Open Journal of Statistics, 2014, 4, 456-465
Published Online August 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.46044

Scientific Research

Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/*t*-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics

Peter Hagell

The PRO-CARE Group, School of Health and Society, Kristianstad University, Kristianstad, Sweden
Email: Peter.Hagell@hkr.se

Received 26 May 2014; revised 30 June 2014; accepted 15 July 2014

Abstract

Psychometric theory requires unidimensionality (*i.e.*, scale items should represent a common latent variable). One advocated approach to test unidimensionality within the Rasch model is to identify two item sets from a Principal Component Analysis (PCA) of residuals, estimate separate person measures based on the two item sets, compare the two estimates on a person-by-person basis using *t*-tests and determine the number of cases that differ significantly at the 0.05-level; if ≤5% of tests are significant, or the lower bound of a binomial 95% confidence interval (CI) of the observed proportion overlaps 5%, then it is suggested that strict unidimensionality can be inferred; otherwise the scale is multidimensional. Given its proposed significance and potential implications, this procedure needs detailed scrutiny. This paper explores the impact of sample size and method of estimating the 95% binomial CI upon conclusions according to recommended conventions. Normal approximation, "exact", Wilson, Agresti-Coull, and Jeffreys binomial CIs were calculated for observed proportions of 0.06, 0.08 and 0.10 and sample sizes from *n* = 100 to *n* = 2500. Lower 95%CI boundaries were inspected regarding coverage of the 5% threshold. Results showed that all binomial 95%CIs included as well as excluded 5% as an effect of sample size for all three investigated proportions, except for the Wilson, Agresti-Coull, and JeffreysCIs, which did not include 5% for any sample size with a 10% observed proportion. The normal approximation CI was most sensitive to sample size. These data illustrate that the PCA/*t*-test protocol should be used and interpreted as any hypothesis testing procedure and is dependent on sample size as well as binomial CI estimation procedure. The PCA/*t*-test protocol should not be viewed as a "definite" test of unidimensionality and does not replace an integrated quantitative/qualitative interpretation based on an explicit variable definition in view of the perspective, context and purpose of measurement.

How to cite this paper: Hagell, P. (2014) Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/*t*-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics. *Open Journal of Statistics*, **4**, 456-465.
http://dx.doi.org/10.4236/ojs.2014.46044

Keywords

Confidence Intervals, Dimensionality, Psychometrics, Rasch Model, Validity

1. Introduction

Rating scales are one of the most commonly used methods of data collection across a range of disciplines such as behavioural, educational, social and health sciences. Rating scales are rooted in the behavioural sciences and their typical purpose is to enable measurement of phenomena that cannot be directly observed and measured, *i.e.*, latent variables. The measurement of such variables is of central and immense importance. For example, within the clinical health sciences rating scales are a prime mode of data collection in descriptive and associative studies as well as in clinical trials of therapeutic interventions. As such, rating scale based data have direct impact on study results, conclusions, and ultimately individual patient care. The quality of rating scales is therefore at the heart of the quality of evidence-based practice and central to the quality of study results and decision-making [1].

The basic logic of rating scale based measurement is that a number of observable manifestations of the target variable are selected and expressed as items in a rating scale, typically a self-report or observer based assessment tool. These items are assumed to represent expressions or manifestations of the latent variable that is intended to be measured [2]-[4]. The extent to which this process is successful is evaluated by means of tests regarding the rating scale's psychometric properties, of which aspects of reliability and validity are the core characteristics. Whereas reliability concerns measurement precision and consistency, validity refers to the extent to which the rating scale (or any other measurement tool) actually measures the variable it purports to measure. In this respect, dimensionality is a central consideration and the topic of the current paper.

Psychometric measurement models require unidimensionality; *i.e.* valid and legitimate summing of rating scale items into an interpretable total score rest on the requirement that the items represent one common underlying (latent) variable [5]. This idea, that useful measurement only represents one attribute at a time is not specific for rating scales and latent variables, and dates back at least to the early 1930-ies, when Thurstone [6] stated that:

"The measurement of any object or entity describes only one attribute of the object measured. This is a universal characteristic of all measurement" (p. 259).

