Article

JHP

Journal of Health Psychology
2014, Vol. 19(4) 544–555

DOI: 10.1177/1359105313475897
hpq.sagepub.com
SAGE

Rasch analysis of the General Self-Efficacy Scale in spinal cord injury

Claudio Peter[1,2], Alarcos Cieza[1,2,3] and Szilvia Geyh[1,4]

Abstract

This study examined the psychometric properties of the General Self-Efficacy Scale by applying Rasch analysis to data from 102 persons with spinal cord injury. Our results suggest that the General Self-Efficacy Scale is a psychometrically robust instrument suitable for application in a spinal cord injury population. The General Self-Efficacy Scale shows an overall fit to the Rasch model (χ^2 = 15.5, *df* = 20, *p* = .75), high reliability (r_p = 0.92), ordered response scale structure, and no item bias by gender, age, education, and lesion levels. However, the analyses indicate a ceiling effect and potential to enhance the differentiation of the General Self-Efficacy Scale across self-efficacy levels.

Keywords

psychometrics, Rasch analysis, rehabilitation, self-efficacy, spinal cord injury

Introduction

Spinal cord injury (SCI) is a health condition with severe consequences on a physical, social, and psychological level (Kirshblum et al., 2002). These difficulties represent huge challenges for the affected persons and may also have a severe impact on daily activities and participation.

The question of whether certain persons with distinct personal attributes may show a quicker recovery than others and less adjustment problems from disease or trauma such as SCI has been extensively examined (Atkinson et al., 2009). Self-efficacy is one personal attribute that has been associated with resilience in the face of an adverse event and is seen as a key psychological resource (Hobfoll, 2002).

Self-efficacy is defined as the person's belief in his or her own ability to perform a behavior (Bandura, 1977). Associations of high self-efficacy with better well-being, mental health, and health behavior were observed for persons with SCI and other chronic health conditions such as cancer or arthritis (Marks, 2001; Park

[1]Swiss Paraplegic Research (SPF), Switzerland
[2]Ludwig-Maximilians-University, Germany
[3]University of Southampton, UK
[4]University of Lucerne, Switzerland

Corresponding author:
Claudio Peter, Swiss Paraplegic Research (SPF), Guido A. Zäch-Strasse 4, CH-6207 Nottwil, Switzerland.
Email: claudio.peter@paraplegie.ch

and Gaffey, 2007). Also, self-efficacy improved in several intervention studies, which emphasizes the relevance of this concept for clinical practice in SCI and other health conditions such as HIV or mental illness (Antoni, 2003; Hansson, 2006). Strengthening resources such as self-efficacy is an important aim in psychotherapy (Flückiger et al., 2010) and can help, for example, to achieve improvements in health behavior (Schwarzer, 2008), which could in turn help to avoid secondary complications in SCI, such as pressure sores or urinary tract infections (Kirshblum et al., 2002; Lin, 2003). Intervention decisions, treatment selection, and evaluation are often supported by standardized assessment (Gadotti et al., 2006; Vianin, 2008). Therefore, researchers and practitioners depend on reliable and psychometrically sound measurement instruments.

A popular measurement instrument of general self-efficacy is the General Self-Efficacy Scale (GSES) (Schwarzer and Jerusalem, 1995). It has been translated into 30 languages and has been extensively used in health research (see www.ralfschwarzer.de). Psychometric studies based on classical test theory (CTT) show satisfactory results, indicating that the GSES is reliable, valid, and unidimensional in both the general population (Luszczynska et al., 2005; Scholz et al., 2002; Schwarzer et al., 1997) and SCI population (Kennedy et al., 2006). However, hardly any research has been conducted employing item response theory (IRT), that is, modern test theoretical techniques such as Rasch analysis, to examine the psychometric properties of this measurement instrument (Scherbaum et al., 2006).

Modern test theoretical models were introduced in the 1960s (Lord and Novick, 1968; Rasch, 1960). Georg Rasch, a Danish mathematician with background in biological and medical statistics, worked in a study examining the progress of reading abilities of pupils (Rasch, 1960). In the data analysis, Rasch developed a model where item parameters could be estimated independent from the person parameters (Rasch, 1960). Rasch's model is a well-known, distinct approach within modern test theory (Andrich, 2004).

Rasch-based analysis has several advantages over CTT (Andrich, 1988; Geyh et al., 2010; Tesio, 2003; Wright and Linacre, 1989). First, Rasch analysis transforms ordinal scale observations into interval scale measures, which is the prerequisite for the meaningful measurement of change in patients, ensuring the additivity of the total score. In CTT, the additivity of the total score is simply postulated.

