigt sich folgendermaßen:
$Cov(\epsilon_x, \tau_y) = E(\epsilon_x\tau_y) - E(\epsilon_x)E(\tau_y) = E(\epsilon_x\tau_y)$. Es ist $E(\epsilon_x\tau_y) = \tau_y E(\epsilon_x) = 0$, da aus der Grundannahme gilt $E(\epsilon_x) = 0$. Es gilt somit auch der Zusammenhang $Cov(\epsilon_x, \tau_y) = 0$. Inhaltlich bedeutet die Formel 5, dass die Korrelation zwischen dem Fehler einer Messung und dem wahren Wert einer anderen Messung null ist, was man auch intuitiv annehmen würde.

4. Die Formel 6 ist eine direkte Folge der angenommenen Unabhängigkeit zwischen X und Y. Für die beiden Tests wird angenommen, dass die Tests an sich sowie die Fehler unkorreliert sind.

Aus der Grundannahme der klassischen Testtheorie und deren Folgerungen kann der Begriff der Reliabilität definiert werden.

4 Reliabilität

Wie bereits beschrieben, handelt es sich bei der Reliabilität um ein Gütekriterium für Messungen. Die Reliabilität steht im Fokus dieser Arbeit. Hierzu erfolgt zunächst eine allgemeine Definition, ehe die Reliabilität bei parallelen und zusammengesetzten Tests betrachtet wird. Schließlich erfolgt eine Herleitung des Koeffizienten α nach Cronbach (1951).

4.1 Allgemeine Definition

Durch die Grundannahme der klassischen Testtheorie und deren Folgerungen kann die Reliabilität eines Test hergeleitet werden. Es wird im Folgenden von Annahmen der klassischen Testtheorie gesprochen, obwohl es um eine Grundannahme und deren Folgerungen handelt. Dies erfolgt um die Darstellungen verbal zu vereinfachen. Die folgenden Ausführungen beziehen sich auf Bühner (2014: 115 ff.), Steyer und

Eid (1993: 104 ff.), Lord (1974: 56 ff.) und Mortensen (2009: 59 ff.).

Unter Reliabilität im Kontext der klassischen Testtheorie ist generell ein Varianzverhältnis zu verstehen, das im Folgenden näher ausgeführt wird. Aus der klassischen Testtheorie folgt, dass die Varianz der beobachteten Werte gleich der Summe der Varianzen aus dem wahren Werten und der Fehlervariablen ist:
$\sigma_x^2 = \sigma^2(\tau + \epsilon) = \sigma_\tau^2 + \sigma_\epsilon^2 + 2Cov(\tau, \epsilon)$. Da $Cov(\tau, \epsilon) = 0$ gilt, ergibt sich folgende verkürzte Varianz: $\sigma_x^2 = \sigma_\tau^2 + \sigma_\epsilon^2$.
Die Gleichung 2 ($X = \tau + \epsilon$) ist eine lineare Regressionsgleichung. Möchte man diese Gleichung in die Form der üblichen linearen Regressionsgleichung überführen, dann ergibt sich: $x = \beta\tau + \alpha + \epsilon$. Für die Gleichung 2 gilt, $\alpha = 0$ und $\beta = 1$. Es ist das Verhältnis von der Varianz der beobachteten Werte und der Varianz der wahren Werte von Interesse. Dieses Verhältnis kann mit der quadrierten Korrelation (ρ^2) zwischen diesen Werten beschrieben werden, es gilt: $\rho_{x,\tau}^2 := \frac{\sigma^2(x,\tau)}{\sigma_x^2\sigma_\tau^2}$. Ferner gilt für die Standardabweichung zwischen dem beobachteten Wert und dem wahren Wert: $\sigma(X, \tau) = \sigma_\tau^2$. Dies lässt sich aus folgendem Zusammenhang begründen: $\sigma(X, \tau) = E(X\tau) - E(X)E(\tau) = E[(\tau + \epsilon)\tau] - E(\tau + \epsilon)E(\tau) = E(\tau^2) + E(\epsilon\tau) - (E(\tau)^2 - E(\epsilon)E(\tau)) = \sigma_\tau^2 + E(\epsilon\tau) - E(\epsilon)E(\tau)$. Da $E(\epsilon) = 0$ und $\sigma(\tau, \epsilon) = 0$ gilt, ergibt sich $\sigma(x, \tau) = \sigma_\tau^2$. Hieraus folgt:

$$\rho_{x,\tau}^2 = \frac{\sigma_\tau^2\sigma_\tau^2}{\sigma_x^2\sigma_\tau^2} \tag{8}$$

Durch einfaches Kürzen der Formel ergibt sich:

$$\rho_{x,\tau}^2 = \frac{\sigma_\tau^2}{\sigma_x^2} \tag{9}$$

Somit entspricht die quadrierte Korrelation zwischen dem beobachteten Wert und dem wahren Wert dem Verhältnis von der Varianz des wahren Wertes und des beobachteten Wertes. Hierfür ist die lineare Form der Gleichung Voraussetzung. Formel 9 kann auch in folgende Formen umgewandelt werden:

$$\rho_{x,\tau}^2 = 1 - (\sigma_\epsilon^2/\sigma_x^2) \tag{10}$$

$$\rho_{x,\epsilon}^2 = \sigma_\epsilon^2/\sigma_x^2. \tag{11}$$

Hieraus kann gefolgert werden, dass $\rho_{x,\tau}^2 + \rho_{x,\epsilon}^2 = 1$. Der definierte Ausdruck $\rho_{x,\tau}^2$ gibt an, was man intuitiv im Kontext der KTT unter Reliabilität versteht, nämlich das Verhältnis von der Fehlervarianz zur Gesamtvarianz. Für $\sigma_\epsilon^2 \to 0$ folgt, dass

$\rho^2_{x,\tau} \rightarrow 1$. In diesem Fall wäre die Messgenauigkeit maximal, da kein Messfehler besteht. Unter Reliabilität definieren wir also:

$$Rel(X) = \rho^2_{x,\tau} = \frac{\sigma^2_\tau}{\sigma^2_x} \tag{12}$$

Eine Berechnung der Reliabilität ist jedoch durch diese Formel nicht möglich, da der wahre Wert und die Varianz des wahren Wertes nicht bekannt sind.

4.2 Reliabilität bei parallelen Tests

Es geht nun darum, einen Weg zu finden, die Reliabilität bestimmen zu können. Eine Möglichkeit besteht darin, parallele Tests anzunehmen. Parallele Tests erfassen jeweils dasselbe zu erfassende Merkmal der Testpersonen. Die intuitive Idee ist, dass die Reliabilität der Korrelation zwischen zwei parallelen Tests entspricht. Wir nehmen an, dass wir die Messwerte X und X$'$ von zwei parallelen Tests haben. Mit diesen Tests sollen die gleichen Merkmale gemessen werden.
Diese Grundidee soll im Rahmen dieses Abschnittes formalisiert werden. Es wird angenommen, dass die Messungen X und X$'$ vorliegen, so dass

$$X = \tau + \epsilon, \quad X' = \tau' + \epsilon' \tag{13}$$

in jeder von Null verschiedenen Subpopulation von P gilt. Es gilt: $E(X) = E(\tau) = \tau_x$, $E(X') = E(\tau') = \tau_{x'}$, $E(\epsilon_x) = 0$ und $E(\epsilon_{x'}) = 0$. Falls nun des Weiteren angenommen wird, dass $\tau = \tau'$ und $\sigma^2_\epsilon = \sigma^2_{\epsilon'}$, dann liegen parallele Tests vor.

Sei nun $\rho(X, X') = \rho_{xx'} = \frac{Cov(X,X')}{\sigma_x \sigma_{x'}}$, $Cov(X, X') = E(XX') - E(X)E(X'), \sigma^2_x = Var(X) = Var(\tau) + Var(\epsilon_x)$ und $\sigma^2_{x'}(X') = Var(X') = Var(\tau') + Var(\epsilon_{x'})$. Es gilt $Cov(X, X') = E[(\tau + \epsilon_x - \tau_x)(\tau' + \epsilon_{x'} - \tau_{x'})] = E(\tau\tau') - \tau_x\tau_{x'}$. Dies ergibt sich durch einfaches Ausmultiplizieren und unter Berücksichtigung der Annahmen $E(\epsilon) = 0$ und $E(\epsilon') = 0$. Aufgrund der aufgeführten Bedingungen von parallelen Tests ergibt sich $Cov(XX') = E(\tau^2) - E^2(\tau) = Var(\tau)$ und $\sigma_x\sigma_{x'} = \sigma^2_x = Var(X)$. Es gilt dadurch:

$$\rho_{xx'} = \frac{Var(\tau)}{Var(X)} = \rho^2_{x\tau} = Rel(X) \tag{14}$$

Theoretisch gehen wir hierdurch davon aus, dass der wahre und der beobachtete Wert gleichzeitig beobachtet werden. Es liegt kein zeitlicher Verzug zwischen den Messungen vor (Yang et al.: 2011: 378). Die Annahme von Parallelmessungen ist jedoch sehr radikal, da sie sehr streng ist und sich empirisch oft nicht bestätigt.

4.3 Reliabilität bei zusammengesetzten Tests

Im Mittelpunkt dieser Arbeit stehen zusammengesetzte Tests. Hierbei handelt es sich um Tests, die sich aus mehr als einem Item zusammensetzen. Hierbei wird sich zeigen, dass die Annahme von parallelen Tests nicht erforderlich ist, um die Reliabilität bestimmen zu können. Zusammengesetzte Tests, die ein theoretisches Konstrukt abbilden sollen, bestehen aus unterschiedlichen Aufgaben bzw. Items. Die Idee besteht nun darin, den inneren Zusammenhang der Items (interne Konsistenz) zu betrachten (Bühner 2004: 118). Der Test wird dabei in so viele Untertests zerlegt, wie er Items hat. Die Berechnung erfolgt mit Hilfe von Korrelationen, Varianzen und Kovarianzen. Ziel ist es, angeben zu können, ob ein theoretisches Konstrukt durch die Items homogen abgebildet werden kann. Die Berechnung einer Maßzahl für die interne Konsistenz (hier: Cronbachs α) ist also nur dann sinnvoll, wenn homogene Merkmalsbereiche betrachtet werden sollen. Ob Homogenität hinsichtlich der Items besteht, wird jedoch nicht durch den Koeffizienten α abgebildet.

Grundsätzlich gibt es viele unterschiedliche Methoden der Reliabilitätsschätzung (Lienert et al: 1998: 180). Beispielsweise sind als Methoden der Testhalbierungskoeffizient, der Retestkoeffizient und die Paralleltestkoeffizient zu nennen. Die vorliegende Arbeit beschränkt sich auf den Koeffizienten α nach Cronbach (1951).

4.4 Koeffizient α

Es wird nun der Koeffizient α im Rahmen der Annahmen der klassischen Testtheorie hergeleitet. Zunächst ist es jedoch erforderlich eine Notation für zusammengesetzte Tests einzuführen, da diese durch den Koeffizienten betrachtet werden sollen (Lord et al. 1974: 85, Komaroff 1997: 337 ff., Mortensen 2009: 59 ff.). Die klassische Testtheorie kann in einem Triple dargestellt $(y_i, \tau_i, \epsilon_i)$ werden. Dabei ist y_i der beobachtete Wert, τ_i der wahre Wert und ϵ_i der Fehler des jeweiligen Items $i = 1, \ldots, n$. Wichtig ist hier zu verstehen, dass sich der Laufindex auf die Items $i = 1, \ldots, n$ bezieht und nicht auf die Personen, die an einem Test teilnehmen. Es werden nun zusammengesetzte Tests betrachtet, sodass es grundsätzlich möglich ist, den Testscore auf unterschiedliche Weise zu berechnen. Es können beispielsweise die Items unterschiedlich gewichtet in den Gesamtscore eingehen. Für die weiteren Betrachtungen nehmen wir an, dass die Scores der einzelnen Items ungewichtet in den Gesamtscore eingehen:

$$X = \sum_{i=1}^{n} y_i. \tag{15}$$

Dabei bildet y_i die Beobachtungswerte der einzelnen Items ab, es gilt: $y_i = \tau_i + \epsilon_i$, für $i = 1, \ldots, n$. In anderen Worten liegen hier also $i = 1, \ldots, n$ Tests vor. Der zusammengefasste Testscore wird des weiteren als X bezeichnet.
Der zusammengesetzte wahre Wert und der zusammengesetzte Fehlerwert werden entsprechend definiert:

$$T = \sum_{i=1}^{n} \tau_i \tag{16}$$

und

$$E = \sum_{i=1}^{n} \epsilon_i. \tag{17}$$

Für das Triple (X,T,E) gilt ebenfalls der im Rahmen der klassischen Testtheorie formulierte lineare Zusammenhang: X = T + E.
Gegeben seien nun die Messungen $y_1, y_2, \ldots, y_n$ mit den wahren Werten $\tau_1, \tau_2, \ldots, \tau_n$ und $X = y_1 + y_2, \ldots, +y_n$. Aus der Grundannahme der klassischen Testtheorie, dass sich ein Itemscore aus einem wahren Wert und einem Fehler zusammensetzt und aus der Idee zu den parallelen Messungen ergibt sich folgendes Maß für die Reliabilität:

$$\rho_{xx'} := \rho^2_{x,\tau} = \frac{\sigma^2(\tau)}{\sigma^2(X)}, \quad mit \quad X = \sum_{i=1}^{n} y_i, \quad und \quad i = 1, \ldots, n \tag{18}$$

Es geht im Folgenden darum, eine untere Schranke für die Reliabilität $\rho_{xx'}$ zu beweisen, da die τ-Werte nicht bekannt sind. Falls essentiell τ-äquivalente Messungen vorliegen, dann entspricht der Koeffizient α der wahren Reliabilität (vgl. Kapitel 4.5). Diese untere Schranke für die Reliabilität wird als α bezeichnet und hat folgende Form:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma^2(y_i)}{\sigma^2(X)}\right) \leq \rho_{xx'}. \tag{19}$$

Beweis:
Um die Formel 19 zu beweisen, werden die Formel $\rho^2_{x,\tau} = \frac{\sigma^2(\tau)}{\sigma^2(X)}$ und die Annahmen aus der klassischen Testtheorie (insbesondere zur Varianz) benötigt, um dies mit der Cauchy-Schwartz-Ungleichung zu kombinieren. Hieraus kann eine untere Schranke für die Reliabilität hergeleitet werden, sodass gilt $\alpha \leq \rho_{xx'}$.
Hierzu definieren wir $\sigma^2(\tau_i) = Var(\tau_i)$ und $\sigma_{ij} = Cov(\tau_i, \tau_j)$.
Die zusammengesetzte Varianz des wahren Wertes τ, des beobachteten Wertes X und des Fehlers ϵ ergeben sich jeweils aus der Varianzsumme der einzelnen Items

und den entsprechenden Kovarianzen:

$$\sigma^2(T) = \sum_{j=1}^{n} \sigma^2(\tau_i) + \sum_{i \neq j} \sigma(\tau_i, \tau_j) \tag{20}$$

$$\sigma^2(X) = \sum_{i=1}^{n} \sigma^2(y_i) + \sum_{i \neq j} \sigma(y_i, y_j) \tag{21}$$

$$\sigma^2(E) = \sum_{i=1}^{n} \sigma^2(\epsilon_i) + \sum_{i \neq j} \sigma(\epsilon_i, \epsilon_j) \tag{22}$$

Es gilt der Zusammenhang: $\sigma^2(X) = \sigma^2(T) + \sigma^2(E)$. Die Kovarianzen in den Gleichungen 20 und 21 entsprechen sich, da angenommen wird: $\sigma(y_i, y_j) = \sigma(\tau_i, \tau_j) = \sigma_{ij}$. Dies folgt aus der Annahme, dass die Fehler unkorreliert sind. Da $\sum\limits_{i \neq j} \sigma(\epsilon_i, \epsilon_j) = 0$ angenommen wird, verkürzt sich $\sigma^2(X)$ auf die unter Gleichung 21 angegebene Form. Es gilt allgemein die Cauchy-Schwarz-Ungleichung (Amann et al. 2006: 167 ff):

$$2\sigma_{ij} \leq \sigma^2(\tau_i) + \sigma^2(\tau_j), i \neq j \tag{23}$$

Aufgrund dessen kann eine untere Schranke für die Reliabilität definiert werden. Es kann aus der Formel 23 gefolgert werden, dass

$$2 \sum_{i \neq j} \sigma_{ij} \leq \sum_{i \neq j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) \tag{24}$$

$\sum\limits_{i \neq j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) = 2(n-1) \sum\limits_{i=1}^{n} \sigma^2(\tau_i)$, da gilt

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) = \sum_{i=j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) + \sum_{i \neq j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) \tag{25}$$

und dies bedeutet

$$2n \sum_{i=1}^{n} \sigma^2(\tau_i) = 2 \sum_{i=1}^{n} \sigma^2(\tau_i) + \sum_{i \neq j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)), \tag{26}$$

denn $\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) = n \sum\limits_{i=1}^{n} \sigma^2(\tau_i) + n \sum\limits_{j=1}^{n} \sigma^2(\tau_j) = 2n \sum\limits_{i=1}^{n} \sigma^2(\tau_i)$, da $\sum\limits_{i=1}^{n} \sigma^2(\tau_i) = \sum\limits_{j=1}^{n} \sigma^2(\tau_j)$ und $\sum\limits_{i=j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) = \sum\limits_{i=1}^{n} (\sigma^2(\tau_i) + \sigma^2(\tau_j) = 2 \sum\limits_{i=1}^{n} \sigma^2(\tau_i))$.
Es gilt dadurch:
$\sum\limits_{i \neq j} (\sigma^2(\tau_i) + \sigma^2(\tau_j)) = 2n \sum\limits_{i=1}^{n} \sigma_i^2 - 2 \sum\limits_{i=1}^{n} \sigma_i^2 = 2(n-1) \sum\limits_{i=1}^{n} \sigma_i^2$.

