J Rehabil Med 2015; 47: 300–304

ORIGINAL REPORT

PARAMETRIC ANALYSES OF SUMMATIVE SCORES MAY LEAD TO CONFLICTING INFERENCES WHEN COMPARING GROUPS: A SIMULATION STUDY

Asaduzzaman Khan, PhD[1], Chi-Wen Chien, PhD[2] and Karl S. Bagraith, BoccThy(Hons)[1,3]

From the [1]School of Health and Rehabilitation Sciences, The University of Queensland, Brisbane, Australia, [2]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (SAR), China and [3]Interdisciplinary Persistent Pain Centre, Gold Coast Hospital and Health Service, Gold Coast, Australia

Objective: **To investigate whether using a parametric statistic in comparing groups leads to different conclusions when using summative scores from rating scales compared with using their corresponding Rasch-based measures.**

Methods: **A Monte Carlo simulation study was designed to examine between-group differences in the change scores derived from summative scores from rating scales, and those derived from their corresponding Rasch-based measures, using 1-way analysis of variance. The degree of inconsistency between the 2 scoring approaches (i.e. summative and Rasch-based) was examined, using varying sample sizes, scale difficulties and person ability conditions.**

Results: **This simulation study revealed scaling artefacts that could arise from using summative scores rather than Rasch-based measures for determining the changes between groups. The group differences in the change scores were statistically significant for summative scores under all test conditions and sample size scenarios. However, none of the group differences in the change scores were significant when using the corresponding Rasch-based measures.**

Conclusion: **This study raises questions about the validity of the inference on group differences of summative score changes in parametric analyses. Moreover, it provides a rationale for the use of Rasch-based measures, which can allow valid parametric analyses of rating scale data.**

Key words: rating scales; parametric statistics; Rasch analysis; simulation; summative scores; Rasch-based measures.

J Rehabil Med 2015; 47: 300–304

Correspondence address: Asad Khan, School of Health and Rehabilitation Sciences, The University of Queensland, Brisbane QLD 4072, Australia. E-mail: a.khan2@uq.edu.au

Accepted Nov 5, 2014; Epub ahead of print Feb 13, 2015

INTRODUCTION

Clinicians and researchers in health and rehabilitation sciences often need to assess patient-reported outcomes that may not be directly measurable; for example, disability, cognitive function, quality of life, satisfaction, or pain intensity. Such patient-reported outcomes are often measured indirectly via their manifestations, and are referred to as "latent variables", "latent traits" or "constructs" (1, 2). Ordinal rating scales are commonly employed to quantify latent traits or constructs. In ordinal rating scales, respondents choose from a series of ordered response options, which generally denote "less" or "more" of a construct (e.g. in measuring pain, none = 0, mild = 1, moderate = 2 and severe = 3). However, responses on such scales simply represent an ordinal graduation, rather than a "real number" or "physical magnitude" with inherent meaning (measurement scales are described in (3)). In addition, a change or difference of 1 point on ordinal scales can vary in meaning across the scale's continuum (4). Thus, ordinal rating scales do not typically produce interval-level data, and the mathematical manipulations (e.g. addition, substraction, or mean) that are routinely applied to ordinal scores in order to assess patient outcomes may not have readily interpretably meaning within or between patients (5, 6).

Rating scales are widely used to assess patient-reported outcomes that are complex in nature. These items can be pooled to generate a single summary score, which is meant to measure the phenomenon quantified by the set of items. Sometimes the items are weighted or standardized before pooling; however, the most common strategy to generate summary scores is via direct summation of the relevant individual item scores (7). Unfortunately, the total summed scores from rating scales simply offer rank-ordered values, as the component items are based on ordinal rating scales. Summing the scores from multiple items to create a total score is based on the assumption that all the items are measured on the same interval scale and each item contributes equally to the final score, irrespective of how well each item contributes to the underlying construct (8). Such summative scores often ignore the reality that some items may be more important in measuring a construct than others, especially when the difficulty levels of items are likely to be varied (9). Thus, the total summed scores obtained from rating scales violate the fundamental assumption of an equidistant interval (i.e. constant unit of measurement), which is required for the application of any parametric statistics (10).