Second, a Rasch-based approach provides refined information on validity by using a reference that is external to the data. This external reference is the model itself, which in its mathematical formulation holds the requirements for fundamental measurement. In CTT, factor analysis is conducted to test for dimensional validity because the item scale values are population and data dependent.

Third, Rasch analysis also enables the evaluation of response scale validity within a probabilistic framework. By yielding sample- and test-independent estimates of person and item parameters placed on the same continuum, Rasch analysis provides an index of reliability that is independent of sample distribution. In contrast, reliability estimations based on CTT are dependent upon the test length and the distribution of the sample. Today, modern test theoretical techniques are considered standard in test theory due to the "more theoretically justifiable measurement principles and the greater potential to solve practical measurement problems" than CTT (Embretson and Reise, 2000: 3).

The objective of this study is to examine the psychometric properties of the GSES using Rasch analysis in a German-speaking sample with SCI living in Switzerland. More specifically, the aims are (a) to test unidimensionality, (b) to test the reliability, (c) to test the structure of the response scale, (d) to examine the targeting of the instrument, and (e) to check for item bias or differential item functioning (DIF) with regard to age, gender, education, and level of injury.

Methods

Study design and participants

The psychometric evaluation of the GSES was conducted using cross-sectional data from a multicenter study including person with SCI living in the community. Participants were recruited through three major SCI rehabilitation centers in Switzerland (University Clinic Balgrist, Paraplegic Centre, Zurich; Swiss Paraplegic Centre, REHAB Basle; Swiss Paraplegic Centre (SPZ), Nottwil). Data were collected by means of a self-report questionnaire sent to the eligible participants by postal mail. Design and study materials were approved by the ethical committees of the cantons Lucerne, Basle, and Zurich.

Persons with SCI were eligible when they were German speaking, older than 18 years, and discharged from first rehabilitation since at least half a year. Persons with a progressive neurological disorder, a neoplasm of the spine, or a concurrent neurological condition that affected mental functions were excluded. Every participant signed a consent form.

In the data collection, the sociodemographic variables such as age, gender, education, and marital status and lesion-related information such as level, completeness, and etiology of injury on each patient were included. The German version of the GSES was used as an outcome measure (Schwarzer and Jerusalem, 1999).

GSES

The GSES consists of 10 items assessing a general belief in the own ability. For example, item 4 is phrased "I am confident that I could deal efficiently with unexpected events" (Schwarzer and Jerusalem, 1995). Items are assessed on a 4-point response scale with 1 = not at all true and 4 = exactly true. The responses to all 10 items are summarized to form a total score, ranging from 10 to 40 points, where a higher score indicates higher self-efficacy. Overall, classical test theoretical examinations of the psychometric criteria report satisfactory reliability and validity (Luszczynska et al., 2005; Scholz et al., 2002; Schwarzer et al., 1997). Cronbach's alpha in a study comparing the GSES scores of 25 countries were ranging from $.75 < \alpha < .91$ (Scholz et al., 2002). Also, correlations with depression or optimism provided evidence for validity (Schwarzer et al., 1997).

Rasch analyses

Rasch analyses were conducted with RUMM2030 software (Andrich et al., 2009). Rasch analysis estimates person parameters, the item parameters, and the parameters of the thresholds of the response scale (e.g. a 4-point Likert scale).

These parameters describe the position of the persons, items, and thresholds on the continuum of the measured unidimensional latent trait, that is, low to high self-efficacy. Therefore, the parameters are directly comparable because they are placed on one continuum sharing a common metric (logit) scale. They are regarded as sufficiently describing the response pattern in an item–person encounter. The estimation of the parameters is, however, dependent on the sample size. The higher the sample size the more stable are the item calibrations (Linacre, 1994). For example, with a sample of 50 persons, the estimated item difficulties are within one logit of their stable value with a 95 percent confidence interval (CI), which is considered close enough for most practical purposes (Wright and Tennant, 1996).

First, unidimensionality of the measurement instrument was studied. Unidimensionality is an important aspect of construct validity. It means that items contribute to the measurement of only one single attribute (Bond and Fox, 2001). If data fit the Rasch model, the person estimates are interval scale-level measures unbiased by the sample distribution, and the additivity of the score is ensured (Tennant and Pallant, 2006). Unidimensionality can be checked by comparing the observed responses in a set of items with the expected values predicted by the Rasch model (Andrich, 1988; Bond and Fox, 2001; Wright and Linacre, 1989; Wright and Masters, 1982). The fit of each item

is indicated by standardized residuals (Z values) and χ^2 test results. Z values exceeding ±2.5 were considered to indicate misfit to the Rasch model. To further test for unidimensionality, principal component analyses (PCAs) of the residuals were conducted. Given the study sample size, an eigenvalue higher than 1.9 indicates a certain structure in the residuals and thus multidimensionality (Raîche, 2005). Additionally, items were grouped according to their positive or negative loading with the first residual PCA factor. These two subgroups of items were compared for each individual using independent t-tests. The expected rate of significant tests due to chance lies at 5 percent. A percentage rate including CI overlapping the 5 percent ($\alpha = .05$) suggests unidimensionality (Smith, 2002).