Hieraus folgt, dass

$$2(n-1)\sum_{i=1}^{n}\sigma^2(\tau_i) \geq 2\sum_{i\neq j}\sigma_{ij} \tag{27}$$

und dies bedeutet

$$\sum_{i=1}^{n}\sigma^2(\tau_i) \geq \frac{1}{n-1}\sum_{i\neq j}\sigma_{ij}. \tag{28}$$

Hieraus kann gefolgert werden:

$$\sigma^2(\tau) = \sum_{i=1}^{n}\sigma^2(\tau_i) + \sum_{i\neq j}\sigma_{ij} \geq \frac{1}{n-1}\sum_{i\neq j}\sigma_{ij} + \sum_{i\neq j}\sigma_{ij} = \frac{n}{n-1}\sum_{i\neq j}\sigma_{ij}. \tag{29}$$

Ferner besteht folgender Zusammenhang:

$$\sigma^2(X) - \sum_{i=1}^{n}\sigma^2(y_i) = \sum_{i\neq j}\sigma_{ij} \tag{30}$$

so dass,

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X)} \geq \frac{n}{n-1}\frac{\sigma^2(X) - \sum\limits_{i=1}^{n}\sigma^2(y_i)}{\sigma^2(X)} = \frac{n}{n-1}\Big(1 - \frac{\sum\limits_{i=1}^{n}\sigma^2(y_i)}{\sigma^2(X)}\Big). \tag{31}$$

Es ergibt sich hieraus folgende Formel zur Bestimmung der unteren Schranke für die Reliabilität:

$$\alpha = \frac{n}{n-1}\Big(1 - \frac{\sum\limits_{i=1}^{n}\sigma^2(y_i)}{\sigma^2(X)}\Big) \leq \rho_{xx'}. \tag{32}$$

Die Formel 32 kann auch folgendermaßen dargestellt werden (Stanley 1971):
$\sigma^2(X) = \sum\limits_{i=1}^{n}\sigma^2(y_i) + \sum\limits_{i\neq j}\sigma_{ij} = \sum\limits_{i=1}^{n}\sigma^2(y_i) + \sum\limits_{i\neq j}\rho_{ij}\sigma_i\sigma_j.$
Dabei ist ρ_{ij} die Korrelation zwischen den Testscore y_i und y_j. Es folgt:
$\sum\limits_{i=1}^{n}\sigma^2(Y_i) = \sigma^2(X) - \sum\limits_{i\neq j}p_{ij}\sigma_i\sigma_j.$
Der Koeffizient α kann dadurch in folgende Formen gebracht werden:

$$\alpha = \frac{n}{n-1}\Big(\frac{\sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j}{\sum\limits_{i=1}^{n}\sigma^2(y_i) + \sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j}\Big) \quad bzw. \quad \alpha = \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i,y_j)}{\sigma^2(X)}\Big) \tag{33}$$

Der Wertebereich von α ist wie folgt definiert: $\alpha \in [0;1]$. Um den Koeffizienten besser zu verstehen, werden die Intervallgrenzen nun näher betrachtet.

Perfekte Heterogenität

Bei perfekter Heterogenität sind die Messwerte y_i unabhängig voneinander, sodass

$p_{ij} = 0$ für alle $i \neq j$ gilt. In diesem Fall ist $\alpha = 0$, denn $\rho_{xx'} \geq 0$ gilt immer. Allgemein kann man sagen: Je geringer die Testitems untereinander korrelieren, desto kleiner wird der Wert von α sein. Falls die Messungen betragsmäßig schwach korrelieren, dann werden auch rein intuitiv betrachtet unterschiedliche latente Eigenschaften oder Einstellungen gemessen und dies schlägt sich in einer geringen Reliabilität des Gesamttests nieder.

Perfekte Homogenität
Es wird nun der Fall von $p_{ij} = 1$ betrachtet. Die Formel 32 kann dann wie folgt geschrieben werden:

$$\alpha = \frac{n}{n-1}\Big(\frac{1}{1 + \sum\limits_{i=1}^{n} \sigma^2(y_i) / \sum_{i \neq j} \sigma_i \sigma_j}\Big) \tag{34}$$

Es zeigt sich hier, dass α von der Varianz $\sigma^2(y_i)$ abhängt. Betrachtet man zusätzlich den Spezialfall der Varianzgleichheit der Items ($\sigma^2(y_i) = \sigma^2$), dann lässt sich zeigen, dass $\alpha = 1$ ist. Falls die Tests bzw. Testitems perfekt miteinander korrelieren und gleichzeitig die Varianzen identisch sind, dann ergibt sich eine perfekte Reliabilität des Gesamttests.

4.5 Essentiell τ-äquivalente Messungen

Unter der Voraussetzung, dass Messungen mindestens essentiell τ-äquivalent sind, ist der Koeffizient α nicht nur eine untere Schranke für die Reliabilität, sondern vielmehr entspricht der Koeffizient der wahren Reliabilität. Hierfür muss zuerst definiert werden, was unter essentieller τ-Äquivalenz zu verstehen ist (Komaroff 1997: 338, Kristof 1974, Zimmerman et al. 1993, Traub 1994, Lord et al. 1974: 50, 90). Wir nehmen an, dass für die wahren Werte der Items $(\tau_i, \tau_j)(i \neq j; i, j = 1, 2, \ldots, n)$ folgender linearer Zusammenhang besteht:

$$\tau_i = a_{ij} + b_{ij}\tau_j \tag{35}$$

Die Indizes i und j beziehen sich wiederum auf die Items und nicht auf die Personen. Dabei sind a_{ij} und b_{ij} Konstanten, die paarweise differieren können oder von den speziellen Paaren τ_i und τ_j abhängen. Nun folgen einige Definitionen von Begrifflichkeiten. Es gilt dabei immer die Annahme aus der klassischen Testtheorie von unkorrelierten Fehlern $\sigma(\epsilon_i, \epsilon_j) = 0$.

1. Kongenerische Messungen
Es liegen kongenerische Messungen vor, wenn $b_{ij} \neq 1$, ausgenommen der Fall $b_{ij} = 0$. Es ist dabei unerheblich, welchen Wert a_{ij} annimmt (Jöreskog: 1974). Die Varianz der Messungen und die Kovarianz zwischen den Itempaaren entsprechen sich dabei nicht ($\sigma^2(\tau_i) \neq \sigma^2(\tau_j) \neq \sigma(\tau_i, \tau_j)$).
2. Essentiell τ-äquivalente Messungen
Die Messungen sind essentiell τ-äquivalent, falls $a_{ij} \neq 0$ und $b_{ij} = 1$ gilt. Eine lineare Transformation unter diesen Bedingungen führt nicht zu unterschiedlichen Varianzen bei den Items und die Kovarianz zwischen Itempaaren entspricht dadurch auch der Varianz der einzelnen Items ($\sigma^2(\tau_i) = \sigma^2(\tau_j) = \sigma(\tau_i, \tau_j)$). Dies gilt ebenfalls für τ-äquivalente und parallele wahre Werte.
3. τ-äquivalente Messungen
Die zusätzliche Annahme zur essentiellen τ-Äquivalenz ist, dass $a_{ij} = 0$.
4. Parallele Messungen
Die zusätzliche Annahme zu den τ-äquivalenten wahren Werten ist, dass $\sigma^2(\epsilon_i) = \sigma^2(\epsilon_j)$.
Es ist eine notwendige und hinreichende Bedingung, dass die Messwerte mindestens essentiell τ-äquivalent sind (Lord 1974: 90, Osburn 2000: 345). In diesem Fall gilt: $\alpha = \rho_{xx'}$. Die Gleichung 35 verkürzt sich somit auf $\tau_i = a_{ij} + \tau_j$.
Im Beweis von Cronbach α wurde die Cauchy-Schwarz-Ungleichung angewendet. Es wurde verwendet, dass $[\sigma(\tau_i) - \sigma(\tau_j)]^2 \geq 0$ und $\mid \sigma(\tau_i, \tau_j) \mid \geq \sigma(\tau_i, \tau_j)$. Falls nun $\beta_{ij} = 1$ und $\sigma(\tau_i) = \sigma(\tau_j)$, dann gilt: $[\sigma(\tau_i) - \sigma(\tau_j)]^2 = 0$ und $\mid \sigma(\tau_i, \tau_j) \mid = 0$. Es ergibt sich hieraus: $\alpha = \rho_{xx'}$. Falls hingegen kongenerische Werte vorliegen, dann gilt $\alpha < \rho_{xx'}$. Es liegt in diesem Fall eine Unterschätzung der Reliabilität vor.
Falls mindestens essentiell τ-äquivalente Messungen vorliegen, bilden die verwendeten Items eindimensional das interessierende theoretische Konstrukt ab (Green 2009: 124). Es wird dadurch ausgeschlossen, dass multidimensionale Faktoren die Skala beeinflussen. Der Koeffizient α trifft jedoch keine Aussagen, ob eine Skala tatsächlich eindimensional ist oder nicht (Waller 2008: 211, Cortina 1993, Green et al. 1977).

5 Vorteile von Koeffizient α

Die Beliebt- und Berühmtheit des Koeffizienten α kann man allein dadurch erkennen, dass bereits über 6.500 Zitate von dem Artikel von Cronbach (1951) erfolgten (Sijtsma 2009: 108). Auf der Website von Psychometrika ist es des Weiteren eines der beliebtesten Paper zum Download. Es handelt sich wohl aus folgenden Gründen

um eine derart berühmte statistische Maßzahl (Yang et al 2011: 377-378):

1.
Koeffizient α scheint einfach zu interpretieren zu sein. Desto näher der Koeffizient dem Wert 1 kommt, desto höher scheint die interne Konsistenz auszufallen. Die Eigenschaft, dass bei den allermeisten Fällen die essentielle τ-Äquivalenz der Messungen nicht gegeben ist, sollte nicht beunruhigen, da dann der Koeffizient in den meisten Fällen eine untere Schranke für die Reliabilität ist. Es kann jedoch auch Situationen geben, in denen der Koeffizient α nicht mehr eine untere Schranke für die Reliabilität ist. Die Schwierigkeiten der vermeintlich einfachen Interpretation werden im folgenden Kapitel näher dargestellt.

2.
Der Koeffizient α zeichnet sich des Weiteren durch einen hohen Grad der Objektivität aus. Die Forscherinnen und Forscher müssen keine subjektiven Entscheidungen treffen. Beispielweise muss beim Split-Half-Koeffizienten entschieden werden, wo die Menge der Beobachtungen aufgeteilt werden muss. Bei dem Test-Retest Koeffizienten muss die Entscheidung darüber getroffen werden, wann die zweite Welle der Daten erhoben werden soll. Erfolgt dies aus fundierten theoretischen Überlegungen, kann, nach Ansicht des Autors dieser Arbeit, jedenfalls nicht von einem subjektiven Element gesprochen werden. Vielmehr entwickeln sich hier die Testergebnisse theoriegeleitet. Derartige Schwierigkeiten ergeben sich beim Koeffizienten α erst überhaupt nicht.

3.
Des Weiteren kann der Koeffizient α für die Entscheidung über die Auswahl von Items verwendet werden (sog. $\alpha-$Maximierung). Beispielsweise können Items, die eine andere Dimension als die gewünschte messen, erkannt werden und aus der Analyse gestrichen werden. Hierdurch kann die Reliabilität gesteigert werden. Diese Vorgehensweise sieht der Autor dieser Arbeit jedoch als kritisch an und schlägt unter Kapitel 7 eine mögliche Alternative vor.

4.
Aufgrund der Popularität erscheint die Angabe des Koeffizienten in einer Publikation als obligatorisch. Kombiniert mit Regeln zur Interpretation des Koeffizienten erhalten die erzielten Ergebnisse Vergleichbarkeit und Transparenz.

6 Schwierigkeiten bei Koeffizient α

In diesem Kapitel sollen die Schwierigkeiten bei dem Koeffizienten α dargestellt werden. Zuerst wird die Abhängigkeit des Koeffizienten von der Anzahl der verwendeten Items beschrieben. Hierauf aufbauend wird Kritik an einer pauschalen Grenze zur Beurteilung der Reliabilität im verbalen Sinn geübt und ein Gegenvorschlag hierzu formuliert. Danach wird die Schwierigkeit, die sich durch alternierende Vorzeichen ergeben kann, dargestellt. Im darauf folgenden Abschnitt erfolgen Ausführungen über essentiell τ-äquivalente Messungen. Die beiden anschließenden Kapitel befassen sich mit der Korrelation von Fehlern bzw. mit der Korrelation zwischen dem wahren Wert und dem Fehler. Hierzu erfolgen analytische Ausführungen und die Darstellung eines Simulationsergebnisses.