Ordinal data generated from rating scales can be analysed using parametric statistics, which are more powerful than non-parametric statistics (11). However, when parametric statistics are applied to ordinal data the inferences about the underlying construct may not be logically valid and the resulting conclusions may therefore be misleading (12, 13). Previous research

 doi: 10.2340/16501977-1941
 ISSN 1650-1977

has demonstrated how the use of parametric statistics (e.g. analysis of variance (ANOVA)) with ordinal scores to compare groups can lead to incorrect inferences about the underlying construct level (14, 15) and spurious interaction effects between groups and conditions (16). Using a Monte Carlo simulation, another study also showed that untransformed summative ordinal scores can produce underestimated main effects and spurious interaction effects in the context of 2-way ANOVA, and may lead to misleading conclusions (17). A more recent study presented evidence that erroneous conclusions can occur when parametric statistics are applied inappropriately to analyse ordinal outcome measures from health-related quality of life measures (18).

In recent years Rasch models have been increasingly applied to operationalize rating scales due to their unique advantages over classical test theory (CTT) approaches (19). The Rasch model provides a means for transforming ordinal rating scale data into interval measures if and only if the data fit the model assumptions (20, 21). Under the Rasch algorithm, measurement errors can be estimated more accurately and the resulting scores hold the properties of sample independence, invariance and additivity (i.e. equality of the measurement at different points on the continuum). Previous research has suggested that scoring based on Rasch methods offers greater accuracy, precision and responsiveness than classical summative scoring when measuring patient outcomes based on ordinal rating scales (22–24). In addition, item characteristics from Rasch analyses are not necessarily sample dependant, as they are in classical test theory analyses, due to the separate estimation of item difficulty and person ability in Rasch analysis when using conditional estimation (22). Despite these advantages, little is known about the inferential benefits of applying Rasch modelling to analyse ordinal rating scale data in clinical practice and research.

In this paper, a Monte Carlo simulation study was designed to examine whether summative ordinal and Rasch-based interval approaches to scoring rating scales could lead to different statistical inferences when parametric statistics were employed to compare change scores between groups. More specifically, we sought to explore whether the inferential differences (if any) between the 2 scoring approaches were associated with the difficulty levels of test items, the ability of the respondents, or the size of the sample.

METHODS

A Monte Carlo simulation study was designed to examine, when parametric statistics were operated, the impact of using the classical summative scores on inferences related to group comparisons of changes and whether such inference was comparable with that of using Rasch-based interval measures. The rationale for using a Monte Carlo simulation methodology was its flexible manipulation, whereby the rating scale responses could be generated under different simulation scenarios with distinct and known distributions of item difficulty and person ability with different sample sizes, which would otherwise be impossible to achieve in a clinical dataset (25). In this study, 3 populations with varying latent abilities (e.g. low, medium and high) were artificially generated, using the item response theory generation software WinGen (26). In particular, in each ability-level population, we generated one treatment group and one control group with a distinct mean difference, in order to present 3 different magnitudes of treatment effects. Polytomous ordinal responses with varying difficulty and ability parameters were generated using the Rasch partial credit model (PCM):

where θ is the ability for person on the construct and b is the item difficulty parameters for each score category with where is the difficulty of item i, and is the relative difficulty of score category x of item i.

For this simulation study, 3 tests with different item difficulty levels (e.g. easy, moderate and hard) were specified and each test included 10 items (as used elsewhere (27)). Response options for the items were fixed at 5 levels (0–1–2–3–4) to reflect scales commonly used in health and behavioural assessments (27). The items in the 3 tests were sampled from normal distributions (as used elsewhere (28–30)), with standard deviation of 1 and means of –1.0, 0.0, and 1.0, respectively, for easy, moderate and hard tests.