A further threat to unidimensionality is local dependency, which implies that items are interrelated other than by the trait they measure (Baghaei, 2008). Residual correlations between all item pairs should thus be zero. A correlation value higher than .3 between a pair of item suggests local dependency (Wright, 1996). It can be controlled for by adding locally dependent items together to create item subgroups (testlets) or by deleting one item of the item pair.

Reliability was examined with the person reliability index. It represents an analogous value to Cronbach's alpha and ranges between 0 and 1, where the value of 1 indicates perfect reproducibility of person placements (Wright and Masters, 1982). The person reliability index is constructed using the measurement error and the observed variance associated with the person parameters to calculate the ratio of "true" variance to the observed variance (Fisher, 1992).

The structure of the response scale was studied with reference to the ordering of the threshold parameters for each individual item's response scale. Thresholds are boundaries between response categories. The threshold parameters should reach increasing values, as they represent successive transition points along the response scale. Reversed thresholds indicate that the response scale does not work as intended (Linacre, 2002). In addition, the distribution of the responses across the response categories is examined. With fewer than 10 observations in a response category, the threshold parameters may be imprecise (Linacre, 2002). Graphical probability curves of every item were studied to examine the structure of the response scale.

The targeting of the GSES was studied. First, the respective distribution of the person, item, and threshold parameters along the latent trait continuum was examined. Second, the percentage of persons with measures below the level of the lowest threshold, and of those with measures above the level of the highest threshold, was calculated. Third, the distance between the mean person location and the mean item location was analyzed. Ninety-five percent CIs around the means were calculated to further evaluate floor and ceiling effects (Bond and Fox, 2001). Fourth, person strata index indicating the number of identified distinct ability levels was calculated using the formula $[(4G + 1)/3]$ (Wright and Masters, 1982).

DIF, or item bias, was examined to check for the invariance of the item parameters across each of four person groups: gender (male vs female), age (young vs old), education (high vs low), and level of lesion (para- vs tetraplegia). DIF analyses allow the validity of items across different patient groups to be assessed. For example, it could be hypothesized that tetraplegic persons experience higher limitations in daily activities and participation as a consequence of their injury, which might also have an effect on their level of self-efficacy. Therefore, items need to be equally suitable and "behave" in the same expected way in both para- and tetraplegic persons. Potential DIF is ascertained for each item by the comparison of the standardized residuals between the groups and across the person parameter continuum using a two-way analysis of variance (ANOVA). A significant main effect of the group (e.g. gender) or an interaction effect in the ANOVA results (e.g. gender × self-efficacy) is an indicator of item bias. Bonferroni-corrected type I error level was used to identify DIF, correcting for the multiple significance tests conducted (Bland and Altman, 1995; Hagquist and Andrich, 2004).

Table 1. Sociodemographic and lesion-related data of the study population (N = 101).

	n	%
Age (mean in years)	56.28	
Gender	101	
Male	76	75.2
Female	25	24.8
Marital status	99	
Single	19	19.2
Separated	9	9.1
Widowed	6	6.1
Married/partnership	65	65.6
Education (mean in years)	13	
Occupational status	99	
Remunerative employment	46	46.5
No employment	9	9.1
Retired	34	34.3
Other (house wife, education, etc.)	10	10.1
Level of lesion	100	
Cervical	37	37.0
Thoracal	41	41.0
Lumbal	19	19.0
Sacral	3	3.0
Completeness and level of lesion	101	
Complete paraplegia	24	23.8
Complete tetraplegia	3	3.0
Incomplete paraplegia	38	37.6
Incomplete tetraplegia	36	35.6
AIS score	93	
A	29	31.2
B	13	14
C	15	16.1
D	36	38.7
Time since injury (mean in months)	43.5	

Numbers do not necessarily add up to 101 because of missing values.
AIS score: The ASIA Impairment Scale (AIS) categorizes motor and sensory impairment in individuals with SCI. A: complete spinal cord injury with no motor or sensory function in the sacral segments; B: incomplete spinal cord injury where sensory but not motor function is preserved below the neurological level; C: incomplete spinal cord injury where motor function is preserved below the neurological level and more than half of key muscles below the neurological level have a muscle grade of less than 3, which indicates active movement with full range of motion against gravity; D: incomplete spinal cord injury where motor function is preserved below the neurological level and at least half of the key muscles below the neurological level have a muscle grade of 3 or more (American Spinal Injury Association, 2011).