6.1 Abhängigkeit von der Anzahl der Items

Die Formel zur Berechnung des Koeffizienten α bringt mit sich, dass der Koeffizient und somit auch die Reliabilität höher wird, desto mehr Items zur Beschreibung eines theoretischen Konstruktes verwendet werden (Mortensen 2009: 70 ff., Peterson 1994, Cho et al. 2014: 10 ff.). Dies soll im Folgenden veranschaulicht werden. Hierzu wird zuerst Folgendes definiert:

$$S = E(\sigma(y_i, y_j)) \qquad \overline{\sigma}^2 = E(Var(y_i)) \tag{36}$$

S und $\overline{\sigma}^2$ stellen dabei Erwartungswerte über alle möglichen Kovarianzen und Varianzen in der Population dar. Es wird angenommen, dass die Auswahl der Items zufällig aus einer Grundgesamtheit von Items erfolgt (vgl. hierzu Kapitel 8.1). Für eine Stichprobe von Items ergibt sich:

$$S_n = S + \triangle S_n \qquad \overline{\sigma}_n^2 = \overline{\sigma}^2 + \triangle \overline{\sigma}_n^2 \tag{37}$$

Die Mittelwerte der Kovarianz und der Varianz der Stichprobe $(S_n, \overline{\sigma}_n^2)$ ergeben sich somit aus den wahren Werten (S und $\overline{\sigma}^2$) und aus folgenden Differenzen: $\triangle S_n = S_n - S$, $\triangle \overline{\sigma}_n^2 = \overline{\sigma}_n^2 - \overline{\sigma}^2$. Für $\triangle S_n$ und $\triangle \overline{\sigma}_n^2$ gilt, dass diese Ausdrücke für $n \to \infty$ gegen 0 konvergieren.
Der Koeffizient α kann auch wie folgt dargestellt werden:

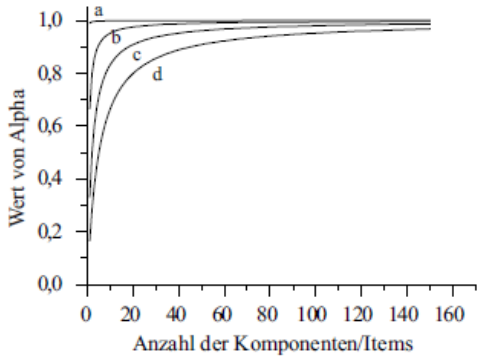
$$\alpha = \frac{1}{(q-1)/n + 1}, 1 \leq q \leq \infty \tag{38}$$

Dabei gilt: $q = \frac{\overline{\sigma}^2}{S}, \quad \overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2, \quad S = \frac{1}{n(n-1)}\sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j.$
Dies gilt, da es n(n-1)$\rho_{ij}\sigma_i\sigma_j$-Paare gibt. Man kann auch schreiben:

$$\alpha_n = \frac{1}{\frac{1}{n}(q_n - 1) + 1} \tag{39}$$

Es wird dabei α_n geschrieben, um die Abhängigkeit von n zu dokumentieren. Der Ausdruck $\lim_{n\to\infty}\frac{1}{n}\left(\frac{\overline{\sigma}^2+\triangle\overline{\sigma}_n^2}{S+\triangle S_n} - 1\right)$ konvergiert dabei gegen Null, so dass gilt: $\lim_{n\to\infty}\alpha = 1$. Dies gilt, da $\frac{1}{n}$ gegen Null konvergiert und $\triangle\overline{\sigma}^2$ sowie $\triangle S_n$ bei wachsendem n ebenfalls gegen Null konvergieren.
Es wurde von Mortensen (2009: 73) hierzu eine Simulation durchgeführt (vgl. Abbildung 1). In der Simulation wurden unterschiedliche q-Werte (vgl. Formel 38) verwendet. Ein hoher q-Wert bedeutet, dass der Anteil der Kovarianz zwischen den Items an der Gesamtvarianz gering ist. Es zeigt sich, dass die α-Werte bei geringeren q-Wert höher liegen, als bei höheren q-Wert. Bei der Simulation wurde nicht berücksichtigt, dass die q-Werte in der Empirie nicht konstant sind. Dadurch ist es auch möglich, dass der Koeffizient α bei einem zusätzlichen Item sogar abnimmt. Unter dieser Betrachtungsweise liegt auch keine monotone Folge der α-Werte gegen 1 vor.

Abbildung 1: Abhängigkeit von α von der Anzahl der Items
a: q = 1.01, b: q = 1.5, c: q = 3.0, d: q = 6
(Mortensen 2009: 73)



Nicht unerwähnt sollte die Meta-Analyse von Peterson (1994) bleiben. In einer Studie wurden 4.286 Alpha-Koeffizienten aus verschiedenen Veröffentlichungen betrachtet. Das Ergebnis hierzu lautet (Peterson 1994: 390): „On average, coefficient alpha does not appear to systematically increase once there are more than three in a sca-

le.“ Durchschnittlich wirkt sich also die Anzahl der Items nicht systematisch auf die Höhe des Koeffizienten aus. Bei der Analyse wurden die Koeffizienten α zwischen verschiedenen Studien betrachtet. Interessant wäre in diesem Zusammenhang auch, wie sich die Höhe der Koeffizienten innerhalb der publizierten Studien in Abhängigkeit der verwendeten Items darstellt. Hier würde sich wohl stärker eine Abhängigkeit von der Anzahl der verwendeten Items zeigen, als bei einer Betrachtung der bereits maximierten α-Werte der verschiedenen Veröffentlichungen.

6.2 Der Koeffizient α sollte größer sein als ...

Beginnen möchte ich dieses Kapitel mit einem Zitat von Cho et al. (2014: 12): „The nature of the decision being made on the basis of a test should be the guide for the acceptable level of reliability“. Die Definition einer Grenze, nach der eine Skala reliabel ist oder nicht, wäre nach dieser Betrachtungsweise nicht mehr erforderlich. In der Literatur wird jedoch meist eine Grenze für die Reliabilität von 0,7 bzw. 0,8 angegeben (Nunally 1978, Nunally et al. 1994, Bortz et al. 2006: 725, Cho et al. 2014: 12). Der Koeffizient α muss nach der Literatur je nach Betrachtungsweise alternativ die Kriterien $\alpha \geq 0,7$ bzw. $\alpha \geq 0,8$ erfüllen, damit eine Skala als reliabel gilt.
Dabei ist wichtig, dass mit dem Begriff der Reliabilität in diesem Kontext die Beurteilung gemeint ist, ob eine Skala als reliabel angesehen wird oder nicht. Dies trifft nach der Literatur zu, falls bestimmte Grenzen überschritten werden. Hiervon ist der Begriff der Reliabilität im Kontext der KTT zu unterscheiden, der ein Varianzverhältnis beschreibt. Es wird also von einer reliablen Skala gesprochen, falls durch den berechneten Koeffizienten α eine definierte Grenze überschritten wird. Der berechnete Koeffizient gibt mit Abweichungen die Reliabilität nach der KTT an.

Als erster definierte Nunally (1978) eine Grenze für einen akzeptablen α-Wert von 0,7. Dies erfolgte um den Forschern eine praktische Hilfe für die Interpretation des Koeffizienten zu geben. Die Definition der Grenze beruhte jedoch auf einer persönlichen Intuition von Nunally und fußte weder auf empirische Ergebnisse noch auf einer klaren logischen Begründung (Churchill et al. 1984, Cho et al. 2014). Die angegebene Grenze ist dadurch (relativ) willkürlich gewählt. Es könnten beispielsweise auch Grenzen von 0,69 oder 0,71 gewählt werden, ohne dass dies irgendeinen inhaltlichen Unterschied machen würde.
Cho et al. (2014: 11, auch Cortina 1993: 101) kritisieren in diesem Zusammenhang, dass die Forscher die beschriebenen Grenzen als ein zu erreichendes Ziel auffassen, das bei einer Einhaltung davon entbindet, sich weitere Gedanken über die Reliabilität zu machen. Dies ist jedoch erforderlich, da eine (zu) hohe Reliabilität nicht

wünschenswert ist. Eine derartige Skala wäre trivial und hätte keine inhaltliche Bedeutung (vgl. hierzu Kapitel 6.4).
Es ist dadurch wünschenswert, dass die Grenze für eine reliable Skala durch die Wichtigkeit der Reliabilität für den Forschungsprozess bestimmt wird. Cho (2014: 12) formuliert dies folgendermaßen: „When the importance of a decision made on the basis of a test score increases, the standard for reliability should also increase.“ Ähnlich äußert sich auch Cortina (1993: 101): „the finer the distinction that needs to be made, the better the reliability must be.“ Der Grenzwert sollte also in den Fällen zunehmen, in denen eine Skala für den Forschungsprozess wichtig ist, wobei berücksichtigt werden muss, dass eine zu hohe Reliabilität ebenfalls nicht wünschenswert ist. Eine fest definierte Grenze würde sich dadurch erübrigen. In der Empirie zeigt sich jedoch, dass meist eine Grenze von 0,7 gewählt wird, unabhängig vom Zweck einer Studie (Lance et al. 2006).
Bei der Anwendung einer universalen Grenze bleibt der Koeffizient natürlich einfach zu interpretieren und dies wurde bereits als ein Vorteil des Koeffizienten dargestellt. Die konkrete Situation kann hierdurch ggf. nicht adäquat abgebildet werden. Dieser „Preis“ wird jedoch von den meisten Forschern bezahlt.
Cortina (1993: 101) weist auf eine weitere Schwierigkeit bei der Definition einer allgemeinen Grenze hin. Der Koeffizient steigt mit der Anzahl der Items an (vgl. Kapitel 6.1). Es wird von Cortina dadurch vorgeschlagen, dass auch die Grenze nach oben korrigiert werden müsste, falls die Anzahl der Items erhöht wird. Der Verwendung von einer Vielzahl von Items nur um den Koeffizienten α zu erhöhen, könnte dadurch entgegengewirkt werden. Ein konkreter Vorschlag, wie dies geschehen sollte, erfolgte nicht.
Die Anzahl der Items sollte nach Ansicht des Autors mit einer adjustierten Grenze berücksichtigt werden.
Der Koeffizient α wird mit der Formel

$$\alpha = \frac{n}{n-1}\left(\frac{\sum\limits_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\right) \tag{40}$$

berechnet. Die Anzahl der Items geht im Wesentlichen durch den Faktor $\frac{n}{n-1}$ in die Berechnung des Koeffizienten ein. Es erscheint dadurch auch als sinnvoll, die Grenze für die Reliabilität auch mit diesem Faktor zu gewichten. Es handelt sich natürlich nur um eine Approximation, da die Anzahl der Items auch den zweiten Teil des Koeffizienten α durch geänderte Kovarianzen und Varianzen betrifft. Es wird daher

folgende Grenze θ vorgeschlagen:

$$\theta = \frac{n-1}{n} * \alpha_0 \tag{41}$$

Abbildung 2: Adjustierte α-Grenze

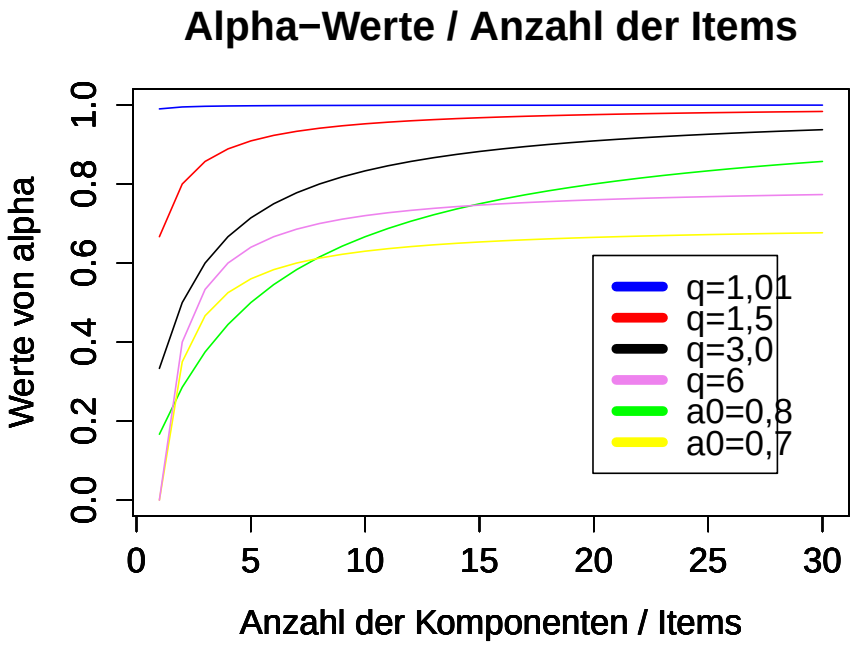


Tabelle 1: Adjustierte Grenze für die Reliabilität, $n = 1, \ldots, 10$

n	1	2	3	4	5	6	7	8	9	10
$\alpha_0 = 0{,}7$	0.35	0.47	0.53	0,56	0.58	0.60	0.61	0.62	0.63	0.64
$\alpha_0 = 0{,}8$	0.40	0.53	0.60	0.64	0.67	0.69	0.70	0.71	0.72	0.73

Tabelle 2: Adjustierte Grenze für die Reliabilität, $n = 11, \ldots, 20$

n	11	12	13	14	15	16	17	18	19	20
$\alpha_0 = 0{,}7$	0.64	0.65	0.65	0,65	0.66	0.66	0.66	0.66	0.67	0.67
$\alpha_0 = 0{,}8$	0.73	0.74	0.74	0.75	0.75	0.75	0.76	0.76	0.76	0.76

Dabei bezeichnet α_0 die gewünschte Grenze für die Reliabilität (z.B. 0,7, 0,8). Durch den gewählten Faktor $\frac{n-1}{n}$ steigt die Grenze durch die Zunahme der Items ähnlich, wie der Koeffizient α, ceteris paribus (vgl. Abbildung 2). In der Abbildung 2 wurden die in Abbildung 1 verwendeten Daten repliziert und ebenfalls die adjustierten Grenzen eingezeichnet. Die Zunahme des Koeffizienten durch zusätzliche Items kann

dadurch mit den adjustierten Grenzen für die Reliabilität berücksichtigt werden. Die von n (= Anzahl der Items) abhängigen Grenzen für $\alpha_0 = 0,7$ und 0,8 können beispielhaft den Tabellen 1 und 2 entnommen werden. Dabei wird die Grenze nicht erhöht, wie Cortina (1993: 101) vorgeschlagen hat, da dies zu einer trivialen Skala führen kann. Vielmehr wird die Grenze reduziert, falls eine geringere Anzahl von Items verwendet wird. Die Abbildung 2 zeigt, dass der Koeffizient bei konstanten Werten von q ansteigt. Würde man unabhängig von der Anzahl der Items immer die gleiche Grenze anwenden, dann würde man die Verwendung von weniger Items bestrafen. Dem kann durch die adjustierte Grenze entgegengewirkt werden. Eine Erhöhung der Itemanzahl zur Erhöhung des Koeffizienten wird durch dieses Verfahren entbehrlich.

6.3 Alternierende Vorzeichen

Aus den bisherigen Ausführungen könnte man annehmen, dass ein niedriger Wert von α eine geringe Korrelation ρ_{ij} zwischen den Items $i \neq j$ mit sich bringt. Dieser Schluss ist jedoch nicht zwingend (Mortensen 2009: 73 ff.). Betragsmäßig geringe Korrelationen zwischen den Testitems sind eine notwendige, jedoch keine hinreichende Bedingung für einen kleinen α-Wert. Ein Wert von $\alpha = 0$ würde sich auch ergeben, wenn eine Testhälfte eine perfekte positive Korrelation mit den Korrelationskoeffizienten $p_{ij} = 1$ und die andere Testhälfte eine perfekte negative Korrelation von $p_{ij} = -1$ aufweisen würde. Der Fall von $\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j \approx 0$ kann somit auch bei alternierenden Vorzeichen eintreten. Der beschriebene Extremfall kann für die Teilmengen M_1 und M_2 formal wie folgt aufgeschrieben werden:

$$\sum_{i,j \in M_1, i \neq j} \rho_{ij}\sigma_i\sigma_j \approx - \sum_{k,l \in M_2, k \neq l} \rho_{kl}\sigma_k\sigma_l \tag{42}$$

Dieser Schwierigkeit kann ggf. dadurch abgeholfen werden, indem die negativ korrelierten Items umformuliert werden, sodass eine positive Korrelation daraus entsteht. Negative Korrelationen zwischen zwei Items werden sich regelmäßig bei Tests zu Einstellungen ergeben, da sich die Antworten zu den Fragen ggf. stark beeinflussen. Bei Leistungstests wird diese Schwierigkeit regelmäßig nicht auftreten. Weiter kann die Kodierung der Antworten zu den Fragen bei der Datenanalyse auch umgedreht werden, so dass keine negativen Korrelationen in die Analyse eingehen.