For each population, pseudo subjects were generated for 2 groups (i.e. treatment and control group) from normal populations with standard deviation (SD) of 1.0, but varying means. To ensure equal latent ability change (k) between the 2 groups (treatment and control) within each population, the treatment group mean was specified as the control group mean + k. A relatively small ability change (e.g. $k = 0.3$ for this analysis) was assumed to demonstrate how the 2 scoring approaches act in identifying small group differences under different simulation scenarios. Ability distributions of control groups in each population were generated from normal distributions with SD of 1 and means of –0.65, –0.15 and 0.35, respectively (Table I). Such distributions for treatment groups were generated from normal distributions with SD of 1 and means of –0.35, 0.15 and 0.65, respectively. Three sample conditions ($n = 250$, 500 and 1000), as used in other simulation studies (27, 31), were considered to generate subjects from normal distribution for the present study, with a set of 20 replications for each scenario.

Generated item responses for each test were summed in order to derive the raw total scores for each subject, referred to as classical summative scores (range 0–40). To derive the corresponding interval scale scores, Rasch PCM analysis was used. Specifically, we pooled the item responses from the control and treatment groups (at the same test difficulty level) together in each Rasch analysis (as described elsewhere (23, 32)). Thus the Rasch-produced ability estimates of each respondent across the control and treatment groups were anchored on the same measurement continuum of item difficulty parameters. These Rasch-based ability estimates were expressed in log-odd units or *logits*, which can be viewed as interval scores (4). As the data for this simulation study were generated using a unidimensional PCM model, no attempt was made to examine the Rasch model assumptions (e.g. unidimensionality, fit statistics). Details about the examination of

Table I. *Three simulated populations for control and treatment groups with different ability and difficulty parameters*

Test difficulty b_x~N(mean, 1)	Person ability	Control group θ~N(mean, 1)	Treatment group θ~N(mean[a]+0.3, 1)
N(–1,1) Easy	Low	θ~N(–0.65,1)	θ~N(–0.35, 1)
	Medium	θ~N(–0.15,1)	θ~N(0.15, 1)
	High	θ~N(0.35,1)	θ~N(0.65, 1)
N(0,1) Moderate	Low	θ~N(–0.65,1)	θ~N(–0.35, 1)
	Medium	θ~N(–0.15,1)	θ~N(0.15, 1)
	High	θ~N(0.35,1)	θ~N(0.65, 1)
N(1,1) Hard	Low	θ~N(–0.65,1)	θ~N(–0.35, 1)
	Medium	θ~N(–0.15,1)	θ~N(0.15, 1)
	High	θ~N(0.35,1)	θ~N(0.65, 1)

b_x: item difficulty for each score category and θ is the ability parameter, both of which are normally distributed with varying mean and a standard deviation of 1.

[a]Control group mean.

Rasch model assumptions and derivation of logit scores (i.e. measures) from rating scales are given elsewhere (33, 34). Winsteps software V 3.68.2 (35) was used to perform the Rasch PCM analyses using joint maximum likelihood estimation.

Observed treatment effects were calculated by subtracting the mean score of the control group from that of the treatment group for each of the 3 populations (e.g. low, medium and high ability) under 3 different test conditions (e.g. easy, moderate and hard). A 1-way ANOVA was conducted to examine whether the treatment effects varied across the 3 populations under 3 different test conditions. *Post-hoc* pairwise comparisons with Bonferroni's adjustment were further attempted in order to identify the pairs (if any) that were significantly different under the 2 scoring approaches. These analyses were independently implemented on the classical summation scores and the Rasch-based interval measures in order to examine whether the conclusions based on the 2 scoring approaches were different under the 3 samples with varying sizes: 250, 500 and 1,000.