Results

A total of 102 persons with SCI from three rehabilitation centers participated in this study. One person did not fill in the questionnaires accurately, leading to a total number of 101 study participants. Sociodemographic and lesion-related data are presented in Table 1. Overall, persons with SCI attained a mean total score of 31.6 (standard deviation (SD) = 6.92) in the GSES (Table 2).

Of the 101 respondents, 9 scores represented extreme cases. Of these nine cases, seven persons achieved the highest possible total score

Table 2. Raw scores and Rasch-based fit statistics, ordering of the response scale thresholds, and reliability (N = 101).

Items	M (SD)	δ	SE	Z	χ^2	df	p	τ	r	Threshold 1	Threshold 2	Threshold 3
Overall	31.6 (6.92)[a]				15.5	20	.75	4-step scale	0.92			
Item 1: If someone opposes me	3.25 (0.73)	−2.810	0.242	2.575	2.979	2	.23	ord		−10.41	−1.38	3.36
Item 2: Manage to solve difficult problems	3.18 (0.76)	−2.610	0.236	0.146	2.258	2	.32	ord		−10.61	−0.74	3.53
Item 3: Stick to aims	2.92 (0.95)	1.018	0.198	−0.315	1.260	2	.53	ord		−1.54	0.37	4.23
Item 4: I am confident	2.94 (0.79)	0.739	0.222	−0.014	2.375	2	.31	ord		−2.67	−0.37	5.25
Item 5: Handle unforeseen situations	3.05 (0.88)	0.485	0.206	−0.332	1.390	2	.50	ord		−2.17	−0.31	3.93
Item 6: I can rely on coping abilities	3.02 (0.86)	0.541	0.209	0.514	0.826	2	.66	ord		−2.58	−0.60	4.27
Item 7: Whatever comes my way	3.06 (0.86)	0.333	0.212	−1.501	0.608	2	.74	ord		−2.78	−0.23	4.01
Item 8: Find several solutions	2.92 (0.90)	1.088	0.204	−0.471	0.359	2	.84	ord		−1.40	−0.25	4.90
Item 9: If I am in trouble	2.94 (0.75)	0.498	0.235	−2.196	3.243	2	.20	ord		−3.94	−0.11	5.54
Item 10: Invest the necessary effort	2.86 (0.74)	0.718	0.237	0.118	0.206	2	.90	ord		−4.08	0.36	5.88

M (SD): mean and standard deviation of raw scores; δ: item location in logits; SE: standard error of item location; Z: standard normal distributed test value Z; df: degrees of freedom; p: probability; τ: ordering of the response scale thresholds; ord: ordered response scale; r: person reliability index; Threshold 1: location of threshold 1 on logit continuum; Threshold 2: location of threshold 2 on logit continuum; Threshold 3: location of threshold 3 on logit continuum.
[a]Kolmogorov–Smirnov, $\alpha > .05$ for GSES total score, indicating normal distribution.

(=40) and two persons the lowest possible total score (=10).

The GSES showed an overall fit to the Rasch model, indicating unidimensionality (Table 2). The χ^2 test was not significant. Likewise, the items fit to the Rasch model. Only item 1 slightly exceeded the critical standardized residual level; however, the χ^2 test for this item was not significant. The residual PCA eigenvalue had a value of 1.76 indicating unidimensionality. The rate of significant *t*-tests with the two item subsets was 9.78 percent (CI: 3.71–15.85) and overlapped with the 5 percent criterion. Testing the 45 possible item pairs for local dependency yielded only two correlation coefficients slightly higher than .3: items 2 and 6 (−.34) and items 4 and 7 (−.39). Local dependency was adjusted for by creating item subgroups and by item deletion. However, residual item correlations ranging from −.30 to −.32 repeatedly emerged. The person reliability index had a value of 0.92 (0.97 with extreme cases included), which indicates high reliability.

The structure of the response scale was studied based on the ordering of the threshold parameters for each individual item's response scale. No reversed thresholds on any item were observed; the thresholds showed the expected pattern of increasing values. With regard to the number of observed responses per category, the first category representing the lowest level of self-efficacy ("not at all true") was selected by less than 10 persons in 8 out of 10 items (items 1, 2, 4, 5, 6, 7, 9, 10) and by no one in 2 items (items 1 and 2). All other categories of all items were selected by at least 10 persons.