6.4 Essentiell τ-äquivalente Messungen

Dieser Abschnitt befasst sich mit dem Thema, ob es realistisch ist, dass die Messungen zu den verwendeten Items mindestens τ-äquivalent sind. Hierdurch würde sich ergeben, dass der Koeffizient α der wahren Reliabilität entspricht. Es wird hierzu zuerst auf die Dimension einer Skala eingegangen. Weiter erfolgt eine Abgrenzung zwischen den Begriffen Homogenität, interne Konsistenz und Reliabilität. Abschließend werden die Ergebnisse einer Simulation dargestellt.

Es können nur essentiell τ-äquivalente Messungen vorliegen, falls eine unidimensionale Skala vorliegt. In der Empirie zeigt sich jedoch, dass meist keine essentiell τ-äquivalenten Messungen vorliegen, da meist multidimensionale Skalen vorliegen (Green 2009: 123). Liegt eine Skala vor, die durch mehrere Faktoren beeinflusst wird, dann kommt es regelmäßig zu einer Unterschätzung der Reliabilität.
Eindimensionale Skalen sind in der Psychometrie ein Ideal, das aber jedoch nur erreicht werden kann, falls sich die Items sehr ähnlich sind (Catell et al. 1964: 6). Die Items würden in diesem Fall nicht das gesamte Spektrum des interessierenden Merkmals abbilden und wären dadurch zu spezifisch. Provokanter formuliert dies Reise et. al (2007: 22), der das Erreichen der Unidimensionalität nur sieht, falls eine triviale Skala mit ähnlichen Fragen eine spezielle Eigenschaft misst. Eine unidimensionale Skala könnte nur einen Aspekt abbilden. Weiter spricht gegen die Unidimensionalität, dass wohl nicht ein Faktor die Antwort zu allen Items in gleicher Weise beeinflusst (Cortina 1993).

Für eine Skala wird daher mehr oder weniger ausgeschlossen, dass sie unidimensional ein theoretisches Konstrukt abbilden kann. Der Grad inwieweit Items unidimensional sind, wird auch als Homogenität bezeichnet (Green et al. 1977). Es ist entscheidend zu verstehen, dass der Koeffizient α nicht darüber reflektiert, inwieweit eine Skala homogen ist, oder nicht (Yang et al 2011: 380, Cortina 1993, Green et al. 1977, Miller 1995, Schmitt 1996). Eine unidimensionale Skala kann dabei zu einem niedrigen oder einem hohen Koeffizienten α führen. Dies gilt genauso für eine multidimensionale Skala (Yang et al. 2011: 380).

Hieraus ergibt sich die Frage, ob überhaupt von einem Maß der Reliabilität gesprochen werden kann, wenn es regelmäßig zu einer Unterschätzung kommt, da meist keine essentiell τ-äquivalenten Messungen vorliegen. Des Weiteren wird die Begrifflichkeit der internen Konsistenz hierzu in Bezug gesetzt.
In der Literatur wird der Koeffizient α oftmals als Standardmethode zur Schätzung

der internen Konsistenz dargestellt (Bühner 2004: 122, Christensen et al. 2011: 144, Rubin et al. 2008: 184, Cho et al. 2014: 8 ff.). Das Konzept der internen Konsistenz ist jedoch unscharf (Sijtsma 2009: 114). Nach Schmitt (1996) bedeutet interne Konsistenz, dass die verwendeten Items in einer Wechselbeziehung zueinander stehen („being interrelated“). Schwierigkeiten ergeben sich zudem, da Forscherinnen und Forscher die Begriffe Homogenität und interne Konsistenz gleichsetzen (Yang et al. 2011: 380, Cortina 1993). Durch die Homogenität wird jedoch angegeben, inwieweit eine Skala unidimensional ist und durch die interne Konsistenz wird darüber Auskunft gegeben, ob die Items in Wechselbezug zueinander stehen.
Der Koeffizient kann jedoch keine Aussagen über die interne Konsistenz geben. Der Schluss von dem Koeffizienten α auf die interne Konsistenz ist logisch betrachtet nicht korrekt, da der Koeffizient von der Anzahl der verwendeten Items abhängt und zusätzlich meist nur eine untere Schranke für die Reliabilität ist. Es sollte daher die Begrifflichkeit der internen Konsistenz vermieden werden.

Es liegen somit regelmäßig keine essentiell τ-äquivalenten Messungen vor. Green et al. (2009) haben im Rahmen einer Simulation untersucht, welche Auswirkungen sich daraus ergeben. Im Rahmen der Studie wird eine Faktorenanalyse verwendet (Raykov 2001). Details hierzu können der zitierten Arbeit entnommen werden, sind für diese Arbeit nicht erforderlich, da die Ergebnisse der Simulation auch ohne diese Betrachtungsweise dargestellt werden können. Es wurde durch die Autoren konkret untersucht, wie hoch der negative Bias ausfällt, wenn keine essentielle τ-Äquivalenz vorliegt. Der negative Bias ist dabei die Differenz zwischen $\rho_{xx'}$ und dem Koeffizienten α. Es wurden dabei zwei Skalen mit sechs bzw. zwölf Items betrachtet. Der negative Bias wurde dabei für unterschiedliche wahre Reliabilitätswerte berechnet, die im Rahmen der Simulation angenommen wurden. Für sechs Items lag die wahre Reliabilität zwischen 0,2 und 0,914, bei zwölf Items zwischen 0,333 und 0,955. Die Reliabilität wurde dabei über Faktorladungen variiert. Die Faktorladungen wurden zwischen 0,2 und 0,8 angenommen. Eine Zunahme einer Faktorladung pro Item bedeutet, dass der wahre Werte bei einer gleichzeitigen Abnahme des Fehlers zunimmt. Bei einer Faktorladung von 1,0 liegt eine mindestens τ-äquivalente Messung vor.
Der negative Bias lag bei der Simulationsanalyse in den meisten Fällen bei weniger als fünf Prozent der wahren Reliabilität. In vereinzelten Fällen lag der negative Bias jedoch bei mehr als zehn Prozent. Die wahre Reliabilität kann dadurch bedeutsam unterschätzt sein. Dies zeigt sich bei der durchgeführten Simulation jedoch nur selten. Nichtsdestotrotz sollte der Koeffizient nicht unreflektiert berichtet werden, sondern vielmehr die Struktur der Skala genau angesehen werden.

6.5 Korrelation zwischen den Fehlern

In diesem Kapitel wird auf die Verletzung der Annahme von unkorrelierten Fehlern eingegangen. Hierzu werden zuerst Situationen beschrieben, in denen eine Korrelation zwischen den Fehlern auftreten kann. Anschießend wird analytisch gezeigt, welche Auswirkungen sich hieraus auf den Koeffizienten α ergeben. Auf dieser Grundlage wird ein adjustierter Koeffizient α_κ entwickelt. Des Weiteren wird der Frage nachgegangen, ob der Koeffizient α noch eine untere Schranke für die Reliabilität sein kann, falls eine Annahmeverletzung vorliegt. Abschließend wird ein Simulationsergebnis zu dieser Thematik dargestellt.

Es wird in der klassischen Testtheorie angenommen, dass

$$\sum_{i \neq j} \sigma(\epsilon_i, \epsilon_j) = 0 \tag{43}$$

gilt. Dies bedeutet, dass die Summe der Kovarianzen und somit auch die Korrelation der Fehler null beträgt. Es kann jedoch Situationen geben, in denen die experimentelle Unabhängigkeit der Fehler nicht mehr erfüllt ist. Dies tritt auf, wenn eine Antwort einer Testperson zu einem Item abhängig ist von einer Antwort zu einem anderen Item, unabhängig des wahren Wertes. Die Testpersonen antworten sozusagen nicht unabhängig zu allen Items (Raykov 2001: 70, Lord et al. 1968). Hieraus ergibt sich, dass die Fehler korrelieren und eine Kovarianz vorliegt. Die getroffene Annahme ist dann nicht mehr erfüllt.

Die Annahme von unkorrelierten Fehlern wird beispielsweise von Rozeboom (1966: 415) als kritisch gesehen:
„Distressingly, however, this is a most unrealstic assumptions. When the individual items on a test are administered at one sitting it is enevitable - or at least we have no reason to think otherwise - that a substantial proportion of the uncontrolled extraneous conditions which are presumably responsible for measrurement error will persist from one item to the next, thus inducing positive error correlation which may be substantial [...] however pleasant a mathematical pastime it may be to shuffle through the internal statistics of a compound test in search of a formula which gives the closet estimate of the test-reliability under conditons of uncollrelated errors, this is for practical applications like putting on a clean shirt to rassle a hog“.

Es geht nun darum konkrete Situationen zu beschreiben, bei denen dies auftritt. Eine vorübergehende Lärmquelle bei einer Testsituation kann beispielsweise bewir-

ken, dass $E(\epsilon_i) = 0$ nicht mehr gilt. Es liegt ein Bias vor, da der Fehler durch die zusätzliche Lärmquelle zunimmt. Hieraus ergeben sich auch positive Korrelationen zwischen einzelnen Fehlertermen ϵ_i und somit auch eine positive Kovarianz zwischen den Fehlern (Rae 2006: 57). Ein zusätzlicher Faktor, der nicht in der originären Testsituation abgebildet wird, hat somit Auswirkungen auf die Kovarianz der Fehler. Raykov (1997: 377) drückt dies folgendermaßen aus: „Error covariances are equivalent to the impact of an additional factor that loads equally on the pertinent manifest variables."

Des Weiteren kommt es bei Tests, bei denen bestimmte Aufgaben unter Zeitdruck mit richtig oder falsch zu beantworten sind, regelmäßig zu einer Verletzung der Annahme von unkorrelierten Fehlern (Green et al. 2009: 124, Rozeboom 1966, Crocker et al. 1986).

Eine wesentliche Rolle in der Psychometrie spielen vorübergehende bzw. gelegentliche Fehler. Das Antwortverhalten variiert sozusagen von Moment zu Moment und dies wird nicht durch eine systematische Komponente hervorgerufen (Ghiselli 1964: 213). Die Antwort zu einem Item kann dadurch fehlerbehaftet sein, da in Bezug zu einem Item Gefühle, mentale Schärfe oder Unkonzentriertheit bestehen (Green et al. 2009: 125, Huysamen 2006: 46). Dies ist durchaus einleuchtend, da das Antwortverhalten zu den Items sicherlich davon abhängt, ob man sich z.B. gerade konzentriert oder motiviert ist. Es wird angenommen, dass die unterschiedliche Verfassung (z.B. Konzentration, Motivation) der Teilnehmer einer Studie durch die Items vermittelt wird. Dies trifft nach der Ansicht des Autors insbesondere bei langwierigen Befragungen zu, da die Konzentration und Motivation mit zunehmender Befragungsdauer wohl abnimmt. Um dem (teilweise) entgegenzuwirken, könnten die Items im Fragebogen in ihrer Reihenfolge zufällig angeordnet werden. Die personenspezifischen Fehler, die nicht durch die Items vermittelt werden, finden sich im Fehler ϵ, für den die Annahme $E(\epsilon) = 0$ gilt.

Messbar können vorübergehende Fehler nur gemacht werden, falls Messwiederholungen durchgeführt werden. Wenn Beobachtungen nur zu einem Zeitpunkt durchgeführt werden, kann der punktuelle Fehler nicht abgeschätzt werden. Da bei dem Koeffizienten α nur einmal Daten erfasst werden, kann dieser Fehler somit nicht berechnet werden.

6.5.1 Adjustierter Koeffizient α_κ

Bei der Herleitung des Koeffizienten α wurde die Annahme getroffen, dass $Cov(y_i, y_j) = Cov(\tau_i, \tau_j)$. Es kann jedoch zu den beschriebenen Situationen kommen, in denen diese Annahme verletzt ist. Die folgenden Ausführungen beziehen sich auf Komaroff (1997).
Es wird hier analytisch gezeigt, wie sich eine Verletzung der Annahme von unkorrelierten Fehlern auswirkt. Simulationsbasierte Ergebnisse hierzu finden sich im Kapitel 6.5.4.
Durch das Vorliegen einer Kovarianz zwischen den Fehlern ergibt sich nun:

$$Cov(y_i, y_j) = \sigma(y_i, y_j) = \sigma(\tau_i, \tau_j) + \sigma(\epsilon_i, \epsilon_j) \tag{44}$$

Bisher wurde angenommen, dass $\sigma(\epsilon_i, \epsilon_j) = 0$ und dadurch gilt auch $\sum_{i \neq j} \sigma(\epsilon_i, \epsilon_j) = 0$, so dass sich $\sigma^2(X)$ wie in Formel 21 ergibt. Wird die Annahme $\sum_{i \neq j} \sigma(\epsilon_i, \epsilon_j) = 0$ aufgegeben, dann ergibt sich:

$$\sigma^2(X) = \sum_{i=1}^{n} [\sigma^2(\tau_i) + \sigma^2(\epsilon_i)] + \sum_{i \neq j} [\sigma(\tau_i, \tau_j) + \sigma(\epsilon_i, \epsilon_j)] \tag{45}$$

Der Koeffizient α kann in folgender Form dargestellt werden:

$$\alpha = \frac{n}{n-1} \Big(\frac{\sum_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)} \Big). \tag{46}$$

Wenn man die vorhergehenden Ausführungen berücksichtigt, dann ergibt sich der Koeffizient in folgender geänderter Form:

$$\alpha = \frac{n}{n-1} \Big(\frac{\sum_{i \neq j} [\sigma(\tau_i, \tau_j) + \sigma(\epsilon_i, \epsilon_j)]}{\sum_{i=1}^{n} [\sigma^2(\tau_i) + \sigma^2(\epsilon_i)] + \sum_{i \neq j} [\sigma(\tau_i, \tau_j) + \sigma(\epsilon_i, \epsilon_j)]} \Big) \tag{47}$$

Es zeigt sich aus dieser Formel, dass sich ein Bias aufgrund der Kovarianz zwischen den Fehlern, sowohl im Zähler als auch im Nenner ergibt. Um diesen Bias zu korrigieren, wird von Komaroff (1997: 343) folgender adjustierter Koeffizient vorgeschlagen:

$$\alpha_\kappa = \frac{n}{n-1} \Big(\frac{\sum_{i \neq j} \sigma(y_i, y_j) - \sum_{i \neq j} \sigma(\epsilon_i, \epsilon_j)}{\sigma^2(X) - \sigma(\epsilon_i, \epsilon_j)} \Big) \tag{48}$$

Hierdurch kann der Bias aufgrund der bestehenden Kovarianz der Fehler herausgerechnet werden. Wie man aus den Formeln leicht entnehmen kann, wird der Zähler

und Nenner des nicht adjustierten Koeffizienten überschätzt, wenn eine positive Kovarianz zwischen den Fehlern entsteht. Falls sich in einer Studie eine negative Kovarianz ergibt, dann liegt eine Unterschätzung des Koeffizienten vor.