RESULTS

Figs 1a, b and c present the mean differences in change scores along with their error bars between the treatment and the control groups for each of the 3 populations (e.g. low, medium and high ability) under 3 different test conditions with sample sizes of 250, 500 and 1,000. While a constant change of 0.3 in ability between the treatment and control groups was set in the generation of their item responses, the observed changes in summative score were found to be different for various populations depending on the test difficulty levels. For summative scores (as shown in Fig. 1), *easy* tests with *low* ability consistently produced the highest differences in change scores compared with the *easy* tests applied to the groups with *medium* or *high* ability. By contrast, the highest mean difference

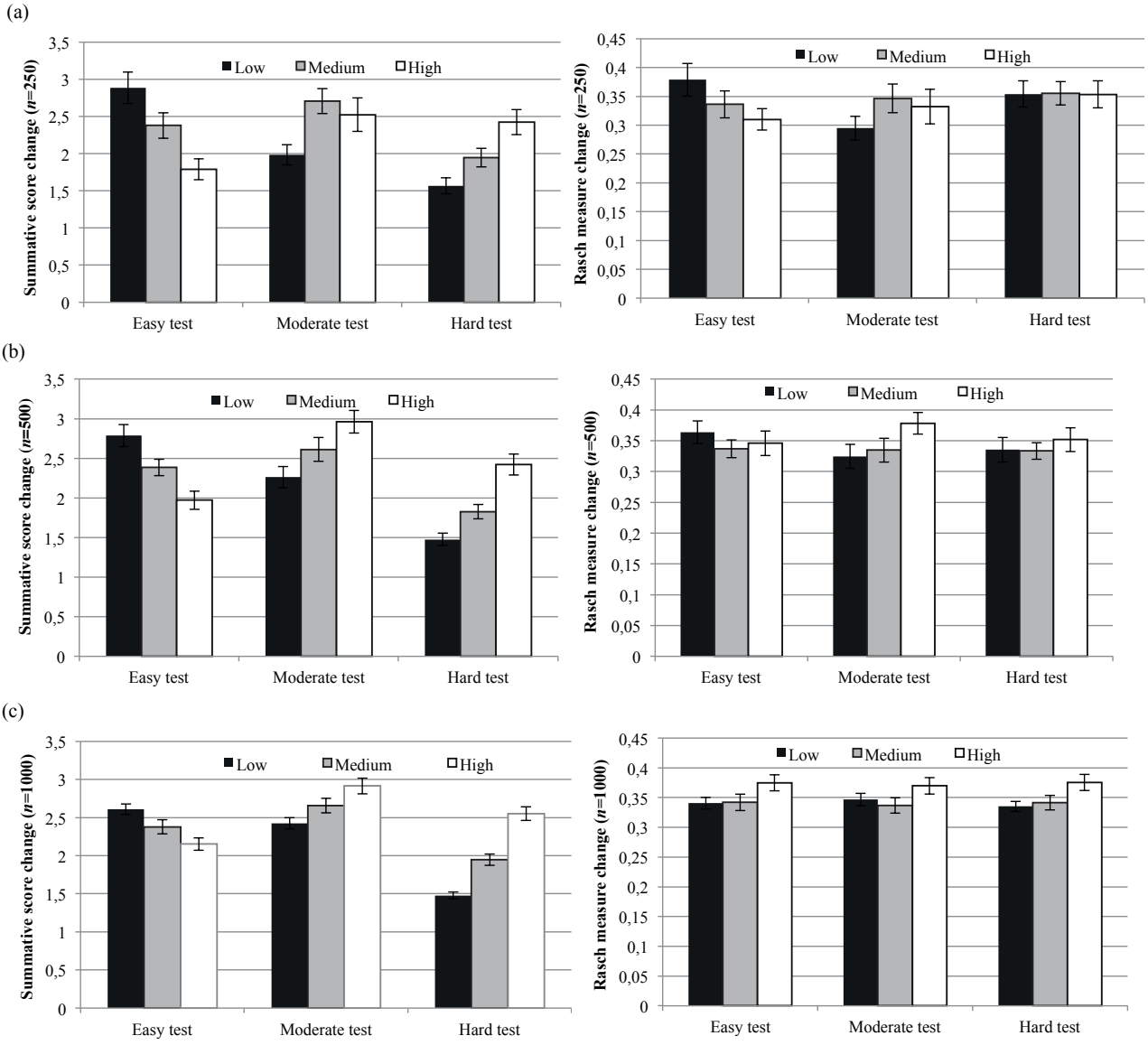

Fig. 1. Mean differences in change scores between treatment and control groups in classical summative scores and Rasch-based measures under different test conditions for a sample of (a) 250, (b) 500 and (c) 1,000.