When total scores are not unidimensional they are technically invalid and their meaning is ambiguous since it is unclear what scores represent. There are at least three related reasons why unidimensionality is important to consider [5] [7] [8]. Firstly, unidimensionality is a basic assumption for valid calculation of total scores according to both classic and modern test theories. Secondly, unambiguous interpretation requires scores to represent a single defined attribute. That is, scores on a scale that is used to measure one variable should not be appreciably influenced by varying levels of one or more other variables. Thirdly, if scores do not represent a common line of inquiry it is unclear if two individuals with the same score can be considered comparable. Similarly, the interpretation of any differences between individuals will be ambiguous since it is unknown in what way(s) they actually differ. This cannot be compensated for by study design or analytical statistics, and hampers the understanding and usefulness of outcomes. For example, if this is not clear in therapeutic clinical trials, it may have consequences for the selection of interventions for individual patients; unambiguous score interpretation is a prerequisite for rating scales to be acceptable as clinical trial endpoints [9].

Andrich [10] emphasized three related points regarding unidimensionality:

"First, unidimensionality is a relative matter—every human performance, action, or belief is complex and involves a multitude of component abilities, interests, and so on. Nevertheless, there are circumstances in which it is considered useful to think of concepts in unidimensional terms (…). Second, a unidimensional variable is constructed—it makes a great deal of ingenuity and knowledge of the subject matter to establish a variable that is unidimensional to a level of precision that is of some practical or theoretical use (…): Where relevant, successful measurement demonstrates a great deal of understanding of the property. Often, devising a measuring instrument is as important in what it teaches about the variable as are the subsequent acts of measurement using the instrument. Third, with unidimensional measurements, comparisons can be made using their differences.

Such differences are differences in degree. Differences that are not differences in degree are said to be differences in kind, and both are important" (pp. 9-10).

Despite the strong case for unidimensionality, it must be emphasized that this property is not an absolute but a relative one. That is, it is relative with respect to variable definition, perspective and purpose of measurement and score interpretation, and with respect to frames of reference. In certain situations, it may be useful and desirable to consider broad constructs and variable definitions, whereas more focused and narrow ones are more appropriate in other situations. Andrich [11] provided an excellent metaphor to this end, using an example from educational measurement of student achievement in mathematics, which may be subdivided into areas such as addition, subtraction, multiplication, division, and so on:

"If one considers a very thick rope, which can of course be straightened to form a linear continuum, there are components that are made of much finer threads (e.g., *items in a rating scale*). These are woven together to form a higher-level component, which could itself be a narrow (thin) rope (e.g., *dimension or subscale of a rating scale*). These relatively thin ropes are then woven together to form a thicker rope (e.g., *an overall rating scale score*), this process can be repeated until one has a rope thick enough for the purpose in hand" (p. 104) (emphases in italics added).

This metaphor is easily applicable to a variety of contexts. Consider for example a rating scale intended to measure patient-reported overall health. Depending on how "overall health" was defined in guiding the development of that scale, it is reasonable to suggest that such a rating scale would comprise items addressing aspects of physical, mental, social health and so on. Items would then be the finer threads in the metaphor above, the physical, mental and social health dimensions would be the thinner ropes, and the overall scale would be the thicker rope. As long as the content provides a valid representation of manifestations of the latent "overall health" variable, the overall scale may meet the unidimensionality requirement and render interpretable scores within that frame of reference. However, if the focus of interest is mental health, an overall total score would be irrelevant. Instead, items representing the mental health dimension would need to be unidimensional to yield valid and interpretable mental health scores from that perspective and within that frame of reference. This mental health dimension may or may not then in turn be subdivided into further subdimensions representing, e.g., mood, cognition, and so on. Clearly, the issue is one related to context and perspective.