6.5.2 Untere Schranke der Reliabilität

Von Interesse erscheint, ob sich bei einer Überschätzung des Koeffizienten noch eine untere Schranke für die Reliabilität durch den Koeffizienten α ergibt. Die Ausführungen gelten also, falls der adjustierte Koeffizient α_κ nicht zur Analyse verwendet wird, aber eine positive Kovarianz zwischen den Fehlern vorliegt. Es gilt nach Kapitel 4.4 Folgendes:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X)}\right) \leq \rho_{xx'} \tag{49}$$

Falls eine vorliegende Kovarianz zwischen den Fehlern nicht berücksichtigt wird, ergibt sich:

$$\alpha^* = \frac{n}{n-1}\left(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j) + \sigma(\epsilon_i, \epsilon_j)}{\sigma^2(X) + \sigma(\epsilon_i, \epsilon_j)}\right) > \alpha = \frac{n}{n-1}\left(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X)}\right). \tag{50}$$

Da sich der Koeffizient α in diesem Fall auf α^* erhöht, stellt sich die Frage, ob $\alpha^* \leq \rho_{xx'}$ gilt. Die nicht bekannte wahre Reliabilität $\rho_{xx'}$ wird durch die zusätzliche Fehlerkovarianz folgendermaßen beeinflusst:

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X) + \sum\limits_{i\neq j}\sigma(\epsilon_i, \epsilon_j)} \tag{51}$$

Es wird dabei die Annahme vertreten, dass der Nenner durch die Kovarianz zwischen den Fehlern erweitert wird und somit $\sigma^2(X)$ erweitert wird (Rae 2006: 58, Raykov 2001). Es handelt sich hierbei um einen Teil der Reliabilität der berücksichtigt werden sollte. Die Gegenposition vertritt Komaroff (1997: 345). Die Ungleichung

$$\alpha^* \leq \rho_{xx'} \tag{52}$$

muss daher nicht mehr gelten. Bei α^* und $\rho_{xx'}$ erhöht sich der Nenner jeweils identisch. Der Zähler erhöht sich jedoch nur bei α^*. Die Varianz der τ-Werte wird durch eine vorliegende Kovarianz zwischen den Fehlern nicht beeinflusst. Im Vergleich zu $\rho_{xx'}$ erhöht sich α^* dadurch. Es kann daher Situationen geben, in denen der Koeffizient α nicht mehr eine untere Schranke für die Reliabilität ist. Dies tritt jedenfalls auf, falls essentiell τ-äquivalente Messungen vorliegen. Es sind jedoch auch andere

Situationen denkbar, in denen der Koeffizient α nicht mehr eine untere Schranke für die Reliabilität ist.

Dies kann auch analytisch betrachtet werden. Hierzu wird der Beweis zu Cronbachs α nochmals betrachtet (vgl. Kapitel 4.4). Das zentrale Element des Beweises von α als eine untere Schranke für die Reliabilität ist die Cauchy-Schwarz-Ungleichung. Es wurde beim Beweis folgende Ungleichung verwendet:

$$2\sum_{i\neq j}\sigma_{ij} \leq \sum_{i\neq j}(\sigma^2(\tau_i)+\sigma^2(\tau_j)). \tag{53}$$

Bei einem Vorliegen einer Kovarianz zwischen den Fehlern muss dies jedoch umformuliert werden, da gilt: $\sigma(y_i,y_j) \neq \sigma(\tau_i,\tau_j) \neq \sigma_{ij}$, denn $\sum\limits_{i\neq j}\sigma(y_i,y_j) = \sum\limits_{i\neq j}\sigma(\tau_i,\tau_j) + \sum\limits_{i\neq j}\sigma(\epsilon_i,\epsilon_j)$. Es ergibt sich:

$$2\sum_{i\neq j}\sigma_{\tau_i,\tau_j} \leq \sum_{i\neq j}(\sigma^2(\tau_i)+\sigma^2(\tau_j)). \tag{54}$$

Nach den bereits unter Kapitel 4.4 durchgeführten Umformungen dieses Ausdruckes ergibt sich unter Berücksichtigung der Kovarianz zwischen den Fehlern folgende Ungleichung:

$$\sigma^2(\tau)=\sum_{i=1}^{n}\sigma^2(\tau_i)+\sum_{i\neq j}\sigma(\tau_i,\tau_j) \geq \frac{1}{n-1}\sum_{i\neq j}\sigma(\tau_i,\tau_j)+\sum_{i\neq j}\sigma(\tau_i,\tau_j) = \frac{n}{n-1}\sum_{i\neq j}\sigma(\tau_i,\tau_j) \tag{55}$$

Die Ungleichung

$$\rho_{xx'}=\frac{\sigma^2(\tau)}{\sigma^2(X)} \geq \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i,y_j)}{\sigma^2(X)}\Big) \tag{56}$$

gilt jedoch bei einem Vorliegen von einer Kovarianz zwischen den Fehlern nicht mehr, denn $\sum\limits_{i\neq j}\sigma(y_i,y_j) \neq \sum\limits_{i\neq j}\sigma(\tau_i,\tau_j)$, da gilt $\sum\limits_{i\neq j}\sigma(y_i,y_j) = \sum\limits_{i\neq j}\sigma(\tau_i,\tau_j) + \sum\limits_{i\neq j}\sigma(\epsilon_i,\epsilon_j)$. Der Zusammenhang

$$\sigma^2(X)-\sum_{i=1}^{n}\sigma^2(y_i)=\sum_{i\neq j}\sigma(\tau_i,\tau_j)=\sum_{i\neq j}\sigma(y_i,y_j) \tag{57}$$

gilt nicht mehr, denn nun ist

$$\sigma^2(X)-\sum_{i=1}^{n}\sigma^2(y_i)-\sum_{i\neq j}\sigma(\epsilon_i,\epsilon_j)=\sum_{i\neq j}\sigma(\tau_i,\tau_j). \tag{58}$$

Der Schluss von der Formel 55 auf 56 ist nicht mehr möglich. Der Beweis funktioniert an dieser Stelle also nicht mehr. Es gilt

$$\sum_{i \neq j} \sigma(y_i, y_j) > \sum_{i \neq j} \sigma(\tau_i, \tau_j) \tag{59}$$

und somit können sich die Situationen

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X)} \geq \frac{n}{n-1}\Big(\frac{\sum\limits_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\Big) \tag{60}$$

und

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X)} < \frac{n}{n-1}\Big(\frac{\sum\limits_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\Big) \tag{61}$$

ergeben. Cronbachs α muss daher nicht mehr eine untere Schranke für die Reliabilität sein, falls die Kovarianz zwischen den Fehlern nicht berücksichtigt wird.
Weiter wurde bereits analysiert, ob α_κ eine untere Schranke für die Reliabilität sein muss (Rae 2006). Das Ergebnis dieser Analyse ist, dass α_κ nicht zwangsläufig eine untere Schranke für die Reliabilität sein muss, falls man annimmt, dass die Kovarianz zwischen den Fehlern ein Teil der wahren Reliabilität ist. Trifft man die Annahme, dass die Kovarianz nicht Teil der wahren Reliabilität ist, so ist α_κ eine untere Schranke für die Reliabilität.

6.5.3 Simulationsergebnisse

Komaroff (1997) hat in seiner Arbeit sowohl den adjustierten Koeffizienten α_κ entwickelt, als auch eine Simulation hierzu durchgeführt. Es wurden die Auswirkungen einer Verletzung der Annahme von unkorrelierten Fehlern betrachtet. Um mögliche Korrelationen zwischen den Fehlern der Items zu berücksichtigen, wurde der bereits beschriebene adjustierte Koeffizient α_κ entwickelt. In einer Simulation wurden für unterschiedliche Situationen die Koeffizienten α, α_κ und $\rho_{xx'}$ miteinander verglichen. Die Berechnung von $\rho_{xx'}$ war möglich, da die τ-Werte bekannt waren.
Um verschiedene Koeffizienten zum Vergleich zu erhalten, wurden Korrelationen bei τ_i, τ_j und ϵ_i, ϵ_j in unterschiedlicher Höhe angenommen. Die Korrelation zwischen den τ-Werten wurden folgendermaßen angenommen: $\rho(\tau_i, \tau_j) = 0, 0.2, 0.5, 0.7, 1$. Die Korrelation zwischen den Fehlern wurde auf folgende Werte festgesetzt: $\rho(\epsilon_i, \epsilon_j) = 0, 0.2, 0.5, 0.7, 1$. Des Weiteren wurde die Anzahl der Items zwischen 6, 12 und 18 variiert.
Bei der Studie konnte belegt werden, dass der Koeffizient α bei unkorrelierten Fehlern in allen Situationen eine untere Schranke für die Reliabilität ist. Dies wurde im

Rahmen dieser Arbeit bereits analytisch gezeigt. Desto geringer $\rho(\tau_i, \tau_j)$ angenommen wird, desto größer ist die Differenz zwischen α und $\rho_{xx'}$. Die Unterschätzung wird also größer, desto kleiner die Korrelation zwischen den τ-Werten wird.

Des Weiteren konnte herausgefunden werden, dass die Inflation von α aufgrund einer Verletzung der Annahme von unkorrelierten Fehler ausgeprägter ist, falls $\rho(\epsilon_i, \epsilon_j) \to 1$. Dies ist unabhängig des Wertes von $\rho(\tau_i, \tau_j)$. Es zeigt sich, dass ab $\rho(\epsilon_i, \epsilon_j) = 0,5$ eine Überschätzung der Reliabilität vorliegt, so dass gilt: $\alpha > \rho_{xx'}$. Der Koeffizient stellt in diesen Fällen keine untere Schranke für die Reliabilität mehr da. Dies gilt bei der Annahme, dass die Kovarianz zwischen den Fehlern nicht ein Teil der wahren Reliabilität $\rho_{xx'}$ ist (Komaroff 1997: 345). Es wäre hier interessant zu sehen, wie die Überschätzung des Koeffizienten ausfallen würde, falls die Kovarianz zwischen den Fehlern auch bei der wahren Reliabilität berücksichtigt werden würde.
Zudem wurde für die beschriebenen unterschiedlichen Situationen der adjustierte Koeffizient α_κ berechnet. Dieser Koeffizient war jeweils eine untere Schranke für die Reliabilität.

6.6 Korrelation zwischen dem wahren Wert und dem Fehler

In diesem Abschnitt wird betrachtet, welche Auswirkungen sich ergeben, falls der wahre Wert und der Fehler korreliert sind. Dies kann beispielsweise auftreten, falls ein Intelligenztest durchgeführt wird und der Fehlerterm von der wahren Intelligenz der Probanden abhängt. Hierzu wird zuerst ein adjustierter Koeffizient α_ψ entwickelt. Zudem wird auch hier betrachtet, ob bei einer Annahmeverletzung immer noch eine untere Schranke für die Reliabilität durch den Koeffizienten α vorliegt.

6.6.1 Adjustierter Koeffizient α_ψ

Es kann also sein, dass die Annahme $\sigma(\tau, \epsilon) = 0$ der klassischen Testtheorie verletzt ist. Dies hat wiederum Auswirkungen auf den Koeffizienten α. Aus Kapitel 4.4 ist für zusammengesetzte Tests bekannt, dass $X = T + E$ gilt. Die Varianz von X setzt sich wie folgt zusammen:

$$\sigma^2(X) = \sigma^2(T + E) = \sigma_\tau^2 + \sigma_\epsilon^2 + 2\sigma_{\tau\epsilon}. \tag{62}$$

Falls angenommen wird, dass keine Kovarianz zwischen dem wahren Wert und dem Fehler besteht, reduziert sich die vorstehende Formel auf

$$\sigma^2(X) = \sigma^2(T + E) = \sigma_\tau^2 + \sigma_\epsilon^2, \tag{63}$$

da angenommen wird: $\sigma_{\tau\epsilon} = \rho_{\tau\epsilon}\sigma_\tau\sigma_\epsilon = 0$, da $\rho_{\tau\epsilon} = 0$. Dies folgt wiederum aus $E(X) = \tau$. Es kann jedoch sein, dass $\sigma_{\tau\epsilon} \neq 0$ gilt. Dies ist der Fall, wenn der Messfehler von der Ausprägung des wahren Wertes abhängt. Die Varianz $\sigma^2(X)$ kann auch wie folgt dargestellt werden (vgl. Formel 21): $\sigma^2(X) = \sum_{i=1}^{n} \sigma^2(y_i) + \sum_{i \neq j} \sigma(y_i, y_j)$. Der zusätzliche Fehler $\sigma(\tau, \epsilon)$ kann sowohl positiv als auch negativ sein. Beispielsweise könnte man davon ausgehen, dass ein hoher τ-Wert - also z.B. eine hohe Intelligenz - mit einem niedrigen ϵ-Wert einhergeht und umgekehrt. Hier würde man davon ausgehen, dass für Personen mit höherer Intelligenz ein kleinerer Messfehler resultiert, da sich diese Personen ggf. besser (über einen längeren Zeitraum) konzentrieren können. Es würde also eine negative Korrelation vorliegen. Falls die Reaktionsschnelligkeit unter Einfluss von Alkohol getestet wird, wird hingegen der Fehler bei höherem Alkoholgehalt im Blut zunehmen. Es liegt hier ein positive Korrelation zwischen dem wahren Wert und dem Messfehler vor.
Berücksichtigt man bei der Zusammenstellung der Varianz die angenommene Kovarianz zwischen dem Fehler und dem wahren Wert bei den einzelnen Items, so ergibt sich bei $\sigma_{\tau\epsilon} \neq 0$ folgender Ausdruck:

$$\sigma^2(X) = \sum_{i=1}^{n} \sigma^2(y_i) + \sum_{i \neq j} \sigma(y_i, y_j) + 2\sum_{i=1}^{n} \sigma(\tau_i, \epsilon_i) \tag{64}$$

bzw.

$$\sigma^2(X) = \sum_{i=1}^{n} [\sigma^2(\tau_i) + \sigma^2(\epsilon_i) + 2\sigma(\tau_i, \epsilon_i)] + \sum_{i \neq j} \sigma(y_i, y_j). \tag{65}$$

Cronbachs α kann in folgender Form dargestellt werden: $\alpha = \frac{n}{n-1}\left(\frac{\sum_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\right)$. Bei einer Berücksichtigung von $\sigma(\tau\epsilon) \neq 0$ stellt sich der Koeffizient folgendermaßen dar:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum_{i \neq j} \sigma(y_i, y_j)}{\sum_{i=1}^{n} [\sigma^2(\tau_i) + \sigma^2(\epsilon_i) + 2\sigma(\tau_i, \epsilon_i)] + \sum_{i \neq j} \sigma(y_i, y_j)}\right). \tag{66}$$

Falls in diesem Fall der zusätzliche Fehler nicht berücksichtigt wird, dann kommt es zu einer Überschätzung ($\sigma_{\tau\epsilon} > 0$) bzw. zu einer Unterschätzung ($\sigma_{\tau\epsilon} < 0$) des Koeffizienten - ceteris paribus.
Um die entstehende Über- bzw. Unterschätzung zu korrigieren, wird folgender adjustierter Koeffizient vorgeschlagen:

$$\alpha_{\psi} = \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X) - 2\sum\limits_{i=1}^{n}\sigma(\tau_i, \epsilon_i)}\Big). \tag{67}$$

6.6.2 Untere Schranke der Reliabilität

Von Interesse erscheint auch hier, ob sich bei einer Überschätzung des Koeffizienten ($\sigma_{\tau\epsilon} > 0$) noch eine untere Schranke für die Reliabilität durch den Koeffizienten α ergibt. Die Ausführungen gelten als somit, falls der adjustierte Koeffizient α_{ψ} nicht zur Analyse verwendet wird, aber eine positive Kovarianz zwischen dem wahren Wert und dem Fehler vorliegt. Es wird im Folgenden überprüft, ob der Beweis zur Herleitung des Koeffizienten α aufrechterhalten werden kann. Es gilt nach Kapitel 4.4 Folgendes:

$$\alpha = \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X)}\Big) \leq \rho_{xx'} \tag{68}$$

Falls eine vorliegende Kovarianz zwischen den Fehlern nicht berücksichtigt wird, ergibt sich:

$$\alpha^{**} = \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X) + 2\sum\limits_{i=1}^{n}\sigma(\tau_i, \epsilon_i)}\Big) > \alpha = \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X)}\Big) \tag{69}$$

Da sich der Koeffizient α in diesem Fall auf α^{**} erhöht, stellt sich die Frage, ob $\alpha^{**} \leq \rho_{xx'}$ gilt. Die nicht bekannte wahre Reliabilität $\rho_{xx'}$ wird durch die zusätzliche Fehlerkovarianz folgendermaßen beeinflusst:

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X) + 2\sum\limits_{i=1}^{n}\sigma(\tau_i, \epsilon_i)} \tag{70}$$

Es wird hier auch die Annahme vertreten, dass der Nenner durch die Kovarianz zwischen dem wahren Wert und dem Fehler erweitert wird und somit $\sigma^2(X)$ erweitert wird. Die Annahme wird entsprechend zur Kovarianz zwischen den Fehlern getroffen. Die Ungleichung

$$\alpha^{**} \leq \rho_{xx'} \tag{71}$$

gilt in diesem Fall des Weiteren. Durch die zusätzliche Kovarianz wird bei α^{**} und $\rho_{xx'}$ jeweils der Nenner in gleicher Weise beeinflusst, so dass sich hinsichtlich der Ungleichung keine Veränderungen ergeben. Die Zähler werden durch den zusätzlichen Term nicht beeinflusst. Falls die beschriebene Kovarianz nicht berücksichtigt

wird, ist der Koeffizient α deshalb des Weiteren eine untere Schranke für die Reliabilität. Hieraus ergibt sich auch, dass α_ψ eine untere Schranke für die Reliabilität ist.