in summative change scores for *hard* tests was found in the population with *high* ability level. This pattern was consistent across the 3 sample size scenarios. However, it is noted that the *moderate* tests applied to the population with *medium* ability had the highest mean difference only in the sample size of 250. It is thus implied that the magnitude of the mean differences in change scores of summative scores within each population might be dependent on the interactive relationship between test difficulty, respondent ability levels, or sample size.

For Rasch-based measures (as shown in Fig. 1), no consistent pattern regarding the difference in change scores was observed across the 3 difficulty tests applied to groups with 3 ability levels in 3 sample size conditions. However, it was found that the changes in Rasch-based measures marginally exceeded the simulated change value of 0.3 in most of the scenarios.

One-way ANOVA revealed significant mean differences in change scores across the 3 population groups with low, medium and high ability when summative scores were considered for easy ($p \leq 0.001$), moderate ($p < 0.05$) and hard ($p < 0.001$) tests for each sample size scenario. On the other hand, no statistical significance was identified when the same analyses were implemented on Rasch-based measures ($p > 0.05$) (Table II). *Post-hoc* pair-wise analyses on summative scores identified significant differences between population groups with *low* vs *high* ability ($p < 0.01$) in all sample size scenarios with the exception of the moderate test with a small sample size of 250.

Table II. *Analysis of variance (ANOVA) results for comparing the mean differences in change scores between population groups under different test conditions for sample sizes of 250, 500 and 1,000*

		p-value			
			Pair-wise comparisons between abilities[b]		
Test	Scoring method	Overall[a]	Low vs medium	Medium vs high	High vs low
n = 250					
Easy	Summative	<0.001	0.14	0.07	<0.001
	Rasch-based	0.13	0.64	0.99	0.14
Moderate	Summative	0.02	0.02	0.99	0.12
	Rasch-based	0.34	0.47	0.99	0.91
Hard	Summative	<0.001	0.17	0.05	<0.001
	Rasch-based	0.99	0.99	0.99	0.99
n = 500					
Easy	Summative	<0.001	0.07	0.06	<0.001
	Rasch-based	0.56	0.88	0.99	0.99
Moderate	Summative	0.004	0.27	0.26	0.003
	Rasch-based	0.11	0.99	0.32	0.14
Hard	Summative	<0.001	0.06	<0.001	<0.001
	Rasch-based	0.73	0.99	0.99	0.99
n = 1,000					
Easy	Summative	0.001	0.16	0.18	0.001
	Rasch-based	0.10	0.99	0.20	0.17
Moderate	Summative	0.002	0.24	0.16	0.001
	Rasch-based	0.17	0.99	0.19	0.60
Hard	Summative	<0.001	<0.001	<0.001	<0.001
	Rasch-based	0.07	0.99	0.21	0.09

[a]*p*-value based on F-statistic; [b]*p*-value based on *t*-statistic.

DISCUSSION

This study examined the extent to which the application of parametric statistics to classical summative scores generated from rating scales could produce conflicting inferences compared with Rasch-based measures. The results revealed that scaling artefacts could arise when using classical summative scores instead of Rasch-based interval measures. More specifically, the greatest difference in the summative scores was observed if the test with different difficulty levels was used in the participants with the corresponding ability levels (e.g. the *easy* test for the population with *low* ability or the *hard* test for the population with *high* ability). However, given that the data generated were pre-determined to have equal differences in latent abilities between the groups, it is evident that such observed differences could have resulted from the scaling artefacts; and a similar finding has been previously reported elsewhere (29). Earlier research also suggests that total summative scores are not linearly related with the underlying abilities, and that such non-linearity is influenced by the difficulty levels of test items (36). The author argued that non-linearity could be largest for abilities that are extreme with respect to the test items, presenting a plausible reason for differential changes in summative scores across various difficulty levels of test items.