Although there are multidimensional measurement models, these are in general a means of accommodating multidimensionality in the analysis; scores produced within the respective dimensions as well as overall summary scores are still required to be unidimensional within their respective frames of reference and according to their respective definitions. This emphasizes the central role of variable definition, frame of reference and perspective in combination with judgement [10]-[13]; unidimensionality is not an "either/or" issue but a matter of degree. This view was expressed already over 70 years ago [14] and has been repeatedly reiterated since [8] [11] [15]-[17]. Presumably due to the central role of dimensionality in measurement in general and in rating scale construction and evaluation specifically, a vast number of quantitative indices have been suggested as tests of the rating scale unidimensionality requirement [8] [15] [17] [18]. Traditionally, these tend to be based on reliability indices, principal component or factor analysis, and indices of fit between the data and the measurement model; the latter most prominently within modern psychometric theory such as the Rasch model (RM).

2. The Rasch Model

The RM [19] mathematically defines what is required from item responses in order for them to express linear measures rather than mere numbers or ordinal scores. It separately locates persons and items on a common logit (log-odd units) metric that is centered by the mean item location, which is set at zero. According to the RM the probability of a certain item response is a logistic function of the difference between the level of the measured construct represented by the item and that possessed by the person, and only a function of that difference. Expressed formally, this gives:

$$P_{ni} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}, \tag{1}$$

where P_{ni} is the probability of person n to endorse item i, e (Euler's number) is the exponential of the natural logarithm, β_n is the level of the construct possessed by the person and δ_i is the level of the construct represented by the item.

For each person (and item) the model thus estimates a location on the latent variable continuum from less to more, as expressed according to the logit metric. For dichotomous items (*i.e.* items with binary response categories such as "yes"/"no" or "agree"/"disagree"), the location equals the relative position on the latent trait (expressed as a logit value) where there is a 50/50 probability of responding with either of the two response categories; the location of polytomous items (*i.e.* items with more than two ordered response categories) is the mean location of the response category thresholds (*i.e.*, positions on the latent trait where there is a 50/50 probability of responding with either of two adjacent response categories; the number of thresholds is one less than the number of response categories). In addition, and in contrast to traditional raw score based test theory, the RM also provides individual person (and item) standard errors based on the amount of information available regarding that person's (item's) location. That is, more items provide more information about the person's location and, hence, greater precision and a smaller standard error (SE). The same logic also applies for items, *i.e.*, larger samples provide greater precision and smaller SEs for item locations. Specifically, precision is optimized when item threshold locations match the person location (and vice verse). When rating scale data (item responses) meet the requirements of the RM, invariant linear measurement is achieved [10] [16]. However, this achievement is conditional on substantively explicit variable definitions according to which items represent manifestations of the latent target variable, where the RM derived item hierarchy is consistent with theory and reflects what is happening as one moves up and down the scale. In contrast, if rating scale items define the variable rather than the other way around, the scale represents an index rather than a measure [2] [4], and the function of the RM is descriptive rather than anything else [20].

The RM is related to the Guttman model [21] [22] but with the important difference that whereas the model proposed by Guttman is deterministic in nature (*i.e.*, individual item response patterns are directly determined by the person's total score), the RM is probabilistic. The basic assumption and logic of the RM was succinctly articulated as follows by Rasch [19]:

"A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one" (p. 117).

Note that Rasch pointed out that the model is not necessarily concerned with particular sets of items, but rather with "any item of the type in question". That is, as long as items represent the same variable ("item of the type in question"), they should be useful and able to yield the same person measure. The RM is therefore useful in calibrating and building item banks, *i.e.* large sets of items that represent the same variable and from which smaller numbers of items can be selected for the measurement of individual persons, although the exact items used by two people are not necessarily the same [23].