Dies kann auch analytisch betrachtet werden. Es wird hierzu ebenfalls der Beweis zu Cronbachs α nochmals betrachtet. Das zentrale Element des Beweises von α als eine untere Schranke für die Reliabilität ist, wie bereits beschrieben, die Cauchy-Schwarz-Ungleichung:

$$2\sum_{i\neq j}\sigma_{ij} \leq \sum_{i\neq j}(\sigma^2(\tau_i)+\sigma^2(\tau_j)). \tag{72}$$

Bei $\sigma_{\tau\epsilon} \neq 0$ bleibt diese Ungleichung unberührt. Es gilt des Weiteren $\sigma(y_i, y_j) = \sigma(\tau_i, \tau_j) = \sigma_{ij}$ und somit ändert sich auch nicht die Cauchy-Schwarz-Ungleichung. Der Term $\sigma_{\tau\epsilon} > 0$ erhöht ausschließlich die Varianz $\sigma^2(X)$ und zwar in der Form, dass gilt:

$$\sigma^2(X) = \sum_{i=1}^{n}\sigma^2(y_i) + \sum_{i\neq j}\sigma(y_i, y_j) + 2\sum_{i=1}^{n}\sigma(\tau_i, \epsilon_i) \tag{73}$$

Bei dem dargestellten Beweis wird der Term $\sigma^2(X)$ sowohl bei der wahren Reliabilität als auch bei dem Koeffizienten α jeweils im Nenner verwendet, die sich dadurch entsprechend verändern:

$$\rho_{xx'} = \frac{\sigma^2(\tau)}{\sigma^2(X) + 2\sum\limits_{i=1}^{n}\sigma(\tau_i, \epsilon_i)} \geq \frac{n}{n-1}\Big(\frac{\sum\limits_{i\neq j}\sigma(y_i, y_j)}{\sigma^2(X) + 2\sum\limits_{i=1}^{n}\sigma(\tau_i, \epsilon_i)}\Big) \tag{74}$$

Bezüglich des Verhältnisses $\rho_{xx'} \geq \alpha$ ergeben sich hierdurch keine Änderungen.

Im Kapitel 9 wird die beschriebene Annahmeverletzung im Rahmen einer Simulation näher betrachtet.

7 α-Maximierung - Ein Gegenvorschlag

Die Itemauswahl für eine Skala kann in der Art und Weise getroffen werden, dass sich ein maximaler α-Wert ergibt. Ein derartiger Forschungsstil zeigt sich bei vielen empirischen Arbeiten. Die Entscheidung, ob Items in eine Skala aufgenommen werden, erfolgt in diesem Fall durch subjektive Entscheidungen des Forschers. Es geht nun darum, ein Verfahren zu entwickeln, das diese Vorgehensweise entbehrlich macht, da die α-Maximierung verschiedene Nachteile mit sich bringt.
Durch das Löschen von Items zur Erhöhung des Koeffizienten α kann es sein, dass die gesamte Qualität der Skala leidet (Yang et al. 2011: 389). Es besteht die Gefahr,

eine Skala „zu Tode“ zu homogenisieren, um einen möglichst hohen Koeffizienten α zu erhalten. Mit einer Verengung der Skala zur Erhöhung des Koeffizienten wird aber das zu messende latente Konstrukt eingeschränkt, sodass dieses ggf. nicht mehr ausreichend abgebildet wird. Im Extremfall könnte es sich ergeben, dass nur noch Items im Index enthalten sind, die sich sehr ähnlich sind und hierdurch kann keine inhaltliche Aussage mehr über das latente Konstrukt erfolgen. In diesem Fall würde sich also ein hoher Koeffizient ergeben, der jedoch eine triviale Skala abbildet.

Die Überlegung, dass durch das Löschen von Items der Koeffizient α erhöht wird, steht dabei nicht in Widerspruch zu Kapitel 6.1. Hier wurde gezeigt, dass der Koeffizient α mit der Anzahl der Items zunimmt. Ein Löschen von Items würde nach dieser Betrachtungsweise mit einer Reduzierung des Koeffizienten α einhergehen. In diesem Kapitel wurden jedoch ausschließlich die Auswirkungen der Zunahme der Items auf den Koeffizienten betrachtet, aber nicht eine Veränderung der Kovarianz zwischen den Items und der Gesamtvarianz. Es kann daher Situationen geben, bei denen es bei einem Löschen von Items zu einer Erhöhung des Koeffizienten kommen kann. Die Auswahl der Items sollte daher aus substanzwissenschaftlichen Überlegungen erfolgen und sich nicht an der Maximierung des Koeffizienten α orientieren, der ggf. die Reliabilität (stark) unter- bzw. überschätzt.
Es sollte dadurch eine Methode entwickelt werden, die eine α-Maximierung entbehrlich macht. Es wird die Berechnung eines durchschnittlichen α-Wertes nach der Idee der Leave-One-Out-Cross-Validation (Fahrmeir et al. 2013: 149) vorgeschlagen.
Der Koeffizient α wird durch die Formel

$$\alpha = \frac{n}{n-1}\Big(\frac{\sum\limits_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\Big) \tag{75}$$

berechnet. Die Daten werden dadurch auf eine Maßzahl komprimiert. Der Koeffizient schwankt mit der Anzahl der Items, der Kovarianz der Items und der Summe der Varianz der verwendeten Items. Falls die Anzahl der Items durch ein objektives Verfahren verändert wird, dann ergeben sich unterschiedliche α-Werte, die zur Berechnung eines durchschnittlichen α-Wertes verwendet werden können. Um die α-Werte zu erhalten, wird die Idee der Leave-One-Out-Cross-Validation angewendet. Die Idee besteht darin, dass verschiedene Koeffizienten α durch das Entnehmen von jeweils einem Item aus der Analyse berechnet werden. Dies wird durch den Index -i dargestellt:

$$\alpha^{-i} = \frac{n-1}{n-2}\Big(\frac{\sum\limits_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\Big) \tag{76}$$

Der Korrekturfaktor wird aufgrund der veränderten Anzahl der Items angepasst. Ein durchschnittlicher α-Wert ergibt sich dann folgendermaßen:

$$\overline{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \tag{77}$$

mit $\alpha_i = (\alpha_1, \alpha_2, \ldots, \alpha_{n-1}, \alpha_n)$,
$(\alpha_1, \ldots, \alpha_{n-1}) = \alpha^{-i} = \frac{n-1}{n-2}\left(\frac{\sum_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\right)$ und $\alpha_n = \frac{n}{n-1}\left(\frac{\sum_{i \neq j} \sigma(y_i, y_j)}{\sigma^2(X)}\right)$.

Hierdurch erfolgt eine objektive Berechnung eines durchschnittlichen α-Wertes. Unsicherheiten oder Fehler bei der Itemselektion können hierdurch ausgeglichen werden. Es werden unterschiedlich zusammengesetzte Skalen berücksichtigt, so dass eine Maximierung des α-Wertes „per Hand “ entbehrlich wird. Der durchschnittliche α-Wert gibt Informationen über unterschiedliche Skalen und hierdurch wird die Beschreibung des latenten Konstruktes durch eine zusätzliche Datenbasis verfeinert.

Hierzu können Ergebnisse aus einer Simulation und aus der Analyse von empirischen Daten dem Kapitel 9 entnommen werden.

8 α-Konfidenzintervall

Mit dem Koeffizienten α erfolgt eine Punktschätzung einer unteren Schranke der Reliabilität, wenn die Messungen nicht mindestens essentiell τ-äquivalent sind. Zusätzlich zu dieser Punktschätzung kann eine Konstruktion eines Konfidenzintervalls hinsichtlich der Items sinnvoll sein. Zudem ist es nach Ansicht des Autors sinnvoll eine Testentscheidung darüber zu treffen, ob eine Skala reliabel im verbalen Sinn ist, oder nicht. Hierzu erfolgt zuerst eine Darstellung der Grundannahme hierzu. Anschließend wird die für die Konstruktion eines Konfidenzintervalls erforderliche Methode des Bootstraps beschrieben, ehe das konkret verwendete Bootstrap-t-Konfidenzintervall näher betrachtet wird. Danach wird die Dualität zwischen einem Konfidenzintervall und einem Test dargestellt. Hierauf aufbauend wird ein Test auf Grundlage des konstruierten Konfidenzintervalls dargestellt.

8.1 Grundannahme

Es wird angenommen, dass die verwendeten Indikatoren bzw. Items zur Beschreibung eines latenten Konstruktes eine zufällige Auswahl über alle möglichen Items sind. Die Items stammen aus einer Grundgesamtheit von Items. Mortensen (2009:

71) formuliert dies folgendermaßen: „[...] zumal es sich bei den n Items jeweils um Stichproben aus einem Universum von Items handelt“. Die Annahme, dass nur bestimmte Items ein theoretisches Konstrukt abbilden können, wird als nicht realistisch eingestuft, da die Itemauswahl mit dem Index der Kontingenz versehen ist. Dies ergibt sich aus der Einschätzung, dass Wissenschaft im Allgemeinen mit dem Index der Kontingenz versehen ist (Nassehi 2008: 182).
Aus diesem Grund erscheinen asymptotische Aussagen über die zufällig ausgewählten Items als sinnvoll. Es kann hierdurch ein entsprechendes Konfidenzintervall konstruiert werden (Pflaumer et al. 2001: 101 ff., Rüger 1999: 125 ff.). Dem Konfidenzintervall ist dabei ein bestimmtes Konfidenzniveau (Vertrauensgrad) γ zugeordnet. Es gilt aufgrund der bestehenden Dualität zwischen Konfidenzintervallen und Tests folgender Satz (Korrespondenzsatz): $\gamma = 1 - \beta$ (Heumann 2013: 56 ff.). Dabei wird β als das Signifikantsniveau definiert. Wird beispielsweise ein Konfidenzniveau von 95 Prozent festgelegt, bedeutet dies, dass mindestens 95 Prozent aller auf Grundlage der erhobenen Daten konstruierten Konfidenzintervalle den wahren Koeffizienten α enthalten. Unabhängig hiervon gilt weiter die beschriebene Schwierigkeit, dass der Koeffizient α die wahre Reliabilität $\rho_{xx'}$ unter- bzw. überschätzt.
Um eine Zufallsstichprobe für α-Werte zu erhalten, wird das im vorherigen Kapitel beschriebene Verfahren zur Berechnung von unterschiedlichen α-Werten (Leave-One-Out-Cross-Validation) verwendet. Die Idee ein parametrisches Verfahren zur Konstruktion eines Konfidenzintervalls zu verwenden scheitert nach Ansicht des Autors, da hinsichtlich der α-Werte keine Verteilungsfunktion ersichtlich ist.

8.2 Bootstrap-Methode

Die folgenden Ausführungen beziehen sich auf Elfron (1979, 1993), Heumann (2014: 3 ff.) Winkler (1998), Fan (2010: 33 ff.), Pruscha (2000: 51 ff.), Pospeschill (2013: 258 ff.) und Pauls (2003: 10 ff.).

Mit der Methode der Leave-One-Out-Cross-Validation wird eine Stichprobe generiert. Mit dieser Stichprobe können Schlüsse auf den Parameterwert α gezogen werden. Für den interessierenden (aber unbekannten) α-Wert wird keine theoretische Verteilungsannahme getroffen, sodass die Voraussetzungen für parametrische Tests nicht gegeben sind. Es muss daher eine andere Methode gefunden werden, um ein Konfidenzintervall für den unbekannten Parameter zu konstruieren. Es soll hierfür die Bootstrap-Methode verwendet werden. Die theoretische Verteilungsfunktion wird hier durch die sogenannte Bootstrap-Verteilungsfunktion ersetzt. Es wird hierzu eine simulierte Verteilungsfunktion aus den empirischen Daten, den Bootstrap-

Stichproben, selbst erzeugt. Die Bootstrap-Methode ersetzt somit die theoretische Verteilungsfunktion einer Zufallsvariable durch eine empirische Verteilungsfunktion der Stichprobe.
Die empirische Verteilungsfunktion kann dabei nicht aus einer einzigen Stichprobe geschätzt werden. Es wird dadurch die Methode des Resampling verwendet, bei der Statistiken auf der Grundlage einer einzigen Stichprobe simuliert werden. Es wird eine vorher definierte Anzahl von Stichproben (z.B. 2.000 Stichproben) simuliert, in denen sich die Beobachtungen zufällig befinden oder nicht. Es können Beobachtungen dadurch einmal, mehrfach oder gar nicht in den einzelnen Stichproben vertreten sein.
Für die α-Werte wird angenommen, dass sie unabhängig und identisch verteilt sind. Die Reihenfolge der Beobachtungen ist nicht von Bedeutung, sodass dies bei der Erzeugung der Bootstrap-Stichproben nicht beachtet werden muss. Aufgrund dieser Annahme kann die Methode nach Elfron (1993) verwendet werden. Durch die Bootstrap-Methode lässt sich ein Bootstrap-Schätzer für den unbekannten Parameter α, ein Bootstrap-Schätzer für die Varianz und ein Bootstrap-Konfidenzintervall für den Parameter α gewinnen.

Wir definieren hierzu eine Zufallsvariable für die α_i-Werte, $\alpha_i = (\alpha_1, \alpha_2, ..., \alpha_n)$. Es handelt sich hierbei um die nach der Methode der Leave-One-Out-Cross-Validation generierten α_i-Werte (vgl. Kapitel 7). Der Index i wird im Folgenden zur Vereinfachung fallen gelassen. Für diese Zufallsvariable wird keine theoretische Verteilung angenommen, so dass diese unbekannt ist: $\alpha \overset{i.i.d}{\sim}$ F (F unbekannt). Es sei die Zufallsstichprobe $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \rightarrow T(x)$ bekannt. Der interessierende Parameter ist somit $\hat{\theta} = T(x)$. F wird durch die empirische Verteilungsfunktion $\hat{F}_n$ ersetzt, mit $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(\alpha_i \leq \alpha)$, I = Indikatorfunktion. Dieses Verfahren wird auch als Plug-In-Prinzip bezeichnet.