The present study found that comparisons of different groups with different test difficulty levels produced conflicting inferences when using classical summative scores as opposed to Rasch-based measures. These findings are consistent with the findings of previous research, which has demonstrated that group comparisons using t-tests on an observed variable can be influenced by the difficulty of tests (14) and that inconsistent inferences about a latent variable can result from factorial ANOVA (15). The conflicting findings of the present study are therefore of interest, especially because the Rasch model, once fitted properly, can generate interval-level estimates that are suitable for parametric analysis and are more precise than summative scores (24).

In contrast to the findings of the present study, a recent study demonstrated that the use of summative scoring appeared robust to violations in the underlying assumptions of parametric analyses and is comparable to the item response theory-based scores when evaluating the relationships between test scores and outcome measures (27). In addition, there is evidence to suggest that although each item in a rating scale reflects an ordinal number, aggregation of such items into a single score to measure a construct may yield, or closely approximate, an interval scaled score (37), and thus facilitates the use of parametric statistics when analysing summative rating scale scores.

While Rasch-based measures have theoretical and potentially clinical advantages over summative scores, deriving Rasch measures requires choosing and applying the appropriate Rasch model with its underlying assumptions. This complicates the operationalization of Rasch modelling to transform summative scores into interval level measurements and may deter researchers and clinicians from using a Rasch-based scoring approach. More research is therefore warranted to either sim-

plify the Rasch transformation process (e.g. by generating a transformation table) or delineate the conditions under which summative scores can be considered as interval measures and hence be appropriately subjected to parametric analyses.

In conclusion, the present study demonstrated that spurious statistical inferences are likely when analysing rating scale data with different item difficulty levels using classical summative scores, and that this has the potential to produce conflicting conclusions when comparing populations with respect to the underlying construct. Given that Rasch-based measures are generally considered an improvement over classical summative scores, this simulation study provides additional evidence for the use of Rasch-based scoring of rating scale data in order to make valid and accurate inferences through applying appropriate statistical analyses. These findings also support the argument of a recent editorial for not using raw scores from ordinal scales in rehabilitation sciences (38). Nevertheless, further real-world evidence is needed to support the findings of this simulation study and to motivate healthcare professionals and researchers to use a Rasch-based scoring approach when rating scales are used to assess patient-reported outcomes.

ACKNOWLEDGEMENTS

This study was partly supported by a research grant from The University of Queensland, Australia. The authors would like to thank Dr Nicola W. Burton for her comments in an earlier version of this paper and Mr Onwell Maconi for helping with the simulation study.