The thresholds of every item were inspected by examination of the graphical probability curves. Overall, the four categories of all items functioned well. However, the graphical probability curve for item 1 (If someone opposes me) and item 2 (Manage to solve difficult problems) only worked when including the extreme cases. If excluding the two extreme cases with a very low self-efficacy level, the first response category would have never been the most probable for both items (since no participant marked this response category).

The mean difference between the location of the thresholds was 3.98 logits (between thresholds 1 and 2) and 4.82 logits (between thresholds 2 and 3). This mean difference lies within the recommended range of 1.4 and 5 logits (Linacre, 1999). However, threshold distances of several items exceeded the suggested range (Table 2). The thresholds of the first two response categories (not at all true, hardly true) of all items were located on the lowest part of the continuum and thresholds of the third (moderately true) and fourth response categories (exactly true) on higher self-efficacy levels.

The width of the latent metric of 16 logits is high. This could reflect a distributional problem and thus be a consequence of null categories within the response options, as the first category ("not at all true") for items 1 and 2 was never chosen. If these two thresholds are left aside, the metric displays a width of 10 logits. Two testlets comprising positive and negative loading items on the first residual PCA factor were created. The width of the metric shrank and lied between −5 and 4 logits.

To specify targeting and to examine floor and ceiling effects, the distribution of the person and item parameters along the latent trait continuum was examined first. Item means were not located along the whole continuum but appeared to be "clustered" in two groups (Figure 1). Eight item means were located within one logit and two item means were located about three logits lower on the self-efficacy continuum. Of the 101 participants, 67 persons (66.3%) were located higher than the highest mean item location (item 8). Item thresholds were spread along the logit continuum. However, a cluster trend with threshold 1 lying between −4 and −2, threshold 2 around 0, and threshold 3 around 4 and 5 on the logit scale was observed (Table 2). Second, the percentage of the persons below the level of the lowest threshold and of those above the level of the highest threshold was calculated (all study participants included). Of the original 101 scores, 2 persons (2%) scored below the lowest threshold, while 17 persons (16.8%) scored higher

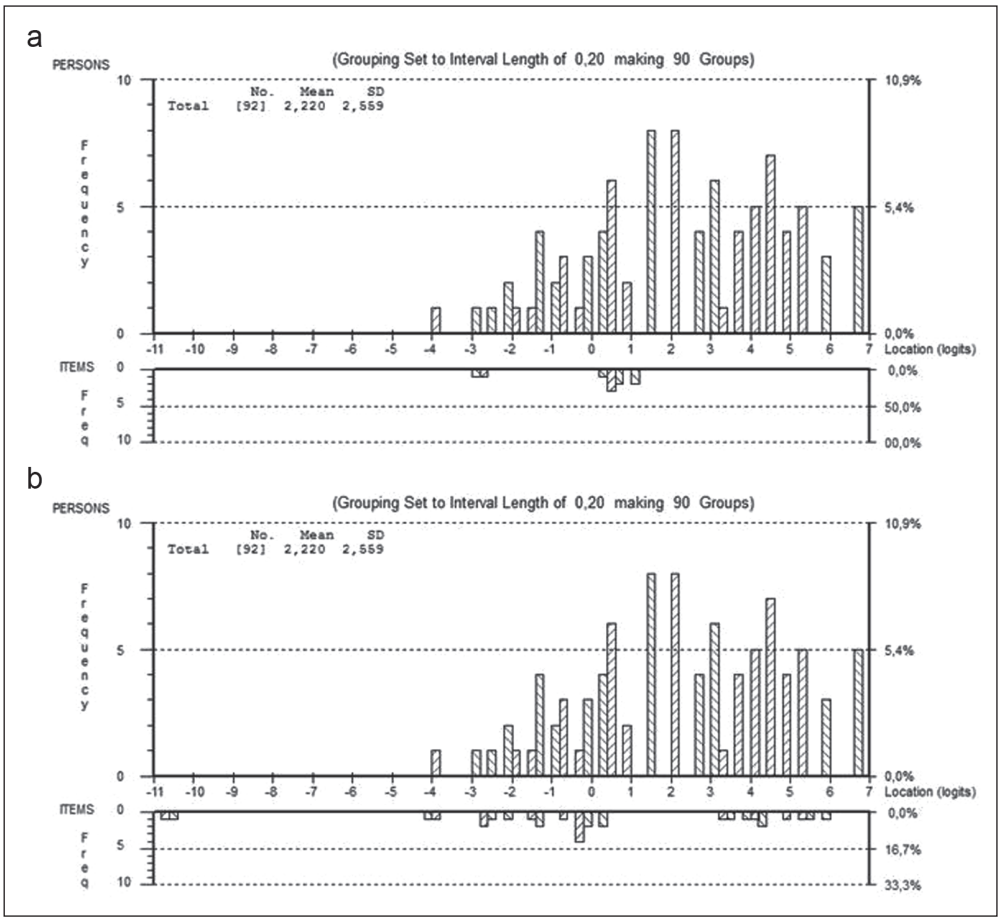

Figure 1. a Person–item location distribution and b person–item threshold distribution (n = 92). Extreme cases are not depicted.