Um Bootstrap-Stichproben zu erhalten, wird n mal aus $\hat{F}_n$ mit Zurücklegen gezogen: $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*) \rightarrow T(\alpha^*) = \hat{\theta^*}(b)$. $T(\alpha^*)$ ist dabei die interessierende Funktion der jeweilig simulierten Stichprobe. Die Bootstrap-Stichproben haben somit jeweils die gleiche Länge n, wie die ursprüngliche Stichprobe. Die einzelnen Funktionen der simulierten Stichproben werden wie folgt bezeichnet: $(T(\alpha^{*1}), T(\alpha^{*2}), \ldots, T(\alpha^{*B}))$, mit $b = 1, \ldots, B$. Für jede Bootstrap-Stichprobe wird dadurch die Statistik T berechnet. Mit den Statistiken $(T(\alpha^{*1}), T(\alpha^{*2}), \ldots, T(\alpha^{*B}))$ lassen sich Aussagen über

die unbekannte Verteilung der Statistik T gewinnen:

$$Var_F(T) \sim \hat{Var}_{Boot}(T) = \Big(\frac{1}{B-1}\sum_{b=1}^{B}\Big[T(\alpha^{*b}) - \overline{T}_{Boot}\Big]^2\Big) \tag{78}$$

mit $\overline{T}_{Boot} = \frac{1}{B}\sum_{b=1}^{B} T(\alpha^{*b})$. Zur Schätzung der Varianz erfolgt somit zuerst eine Erzeugung von B replizierten Bootstrap-Stichproben, auf deren Grundlage jeweils ein Schätzer berechnet wird. Die dadurch mögliche Berechnung der Stichprobenvarianz $\hat{Var}_{Boot}(T)$ der B Schätzungen dient als Approximation der Varianz $Var_F(T)$ der Zufallsvariable.

8.3 Bootstrap-t-Konfidenzintervall

Von Interesse ist nun, wie ein Konfidenzintervall für den interessierenden Parameter konstruiert werden kann. Zuerst soll eine geeignete Teststatistik definiert werden. Hierfür wird das arithmetische Mittel als sinnvoll angesehen:

$$T(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i = \hat{\theta} \tag{79}$$

bzw.

$$T(\alpha^*) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i^{*b} = \hat{\theta^*}(b), \tag{80}$$

mit den Beobachtungen $i = 1, \ldots, n$ und den Bootrap-Stichproben $b = 1, \ldots, B$. Betrachtet wird nun:

$$Z = \frac{\hat{\theta} - \theta}{\hat{se}}, \tag{81}$$

wobei $\hat{se}$ eine sinnvolle Schätzung des Standardfehlers von $\hat{\theta}$ sein soll. Für Z wird keine Verteilung angenommen. Es wird daher ein Weg gesucht, um die Verteilung von Z zu schätzen. Hierzu werden B Bootstrap-Stichproben $\alpha^{*1}, \alpha^{*2}, \ldots, \alpha^{*B}$ gezogen. Hieraus soll

$$Z^*(b) = \frac{\hat{\theta^*}(b) - \hat{\theta}}{\hat{se}^*(b)} \tag{82}$$

berechnet werden. Die Plug-In-Schätzung des Standardfehlers für $\hat{\theta^*}(b)$ erfolgt mit

$$\hat{se}^*(b) = \frac{1}{n}\Big(\sum_{i=1}^{n}(\alpha_i^{*b} - \overline{\alpha}^{*b})^2\Big)^{\frac{1}{2}}. \tag{83}$$

Die Werte von $Z^*(b)$ sind sodann ihrer Größe nach zu ordnen. Aus dieser geordneten Reihe kann das $\hat{t}^{\beta/2}$ und das $\hat{t}^{1-\beta/2}$-Quantil für ein (1-β) Konfidenzintervall geschätzt

werden.:

$$\frac{\#(Z^*(b) \leq \hat{t}^{\beta})}{B} = \beta \tag{84}$$

Dabei bezeichnet $\#(Z^*(b) \leq \hat{t}^{\beta})$ die Kardinalität dieser Menge. Wählt man B = 2000, wie bei der Konstruktion eines Konfidenzintervalls in diesem Kontext üblich, ist $\hat{t}^{0,025}$ der 50. Wert und $\hat{t}^{0,975}$ der 1.950. Wert jeweils der geordneten $Z^*(b)$-Werte. Das Konfidenzintervall ergibt sich dadurch in folgender Form

$$\left[\hat{\theta} - \hat{t}^{1-\beta/2} * \hat{se}, \hat{\theta} - \hat{t}^{\beta/2} * \hat{se}\right], \tag{85}$$

mit $\hat{se} = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(.)\right]^2\right)^{\frac{1}{2}}$ und
$\hat{\theta}^*(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$. Es gilt: $\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}_n}(\hat{\theta}^*)$. Die Approximation konvergiert somit gegen die ideale Bootstrap-Schätzung des Standardfehlers und stellt somit eine sinnvolle Schätzung des Standardfehlers dar.

8.4 Dualität zwischen Konfidenzintervall und Test

Es ist zusätzlich von Interesse, einen Zusammenhang zwischen dem konstruierten Konfidenzbereich und einem entsprechenden Test herzustellen. Es besteht eine Dualität zwischen einem Konfidenzintervall und einem Test (Heumann 2013: 56 ff.). Der Test soll darauf orientiert werden, ob eine zu bestimmende Grenze (α_0) für die Reliabilität im verbalen Sinn durch die Schätzungen erreicht wird oder nicht. Dies wird in eine Nullhypothese und eine Alternativhypothese umgesetzt, die folgendermaßen lauten:

Nullhypothese:

$$H_0 : \alpha \leq \alpha_0 \tag{86}$$

Die Nullhypothese lautet, dass der Koeffizient α eine definierte α-Grenze nicht übersteigt. Falls das unter Kapitel 8.3 konstruierte Konfidenzintervall den Wert von α_0 einschließt, oder unterschreitet, dann kann die Nullhypothese nicht verworfen werden.

Alternativhypothese:

$$H_1 : \alpha > \alpha_0 \tag{87}$$

Die Alternativhypothese (bzw. Forschungshypothese) lautet, dass der Koeffizient α eine definierte α-Grenze übersteigt. Dies trifft zu, falls das Konfidenzintervall über

dem Grenzwert von α_0 liegt.
Dieser Test kann auch mit der unter Kapitel 6.2 entwickelten adjustierten α-Grenze (vgl. Formel 41) kombiniert werden. Es ergibt sich hier:

Nullhypothese:

$$H_0 : \alpha \leq \theta \tag{88}$$

Alternativhypothese:

$$H_1 : \alpha > \theta \tag{89}$$

Aufgrund der bestehenden Dualität zwischen einem Konfidenzintervall und einem Test kann also die Nullhypothese zu einem vorher definierten Signifikantsniveau β verworfen werden, falls das Konfidenzintervall komplett über der definierten Grenze α_0 bzw. θ liegt. Die Nullhypothese wird beibehalten, falls das Konfidenzintervall die definierte Grenze einschließt oder unter dieser Grenze liegt.

9 Simulationsergebnisse, Auswertung von empirischen Daten

Es werden nun Simulationen und eine Auswertung von empirischen Daten über verschiedene Themen dieser Arbeit durchgeführt. Zuerst werden die entwickelten Verfahren nach Kapitel 7 und 8 dargestellt. Im Anschluss wird aufgezeigt, wie sich eine Verletzung der Annahme $\sigma(\tau, \epsilon) = 0$ auf den Koeffizienten α auswirken kann. Die Simulationen erfolgten mit R (Version 3.1.2).

9.1 $\overline{\alpha}$, α-Konfidenzintervall

Im Kapitel 7 wurde die Berechnung eines durchschnittlichen α-Wertes dargestellt. Dazu müssen zuerst die α_i-Werte generiert werden. Im Rahmen der Simulation werden hierzu die α_i-Werte aus einer Normalverteilung gezogen, die folgendermaßen definiert ist: $N(\mu, \sigma^2)$. Die Verwendung der Normalverteilung erscheint hierfür sinnvoll zu sein, da es sich um eine symmetrische Verteilung handelt. Es wird davon ausgegangen, dass die α_i-Werte symmetrisch um einen α-Wert schwanken, der das latente Konstrukt abbildet. Es werden dabei unterschiedliche Werte von μ und σ^2 festgesetzt. Die verschiedenen α_i-Werte können der Tabelle 3 entnommen werden. Auf deren Grundlage wurde jeweils $\overline{\alpha}$ und das α-Konfidenzintervall berechnet. Es wird ein Signifikanzniveau von $\beta = 0{,}05$ und eine Grenze α_0 von 0,8 gewählt. Des

Weiteren wird der Testentscheidung anhand der adjustierten α-Grenze getroffen.

Tabelle 3: Simulation, $\overline{\alpha}$ und α-Konfidenzintervall

N(μ, σ^2)	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	$\overline{\alpha}$	KI
1 (0,78;0,02)	0.77	0.76	0.76	0,80	0.82	0.73	0.79	0.78	0.76	0.75	0.77	0.76-0.79
2 (0,78;0,04)	0.80	0.72	0.81	0.78	0.79	0.77	0.83	0.73	0.81	0.78	0.78	0.76-0.80
3 (0,80;0,02)	0.79	0.85	0.82	0.79	0.77	0.81	0.82	0.80	0.80	0.76	0.80	0.79-0.82
4 (0,80;0,04)	0.77	0.90	0.84	0.78	0.75	0.83	0.84	0.80	0.79	0.72	0.80	0,77-0.84
5 (0,82;0,02)	0.81	0.81	0.85	0.83	0.82	0.83	0.79	0.80	0.82	0.81	0.82	0.81-0.83
6 (0,82;0,04)	0.79	0.92	0.86	0.80	0.77	0.85	0.86	0.82	0.81	0.74	0.82	0.79-0.86

Testentscheidung bei $\alpha_0 = 0,8$:
Die Nullhypothese (vgl. Kapitel 8.4) kann lediglich für das Sample 5 zu einem Signifikanzniveau von β = 0,05 verworfen werden (siehe Tabelle 3). Für die anderen Sample kann die Nullhypothese nicht verworfen werden und somit liegt keine reliable Skala im verbalen Sinn vor.

Testentscheidung bei einer adjustierten α-Grenze:
Die adjustierte α-Grenze beträgt 0,73, für n = 10. Die Anzahl der Items wird bei diesem gewählten Beispiel auf zehn festgesetzt, da zehn α_i-Werte berechnet wurden. Hier zeigt sich, dass für die Sample 1 - 6 die Nullhypothese verworfen werden kann. Des Weiteren wurde der in dem R-Package „psych“ implementierte Datensatz „iqitems“ zur Analyse verwendet. Es handelst sich hierbei um einen Eignungstest mit 16 Items vom Synthetic Aperture Personality Assessment (SAPA), ein web-basiertes Projekt zur Abschätzung der Persönlichkeit der Probanden. Das Sample umfasst 1525 Versuchspersonen. Es wurden hier entsprechend die α_i-Werte berechnet (0.75, 0.74, 0.73, 0.74, 0.73, 0.73, 0.74, 0.74, 0.74, 0.73, 0.75, 0.75, 0.74, 0.74, 0.76, 0.75, 0.73). Es ergibt sich ein Wert von $\overline{\alpha}$ von 0.74. Das Konfidenzintervall liegt zu einem Konfidenzniveau von 0,95 zwischen [0.739, 0.745].

Testentscheidung bei $\alpha_0 = 0,8$:
Die Nullhypothese kann nicht verworfen werden. Die Skala ist hierdurch nicht reliabel im verbalen Sinn.

Testentscheidung bei einer adjustierten α-Grenze:
Die adjustierte α-Grenze beträgt 0,75, für n = 16. Die Nullhypothese kann auch in diesem Fall nicht verworfen werden, falls man als Ausgangsbasis ein α-Grenze von 0,8 wählt.

9.2 Korrelation zwischen dem wahren Wert und dem Fehler

In diesem Abschnitt wird analysiert, welche Auswirkungen sich auf den Koeffizienten α bzw. auf die α_i ergeben, falls $\sigma(\tau, \epsilon) \neq 0$. Des Weiteren wird simuliert, welche Konsequenzen sich auf den durchschnittlichen α-Wert und auf das α-Konfidenzintervall ergeben. Zur Analyse werden die Sample 3 und 5 aus der Tabelle 3 verwendet. Von Interesse erscheint die Frage, wie hoch die (positive bzw. negative) Kovarianz zwischen dem wahren Wert und dem Fehler sein muss, so dass sich die Testentscheidung ändert. Es wird $\alpha_0 = 0.8$ festgesetzt.

Für die beiden Sample wird im Rahmen der Simulation eine negative bzw. positive Kovarianz zwischen dem wahren Wert und dem Fehler in unterschiedlicher Höhe angenommen. Es wird eine Kovarianz $\sigma(\tau, \epsilon)$ von 0,5 % - 2 % der α_i-Werte angenommen. Die ursprünglichen α_i-Werte werden entsprechend der Formel 67 adjustiert (α_ψ) und können den Tabellen 4 und 5 entnommen werden. Die auf Grundlage der adjustierten α_i-Werte berechneten adjustierten Konfidenzintervalle ($= KI_\psi$) sind so dann mit den Ergebnissen aus den ursprünglichen Stichproben 3 und 5 (siehe Tabelle 3) zu vergleichen. Hierbei handelt es sich in diesem Kontext um die nicht adjustierten Werte, die eine nicht korrigierte Kovarianz zwischen dem wahren Wert und dem Fehler enthalten.

Tabelle 4: Simulation, $\overline{\alpha}$ und α-Konfidenzintervall bei $\sigma(\tau, \epsilon) < 0$, Sample 3

$\sigma(\tau, \epsilon)$	$\alpha_{\psi 1}$	$\alpha_{\psi 2}$	$\alpha_{\psi 3}$	$\alpha_{\psi 4}$	$\alpha_{\psi 5}$	$\alpha_{\psi 6}$	$\alpha_{\psi 7}$	$\alpha_{\psi 8}$	$\alpha_{\psi 9}$	$\alpha_{\psi 10}$	$\overline{\alpha_\psi}$	KI_ψ
0.5%	0.79	0.85	0.82	0.79	0.78	0.82	0.82	0.80	0.80	0.76	0.80	0.790-0.820
1%	0.79	0.86	0.82	0.80	0.78	0.82	0.83	0.81	0.80	0.77	0.81	0.793-0.824
1.5%	0.79	0.86	0.83	0.80	0.78	0.82	0.83	0.81	0.81	0.77	0.81	0.796-0.826
2%	0.80	0.87	0.83	0.80	0.79	0.83	0.83	0.81	0.81	0.77	0.81	0.801-0.831

Tabelle 5: Simulation, $\overline{\alpha}$ und α-Konfidenzintervall bei $\sigma(\tau, \epsilon) > 0$, Sample 5

$\sigma(\tau, \epsilon)$	$\alpha_{\psi 1}$	$\alpha_{\psi 2}$	$\alpha_{\psi 3}$	$\alpha_{\psi 4}$	$\alpha_{\psi 5}$	$\alpha_{\psi 6}$	$\alpha_{\psi 7}$	$\alpha_{\psi 8}$	$\alpha_{\psi 9}$	$\alpha_{\psi 10}$	$\overline{\alpha_\psi}$	KI_ψ
0.5%	0.81	0.81	0.84	0.83	0.82	0.83	0.78	0.79	0.82	0.81	0.81	0.80-0.82
1%	0.81	0.80	0.84	0.83	0.81	0.83	0.78	0.79	0.81	0.81	0.81	0.7996-0.82

Falls also für das Sample 3 angenommen wird, dass eine negative Kovarianz von 2 % (oder höher) vorliegt, dann ergibt sich eine falsche Testentscheidung, da das adjustierte KI_ψ über der α_0-Grenze von 0.8 liegt (vgl. Tabelle 4). Wird also eine negative Kovarianz von mindestens 2 % nicht berücksichtigt, dann ergibt sich eine Unterschätzung der Reliabilität, die die Testentscheidung ändert. Die Simulation

wurde bei 2 % abgebrochen. Bis zu einer negativen Kovarianz von 1.5 % der α_i-Werte ändert sich die Testentscheidung nicht.