REFERENCES

1. Martinez-Martin P. Composite rating scales. J Neurol Sci 2010; 289: 7–11.
2. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
3. Hays W. Chapter 2.1 Measurement scales in statistics. Fort Worth, TX: Harcourt Brace College Publishers; 1994, p. 71–77.
4. Wright B, Masters N. Rating scale analysis. Chicago: MESA Press, 1982.
5. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001; 33: 47–48.
6. Kucukdeveci A, Tennant A, Grimby G, Franchignoni F. Strategies on assessment and outcome measurement in physical and rehabilitation medicine: an educational review. J Rehabil Med 2011; 43: 661–672.
7. DeVellis R. Classical test theory. Med Care 2006; 44: S50–S59.
8. Streiner D, Norman G. Health Measurement scales: a practical guide to their development and use. 1st ed. New York: Oxford University Press; 2008.
9. Tesio L, Simone A, Bernardinello M. Rehabilitation and outcome measurement: where is Rasch analysis-going? Eura Medicophys 2007; 43: 417–426.
10. Hobart J, Cano S, Zajicek J, Thompson A. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007; 6: 1094–1105.
11. Drisko J, Grady M. Evidence-based practice in clinical social work. New York: Springer-Verlag New York Inc.; 2012
12. Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989; 70: 308–312.
13. Forrest M, Andersen B. Ordinal scale and statistics in medical research. Br Med J 1986; 292: 537–538.
14. Maxwell S, Delaney H. Measurement and statistics: an examination of construct validity. Psychol Bull 1985; 97: 85–93.
15. Davison M, Sharma A. Parametric statistics and levels of measurement:factorial designs and multiple regression. Psychol Bull 1990; 107: 394–400.
16. Embretson S. Item response theory models and spurious interaction effects in factorial ANOVA designs. Appl Psychol Meas 1996; 20: 201–212.
17. Romanoski J,Douglas G. Rasch-transformed raw scores and two-way ANOVA: a simulation analysis. J Appl Meas 2002; 3: 421–430.
18. Kahler E, Rogausch A, Brunner E, Himmel W. A parametric analysis of ordinal quality-of-life data can lead to erroneous results. J Clin Epidemiol 2008; 61: 475–480.
19. Belvedere S, de Morton N. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J Clin Epidemiol 2010; 63: 1287–1297.
20. Embretson S, Reise S. Item response theory for psychologists. 1st ed. NJ: Lawrence Erlbaum; 2000.
21. Hobart J. Measuring disease impact in disabling neurological conditions: are patients' perspectives and scientific rigor compatible?. Curr Opin Neurol 2002; 15: 721–724.
22. Hays R, Morales L, Reise S. Item response theory and health outcomes measurement in the 21'st century. Med Care 2000; 38: S28–S42.
23. Lyren P, Atroshib I. Using item response theory improved responsiveness of patient-reported outcomes measures in carpal tunnel syndrome. J Clin Epidemliol 2012; 65: 325–334.
24. Khan A, Chien CW, Brauer S. Rasch-based scoring offered more precision in differentiating patient groups in measuring upper limb function. J Clin Epidemiol 2013; 66: 681–687.
25. Wood M. The role of simulation approaches in statistics. J Stat Educ 2005; No. 3. Available from: www.amstat.org/publications/jse/v13n3/wood.html.
26. Han K. WinGen: Windows software that generates IRT parameters and item responses. Appl Psychol Meas 2007; 31: 457–459.
27. Xu T, Stone CA. Using IRT Trait estimates versus summated scores in predicting outcomes. Educ Psychol Meas 2012; 72: 453–468.
28. Dawber T, Rogers W, Carbonaro M. Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models. Alberta J Edu Res 2009; 55: 512–533.
29. Embretson S. Comparing changes between groups: some perplexities arising from psychometrics. In: Laveault D, Zumbo B, Gessaroli M, editors. Modern theories of measurement: problems and issues. Ottawa, ON: University of Ottawa; 1994, p. 214–248.
30. Ree M. Estimating item characteristics curves. Appl Psychol Meas 1979; 3: 371–385.
31. Stone C. Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: an evaluation of MULTILOG. Appl Psychol Meas 1992; 16: 1–16.
32. Chien CW, Brown T, McDonald R. Cross-cultural validity of a naturalistic observational assessment of children's hand skills: a study using Rasch analysis. J Rehabil Med 2011; 43: 631–637.
33. Lannsjo M, Borg J, Bjorklund G, Geijerstam J, Lundgren-Nilsson A. Internal construct validity of the Rivermead Post-Concussion Symptoms Questionnaire. J Rehabil Med 2011; 43: 997–1002.
34. Tennant A, Conaghan P. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. Arthritis Rheum 2007; 57: 1358–1362.
35. Linacre J. A User's guide to Winsteps: Rasch-model computer program. Chicago: MESA Press; 2008.
36. Lord F. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum; 1980.
37. Carifio J, Perla R. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. J Biosoc Sc 2007; 3: 106–116.
38. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med 2012; 44: 97–98.