than the highest threshold. Third, the distance between the mean person parameter and the mean item parameter was examined. The mean person parameter had a value of 2.24 logits (CI: 1.70–2.78 logits). The mean item parameter is 0 by definition, the CI ranged from −0.30 to 0.30. Fourth, person strata were calculated. Five strata could be distinguished. Altogether, these results indicate a ceiling effect. The participant's self-efficacy was higher than that captured by the items.

Overall, DIF was not indicated. The ANOVA of the residuals did not show any effects for age, gender, education, and level of lesion. A significant age effect was discovered for item 7. However, the deviation from the ICC was marginal and in the higher interval. A removal of the item is not indicated.

With the clustering of the mean difficulties and thresholds, the items might appear redundant. It can be argued that the GSES could be shortened. To examine this, we performed a post hoc exploratory Rasch analysis including five items of the GSES selected to maximize spread across the logit continuum (items 1, 4, 5, 7, 8), which resulted in a satisfactory reliability of 0.82, ordered thresholds, and DIF for age in items 4 and 7, but suggested local dependency

for items 1, 5, and 8 ($-.37 < r < -.32$) and items 4 and 7 ($r = -.46$).

Discussion

The current study was the first examination of the psychometric quality of the GSES applying a Rasch-based methodology. The GSES proved to be a unidimensional and reliable instrument in SCI. The response scale structure was ordered. All items worked consistently across gender, age, education, and lesion levels. However, the results indicated that targeting of the GSES is problematic and the differentiation across self-efficacy levels could be enhanced.

First, the items were too easy and demonstrated a ceiling effect given the level of self-efficacy in the current sample of persons with SCI. This is consistent with the findings from a study, which examined the metric properties of the GSES in psychology students also using an IRT approach (Scherbaum et al., 2006).

Second, most items did not differ in their level of difficulty, that is, all but two item mean difficulties laid close to each other within the range of one logit. Thus, the item mean difficulties did not constitute a linear continuum progressing from low to high self-efficacy but were clustered around one point of the self-efficacy logit scale. This might be explained by the similarities of the semantic structure and almost synonymous phrasing of the items. The ceiling effect and the low variation in *item mean difficulty* might pose a threat to the content validity of the GSES, that is, the extent to which the entire universe of the domain to be measured is represented.

In contrast, the thresholds, which specify the transition points between the response options (from "not at all true" to "hardly true," from "hardly true" to "moderately true," and from there to "exactly true") and which together constitute the item mean difficulty, have been found to be considerably distributed across a range of 16 logits. However, the distances between the thresholds were large with a mean of 4.2 logits, which indicates that additional response options might be advantageous and could enhance the precision of measurement (Linacre, 1999; Pishghadam et al., 2011). For most items, the thresholds were also clustered, that is, the third threshold laid consistently around the level of 4–5 on the logit scale, the second threshold around 0, the first around −2 to −4. More variation again would allow for a more fine-grained differentiation of the self-efficacy level.

Overall, while the items tended to cluster around one point on the continuum of the self-efficacy logit scale, the response options showed considerable spread. In terms of reliability, the findings indicate that the summary score of the GSES is capable of discerning five person strata, which supports the usefulness of the measure despite the problems in targeting.

Within CTT, reliability depends on the number of items while reliability is calculated independent of the number of items within probabilistic test theory and Rasch analyses (Embretson and Reise, 2000). The exploratory Rasch analysis with five items of the GSES resulted in a satisfactory reliability. A shortened GSES version could be of use in large surveys by reducing respondent burden and potentially increasing response rate. However, further studies are required to confirm if a reduced GSES would still provide measurements with robust psychometric quality.

Local dependency indicates that the items are to some extent redundant, which supports the idea of scale reduction (Kucukdeveci et al., 2012). However, handling these locally dependent items and reducing the scale produced new local dependency among items (or item testlets). Although local dependency was comparably small, it might have inflated reliability and influenced parameter estimates and the metric (Andrich et al., 2012; Baghaei, 2008; Lundgren-Nilsson and Tennant, 2011).