Falls für das Sample 5 angenommen wird, dass eine positive Kovarianz von 1.0 % (oder höher) vorliegt, dann ergibt sich in diesem Fall eine falsche Testentscheidung, da das adjustierte KI_ψ die α_0-Grenze von 0.8 einschließt (vgl. Tabelle 5). Wird also eine positive Kovarianz von mindestens 1.0 % nicht berücksichtigt, dann ergibt sich eine Überschätzung der Reliabilität, die die Testentscheidung ändert. Die Simulation wurde bei 1.0 % abgebrochen. Bis zu einer positiven Kovarianz von 0.5 % der α_i-Werte ändert sich die Testentscheidung nicht.

10 Resümee

Abschließend erfolgt ein Resümee über die Arbeit. Zu Beginn der Arbeit erfolgte eine Definition der Grundbegriffe (Überbrückungsproblem, psychometrische Tests, Indexbildung, wahrer Wert und Messfehler, Gütekriterien von Messungen). Anschließend wurde die klassische Testtheorie dargestellt und hieraus die Reliabilität definiert. Um die Reliabilität für empirische Daten berechnen zu können, wurde der Koeffizient α hergeleitet. Nach einer kurzen Beschreibung der Vorteile des Koeffizienten befasste sich die Arbeit mit den Schwierigkeiten des Koeffizienten. Die Ergebnisse hierzu werden hier nochmals kurz zusammengefasst.

Es wurde zuerst dargestellt, dass der Koeffizient α mit der Anzahl der verwendeten Items schwankt. Anschließend wurde die Verwendung einer pauschalen Grenze für die Reliabilität im verbalen Sinn kritisch betrachtet. Aufgrund der aufgezeigten Abhängigkeit des Koeffizienten α von der Anzahl der verwendeten Items sollte die Grenze für die Reliabilität im verbalen Sinn in Abhängigkeit von der Anzahl der verwendeten Items adjustiert werden. Ein entsprechender Vorschlag hierzu wurde im Rahmen der Arbeit formuliert.

Im Folgenden wurden die Auswirkungen von alternierenden Vorzeichen auf den Koeffizienten betrachtet, ehe auf die Messungen näher eingegangen wurde. Falls die Messungen nicht mindestens essentiell τ-äquivalent sind, dann entspricht der Koeffizient α nicht der Reliabilität im Sinne der klassischen Testtheorie, sondern gibt vielmehr eine untere Schranke für diese Reliabilität an.

Danach wurde der Fokus darauf gelegt, welche Auswirkungen sich bei korrelierten

Fehlern ergeben. Falls eine positive Kovarianz zwischen den Fehlern vorliegt, dann kommt es zu einer Überschätzung des Koeffizienten. In diesem Fall muss der Koeffizient α nicht mehr eine untere Schranke für die Reliabilität sein.

Es wurde entsprechend bei einer Verletzung der Annahme $\sigma(\tau, \epsilon) = 0$ vorgegangen. Falls sich hieraus eine Überschätzung des Koeffizienten ergibt, dann liegt jedoch des Weiteren eine untere Schranke für die Reliabilität vor. Im Anschluss erfolgte eine Betrachtung der sogenannten α-Maximierung, die aus Sicht des Autors dieser Arbeit kritisch zu sehen ist. Als Gegenvorschlag erfolgt die Berechnung eines durchschnittlichen α-Wertes. Durch die Idee der Leave-One-Out-Cross-Validation konnte ein α-Konfidenzintervall mit der Bootstrap-Methode konstruiert werden und ein entsprechender Test formuliert werden. Es kann hierdurch getestet werden, ob eine Skala reliabel im verbalen Sinn ist oder nicht.

Abschließend erfolgt eine Simulation. Es wurden für unterschiedliche Situationen α-Konfidenzintervalle berechnet und entsprechende Testentscheidungen über die Reliabilität im verbalen Sinn getroffen. Weiter wurde betrachtet, wie sich eine Verletzung von $\sigma(\tau, \epsilon) = 0$ auf das α-Konfidenzintervall auswirkt. Es wurde zudem analysiert, ab welcher Höhe der Kovarianz zwischen dem wahren Wert und dem Fehler eine differierende Testentscheidung getroffen werden muss.

Durch die Entwicklung der adjustierten α-Grenze, des adjustierten Koeffizienten α_ψ, des durchschnittlichen α-Wertes, des α-Konfidenzintervalls und des entsprechenden Tests konnte durch diese Arbeit ein Beitrag zur Weiterentwicklung des Koeffizienten α geleistet werden.

11 Literaturverzeichnis

- Amann, Herbert; Escher, Joachim (2006): Analysis I. 3. Auflage. Birkhäuser Verlag: Basel (Schweiz).
- Beauducel, Andre; Leue, Anja (2014): Psychologische Diagnostik. Hogrefe Verlag GmbH Co. KG: Göttingen.
- Bortz, J.; Döring, N.(2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler (4. Auflage), Springer Medizin Verlag: Heidelberg.
- Bücker, Rüdiger (2003): Statistik für Wirtschaftswissenschaftler. Oldenbourg Wissenschaftsverlag GmbH: München.
- Bühner, Markus (2004): Einführung in die Test- und Fragebogenkonstruktion. Pearson Studium: München.
- Cattell, R.B.; Tsujioka, B. (1964): The importance of factor-trueness and validity, versus homogenity and orthogonality in test scales. Educational and Psychological Measurement, 24, 3-30.
- Cho, Eunseong; Kim, Seonghoon (2014): Cronbachs Coefficient Alpha. Well Known but Poorly Understood. Organizational Research Methods, 1-24.
- Christensen, L. B.; Johnson, R. B.; Turner, L. A. (2011): Research methods, design, and analysis (11th ed.). Pearson: Boston.
- Churchill, G. A.; Peter, J. P. (1984): Research design effects on the reliability of rating scales: A meta-analysis. Journal of Marketing Research, 21, 360-375.
- Cortina, J. M. (1993): What is coefficient alpha? An examination of theory and application. Journal of Applied Psychology, 78, 98-104.
- Cronbach, L. J. (1951): Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.
- Crocker, L.; Alinga, J. (1986): Introduction to classical and modern test theory. Holt, Rimehart, and Winston: New York.
- Danner, Daniel (2015). Reliabilität: die Genauigkeit einer Messung. GESIS Leibniz-Institut für Sozialwissenschaften: Mannheim (SDM Survey Guidelines). DOI: 10.15465/sdm-sg011.

- Davison, A.C. (2003): Recent Developments in Bootstrap Methodology, Statistical Science, Vol. 18, No. 2, 141-157.
- Diekmann, Andreas (2010): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Rowohlt Taschenbuch Verlag: Hamburg.
- Elfron, B. (1979): Bootstrap-Methods: Another Look at the Jacknife. The Anals of Statistics. Volumne 7, Number 1, 1-26.
- Efron, B.; Tibshirani R.J. (1993): An Introduction to the Bootstrap. Chapman and Hall: New York.
- Fan, Sue, Man (2010): Multiple Tests für die Evaluation von Regressionsmodellen. Eine Analyse am Beispiel der Prognose von Vermögenspreisen. Eckart Bornshof, Wim Köstens, Winfried Mathes und Mark Trede (Hrsg.). Josef Eul Verlag GmbH: Köln.
- Ghiselli, Edwin (1964): Theory of psychological measruement. McGraw-Hill Book Company: United States of America.
- Gulliksen, Harold (1967): Theory of Mental Tests. John Wiley, Sons: New York.
- Green, Samuel B.; Lissitz, R.W.; Mulaik, S. (1977): Limitations of coefficient alpha as an index of text unidimensionality. Educational and Psychological Measurement, 37, 827-839.
- Green, Samuel B.; Yang Yanyun (2009): Commentary on coefficient alpha: A cautionary tale. Psychometrika, 74, No. 1, 121-135.
- Henrad, D., H. (2000): Item Response Theory. In L.G. Grimm, P.R. Yanold (Hrsg.) Reading and understanding MORE multivariate statistics. American Psychological Association: Washington, DC.
- Heumann, Christian (2013): Schätzen und Testen I. http://www.statistik.lmu.de/ bothmann/st1_201314/Vorlesung/Skript/Kapitel_2.pdf, Abruf: 09.02.2015.
- Heumann, Christian, Schmid, Volker (2014): Schätzen und Testen II. http://www.statistik.lmu.de/ bothmann/st2_2014/Vorlesung/Skript/Kapitel_5.pdf, Abruf: 09.02.2015.

- Humphreys, L. (1956): The normal curve and the attenuation paradox in test theory. Psychological Bulletin, 53(6), 472-476.
- Huysamen, G.K (2006): Recent proposals to estimate the classical Test Theory Tradition. SA Journal of Industrial Psychology, 2006, 32 (4), 41-47.
- Jöreskog, K.G. (1971): Statistical analysis of sets of congeneric tests. Psychometrika, 36, 153-160.
- Kline, P. (1986): A handbook of test construction: Introduction to psychometric design. Methuen: London.
- Komaroff, E. (1997): Effect of simultaneous violations of essential tau equivalence and uncorrelated errors on coefficient alpha. Applied Psychological Measurements, 21, 337-348.
- Kopalle, P. K.; Lehmann, D. R. (1997): Alpha inflation? The impact of eliminating scale items on Cronbachs alpha. Organizational Behavior and Human Decision Processes, 70(3), 189-197.
- Krosnick, J. A. (1999): Survey research. Annual review of Psychology, 50, 537 - 567.
- Kristof, W. (1974): Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 39, 494 - 499.
- Lienert, G.A.; Raatz, U. (1998): Testaufbau und Testanalyse, 6. Aufl., Beltz: Weinheim.
- Liu, Yan; Zumbo, Bruno, D. (2007): The Impact of Outliers on Cronbachs Coefficient Alpha Estimate of Reliability: Visual Analogue Scales. Educational and Psychological Measurement 2007, 67, 620-634.
- Loevinger, J. (1954): The attenuation paradox in test theory. Psychological Bulletin, 51(5), 493-504.
- Loevinger, Jane (1957): Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.
- Mortensen, U. (2009): Einführung in die Theorie psychometrischer Tests. Skriptum zum Kompaktkurs Testtheorie am Psychologischem Institut der Johannes Gutenberg, Universität Mainz.
  http://psymet03.sowi.uni-mainz.de/meinharg/Lehre/Scripte/TT/TT-ItemTheorie-MainzApril2009-Mortensen.pdf. Abruf: 09.02.2015.

- Nassehi, Armin (2008): Soziologie. Zehn einführende Vorlesungen. Verlag für Sozialwissenschaften: Wiesbaden.
- Nunnally, J. C. (1978): Psychometric theory (2nd ed.). McGraw-Hill: New York.
- Nunnally, J. C.; Bernstein, I. H. (1994): Psychometric theory (3rd ed.). McGraw-Hill: New York.
- Osburn, H.G. (2000): Coefficient alpha and related internal consistency reliability coefficients. Psychological Methods, 5, (3), 343-355.
- Pauls, Thorsten (2013): Resampling-Verfahren und ihre Anwendungen in der nichtparametrischen Testtheorie. Books on Demand GmbH: Norderstedt.
- Peterson, Robert, A. (1994): A Meta-analysis of Cronbachs Coefficient Alpha. Journal of Consumer Research, Vol. 24, 381-391.
- Pflaumer, Peter; Heine, Barbara; Hartung, Joachim (2001): Statistik für Wirtschafts- und Sozialwissenschaften: Induktive Statistik. R. Oldenbourg Verlag: München, Wien.
- Pospeschill, Markus (2013): Empirische Methoden in der Psychologie. Ernst Reinhardt GmbH und Co. KG: München.
- Pruscha, Helmut (2000): Statistisches Methodenbuch. Verfahren, Fallstudien, Programmcodes. Springer: Berlin, Heildelberg, New York.
- Rae, Gordon (2006): Correcting Coefficient Alpha for Correlated Errors: Is α_κ a lower bound to reliability? Applied psychological measurement, 2006, 30, 56-59.
- Raykov, T. (1997): Estimation of composite reliability for congeneric meassures. Applied Psychological Measurement, 21, 173-184.
- Raykov, T. (2007): Reliability if deleted, not alpha if deleted: Evaluation of scale reliability following component deletion. British Journal of Mathematical and Statistical Psychology, 60(2), 201-216.
- Reise, S. P.; Waller, N. G.; Comrey, A. L. (2000): Factor analysis and scale revison. Psychological Assesment 12, 287-297.

- Reise, S. P.; Morizot, J.; Hays R. D. (2007): The rule of the bifactor model in resolving dimensionality issus in health outcomes measures. Quality of Life Research 16, 19-31.
- Rost, Jürgen (1996): Lehrbuch Testtheorie, Testkonstruktion. Verlag Hans Huber: Bern.
- Rozeboom, W. (1966): Foundations of the theory of prediction. Homewood: Dorsey.
- Rubin, A.; Babbie, E. R. (2008): Research methods for social work (6th ed.). Thompson Books/Cole: Belmont.
- Rüger, Bernhard (1999): Test- und Schätztheorie, Band I: Grundlagen. R. Oldenbourg Verlag: München, Wien.
- Schmitt, N. (1996): Uses and abuses of coefficient alpha. Psychological Assessment, 8, 350-353.
- Schnell, Rainer; Hill, Paul, B.; Esser, Elke (1995): Methoden der empirischen Sozialforschung. R. Oldenbourg Verlag: München, Wien.
- Sijtsma, Klaas (2009): On the Use, the misuse, and the very limited usefulness of Cronbachs Alpha. Psyhometrika, 74, No. 1, 107-120.
- Stanley, J., C. (1971): Reliability. in R.L. Thorndike (Hrsg.), Educational measurement, 356-442. American Council in Education: Washington DC.
- Statistisches Bundesamt (2004): Datenreport 2004. Zahlen und Fakten über die Bundesrepublik Deutschland. In Zusammenarbeit mit dem Wissenschaftszentrum Berlin für Sozialforschung (WZB) und dem Zentrum für Umfragen, Methoden und Analysen, Mannheim (ZUMA). Bonn.
- Steinberg, L. (2001): The consequences of pairing questions: Context effects in personality measurement. Jounal of Personality and Social Psychology, 81, 332-342.
- Steyer, Rolf; Eid, Michael (1993): Messen und Testen. Springer-Verlag: Berlin Heidelberg.
- Thorndike, R. L. (1964): Reliability. In Proccedings of the 1963 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 23-32.

- Traub, R., E. (1994): Reliability for the social sciences. Thousand Oaks CA: Sage.
- Waller, Niels (2008): Commingled Samples: A Neglected Source of Bias in Reliability Analysis. Applied Psychological Measurement, Vol. 32, No. 3, 211-223.
- Winkler, Bernd (1998): Bootstrap-Methoden bei nichtparametrischer Regression. Forschungsbericht aus dem Institut für Statistik der Universität München. Serie WiSo No. 1.
- Weber, Max (1980): Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie, 5. Auflage, Mohr: Tübingen.
- Yen, W.M. (1993): Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurements, 30, 187-214.
- Yang, Yanyun; Grenn, Samuel B. (2011): Coefficient Alpha: A Reliability Coefficient for the 21st Century? Journal of Psychoeducational Assessment, 29(4), 377-392.
- Zimmerman, D. W.; Zumbo, B.D.; Lalonde, C. (1993): Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and Psychological Measurements, 53, 33-49.

Eigenständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Masterarbeit eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Masterarbeit ist in dieser oder einer ähnlichen Form in keinem anderen Kurs und/oder Studiengang als Studien- oder Prüfungsleistung vorgelegt worden. Hiermit stimme ich zu, dass die vorliegende Arbeit von der Prüferin/ dem Prüfer in elektronischer Form mit entsprechender Software überprüft wird.

München, den .....

.....

(Andreas Bauer)