Across the analyses, items 1 and 2 appeared to behave distinctly from the others. Their item mean difficulty was lower and thus made up a second cluster of items. This is explained by the exceptionally low level of the first threshold, which, in turn, is a consequence of the fact that

the first response option "not at all true" was never selected for these two items, representing a targeting problem. This irregularity also contributed to the inflation of the width of the metric but cannot be attributed to a difference in the content of the items. It could be hypothesized that the ordering of the questions led to a bias, as they were prominently positioned as the first two items of the questionnaire, which might have affected the response pattern. Rotation of the item order could be used to test this assumption.

The sample size of this study is rather small. This may be connected with less precise and robust estimates and less powerful fit analysis (Linacre, 1994). The standard errors (Table 2) and the CIs of all items in our analyses were small, indicating robust parameters. However, ANOVA may have missed to detect DIF due to the small sample size or due to the sample imbalance (e.g. with regard to gender). The concurrent use of more than one approach was proposed to examine DIF in small samples (Lai et al., 2005). Thus, further testing with larger samples applying other approaches is needed to confirm the findings of this study.

From the analyses, several suggestions for potential improvement of the GSES can be derived. To enhance the coverage of the whole self-efficacy continuum, to avoid ceiling effects and clustering of the items, further items could be introduced, which are located at a lower or higher self-efficacy continuum level; items could be rephrased and restructured to counter the semantic similarities, for example, reversed items could be added; and redundant items removed. Because of the large threshold distances and the null categories within the response options for items 1 and 2, an adaptation of the response format could be indicated, for example, by introducing additional response categories.

Adaptation of the GSES might prove useful especially in clinical practice and rehabilitation. Enhancing self-efficacy can be an important aim in SCI rehabilitation as positive effects on health behavior and participation can be expected (Arbour-Nicitopoulos et al., 2009; Latimer et al., 2006). Assessment instruments can be used, for example, to identify persons with low self-efficacy who are at risk for unfavorable outcomes and who could benefit from self-efficacy interventions. They can also be used to monitor progress and evaluate intervention success (Gadotti et al., 2006; Vianin, 2008).

This study is subject to several limitations. The representativity of the study sample can be questioned because of the low response rate. However, responders and nonresponders did not differ in age, level, and completeness of injury, but nonresponders were more frequently women (data not shown). A comparably small sample size was used in this study. In addition, the study examined only basic psychometric properties of the GSES but could not attend to criteria such as sensitivity to change.

Overall, the GSES seems to be a psychometrically sound instrument. However, the analyses indicate that targeting could be improved. Future research should apply modern test theoretical approaches such as the Rasch methodology to complement traditional approaches and reevaluate and improve assessment. In the context of clinical practice as well as research, such reexaminations could benefit all users of the measurement instruments.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

American Spinal Injury Association (2011) American Spinal Injury Association (ASIA). Available at: http://www.asia-spinalinjury.org/index.php

Andrich D (1988) *Rasch Models for Measurement*. Newbury Park, CA: SAGE.

Andrich D (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care* 42(1 Suppl.): I-7–I-16.

Andrich D, Humphry SM and Marais I (2012) Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement* 36(4): 309–324.

Andrich D, Sheridan B and Luo G (2009) *RUMM 2030 (Beta Version for Windows)*. Perth, WA, Australia: RUMM Laboratory Pty Ltd.

Antoni MH (2003) Stress management and psychoneuroimmunology in HIV infection. *CNS Spectrums* 8(1): 40–51.

Arbour-Nicitopoulos KP, Ginis KA and Latimer AE (2009) Planning, leisure-time physical activity, and coping self-efficacy in persons with spinal cord injury: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation* 90(12): 2003–2011.

Atkinson PA, Martin CR and Rankin J (2009) Resilience revisited. *Journal of Psychiatric and Mental Health Nursing* 16: 137–145.

Baghaei P (2008) Local dependency and Rasch measures. *Rasch Measurement Transactions* 21(3): 1105–1106.

Bandura A (1977) Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84: 191–215.

Bland JM and Altman DG (1995) Multiple significance tests: The Bonferroni method. *British Medical Journal* 310: 170.

Bond TG and Fox CM (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson SE and Reise SP (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions*, 6(3), 238.

Flückiger C, Wüsten G, Zinbarg RE, et al. (2010) *Resource Activation: Using Clients' Own Strengths in Psychotherapy and Counseling*. Cambridge, MA: Hogrefe Publishing.

Gadotti IC, Vieira ER and Magee DJ (2006) Importance and clarification of measurement properties in rehabilitation. *Revista Brasileira de Fisioterapia* 10(2): 137–146.

Geyh S, Fellinghauer BAG, Kirchberger I, et al. (2010) Cross-cultural validity of four quality of life scales in persons with spinal cord injury. *Health and Quality of Life Outcomes* 8: 94.

Hagquist C and Andrich D (2004) Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences* 36: 955–968.

Hansson L (2006) Determinants of quality of life in people with severe mental illness. *Acta Psychiatrica Scandinavica* 113(Suppl. 429): 46–50.

Hobfoll SE (2002) Social and psychological resources and adaptation. *Review of General Psychology* 6(4): 307–324.

Kennedy P, Taylor N and Hindson L (2006) A pilot investigation of a psychosocial activity course for people with spinal cord injuries. *Psychology, Health & Medicine* 11(1): 91–99.

Kirshblum S, Campagnolo DI and DeLisa JA (2002) *Spinal Cord Medicine*. Philadelphia, PA: Lippincott Williams & Wilkins.

Kucukdeveci AA, Kutlay S, Yildizlar D, et al. (2013) The reliability and validity of the World Health Organization Disability Assessment Schedule (WHODAS-II) in stroke. *Disability and Rehabilitation*. 35(5): 214–220.

Lai J, Teresi J and Gershon R (2005) Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions* 28(3): 283–294.

Latimer AE, Ginis KAM and Arbour KP (2006) The efficacy of an implementation intention intervention for promoting physical activity among individuals with spinal cord injury: A randomized controlled trial. *Rehabilitation Psychology* 51(4): 273–280.

Lin VW (2003) *Spinal Cord Medicine: Principles and Practice*. New York: Demos.

Linacre JM (1994) Sample size and item calibration stability. *Rasch Measurement Transactions* 7(4): 328.

Linacre JM (1999) Investigating rating scale category utility. *Journal of Outcome Measurement* 3(2): 103–122.

Linacre JM (2002) Optimizing rating scale category effectiveness. *Journal of Applied Measurement* 3(1): 85–106.

Lord FM and Novick MR (1968) *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company.

Lundgren-Nilsson A and Tennant A (2011) Past and present issues in Rasch analysis: The functional independence measure (FIM) revisited. *Journal of Rehabilitation Medicine* 43: 884–891.

Luszczynska A, Scholz U and Schwarzer R (2005) The General Self-Efficacy Scale: Multicultural validation studies. *The Journal of Psychology* 139(5): 439–457.

Marks R (2001) Efficacy theory and its utility in arthritis rehabilitation: Review and recommendations. *Disability and Rehabilitation* 23(7): 271–280.

Park CL and Gaffey AE (2007) Relationships between psychosocial factors and health behavior change in cancer survivors: An integrative review. *Annals of Behavioral Medicine* 34(2): 115–134.

Pishghadam R, Baghaei P, Ali Shams M, et al. (2011) Construction and validation of a narrative intelligence scale with the Rasch rating scale model. *The International Journal of Educational and Psychological Assessment* 8(1): 75–90.

Raîche G (2005) Critical Eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions* 19(1): 1012.

Rasch G (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Scherbaum CA, Cohen-Charash Y and Kern MJ (2006) Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and Psychological Measurement* 66: 1047–1063.

Scholz U, Doña BG, Sud S, et al. (2002) Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment* 18(3): 242–251.

Schwarzer R (2008) Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied Psychology* 57: 1–29.

Schwarzer R and Jerusalem M (1995) Generalized self-efficacy scale. In: Weinman J, Wright S and Johnston M (eds) *Measures in Health Psychology: A User's Portfolio* (Causal and Control Beliefs). Windsor: NER-NELSON, pp. 35–37.

Schwarzer R and Jerusalem M (1999) *Skalen zur Erfassung von Lehrer- und Schülermerkmalen: Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Freie Universität Berlin.

Schwarzer R, Bäßler J, Kwiatek P, et al. (1997) The assessment of optimistic self-beliefs: Comparison of the German, Spanish, and Chinese versions of the General Self-Efficacy Scale. *Applied Psychology* 46(1): 69–88.

Smith EV Jr (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement* 3: 205–231.

Tennant A and Pallant JF (2006) Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions* 20: 1048–1051.

Tesio L (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine* 35: 105–115.

Vianin M (2008) Psychometric properties and clinical usefulness of the Oswestry Disability Index. *Journal of Chiropractic Medicine* 7: 161–163.

Wright BD (1996) Local dependency, correlations and principal components. *Rasch Measurement Transactions* 10(3): 509–511.

Wright BD and Linacre JM (1989) Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation* 70: 857–860.

Wright BD and Masters GN (1982) *Rating Scale Analysis*. Chicago, IL: MESA.

Wright BD and Tennant A (1996) Sample size again. *Rasch Measurement Transactions* 9(4): 468.