The extent to which data accord with the RM is assessed through analyses of fit, *i.e.* by examining the accordance between expected and observed responses across locations on the measured variable [10] [24]. This is typically determined through a combination of approaches including inspection and analyses of residuals using chi-square or ANOVA based statistics as well as graphical methods. Residuals represent the discrepancy between observed and expected item responses. In general, large positive residuals primarily suggest violation of unidimensionality, whereas large negative residuals signal local dependency (*i.e.*, item responses are dependent on responses to other items, suggesting item redundancy). However, fit statistics can be somewhat insensitive in detecting multidimensionality, particularly if two dimensions are represented by about the same number of items [7] [25] [26]. Approaches beyond that incorporated in the traditional study of model fit have therefore been proposed in order to assess the dimensionality of rating scales within the RM framework. These include, for example, principal component analysis (PCA) of residuals (*i.e.*, what is "left over" once the RM has accounted for the main dimension), likelihood ratio tests, tests based on the association between observed and expected measures, estimation of theoretical correlations, and testing the effect of subtest construction on reliability estimates [7] [8] [17] [18] [27].

3. The Principal Component Analysis (PCA)/*t*-Test Protocol for Testing Unidimensionality in the Rasch Model

One approach that has been advocated in testing for unidimensionality within the RM framework is a PCA based method first proposed by Smith [7]. This approach attempts to assess whether scales are sufficiently unidimensional to be treated as such in practice [7] [26]. First, two item sets potentially representing different sub-

dimensions in the data are identified from a PCA of residuals according to item residual loadings on the first principal component. Specifically, item residuals with loadings of +0.3 or more and −0.3 or less are taken as potential representatives of subdimensions [26]. The two thus identified item sets are then used to estimate two separate sets of person measures. A series of *t*-tests is then conducted to compare the two estimates on a person-by-person basis in order to determine the proportion of instances in which the two item sets yield different person measures. This is possible because the RM yields individual person SEs. If violation of the unidimensionality requirement is trivial, the proportion of person locations that differ between the two item sets is small. Specifically, if ≤5% of the *t*-tests are significant, or the lower bound of a binomial 95% confidence interval (CI) of the observed proportion overlaps 5%, then it has been suggested that unidimensionality can be inferred, otherwise the scale is multidimensional [7] [26] [28] [29]. This approach is based on the same heuristic that underpins any hypothesis testing, *i.e*. that up to 5% is what would be expected to occur by chance under the null hypothesis [30]. Simulation studies have suggested that the PCA/*t*-test protocol performs well as a unidimensionality test in comparison to traditional RM fit analysis, raw score PCA and factor analyses as well as to PCA of Rasch residuals alone, particularly when based on at least 12 item response thresholds [26] [29]. Assumedly due to this, in combination with the implementation of the procedure in popular RM software [31] and its propagation as a test of strict unidimensionality [28], this test has become increasingly popular in RM based rating scale evaluations and it is often implied to provide "definite" evidence for or against the unidimensionality of rating scales (see, e.g. [32]-[37]).

However, despite its seemingly intuitive logic the approach is not without problems. Given its proposed significance and potential implications of its acceptance, this procedure is therefore in need of detailed scrutiny. One apparent critical issue is the basis for the decision as to whether a rating scale meets the unidimensionality requirement or not, *i.e*. the binomial 95% CI. First, there are a number of procedures available for the estimation of the 95% binomial CI [38] [39] and secondly, sample size impacts the width of CIs and therefore also the resulting conclusions [40] [41]. However, neither of these aspects has been considered in the propagation or application of the PCA/*t*-test protocol for testing unidimensionality in the RM. For example, none of the methodological papers that propagate the procedure has commented on the influence of sample size or the type of binomial CI [26] [28] [29]. This paper therefore explores the impact of sample size and estimation method for the 95% binomial CIs and the resulting conclusions according to recommended conventions [26] [28] [29] when using the PCA/*t*-test protocol for testing unidimensionality in the RM.

4. Methods

Binomial 95% CIs were calculated according to four commonly used methodologies: the normal approximation 95% CI (the "Wald" method), the "exact" binomial CI according to Clopper-Pearson [42], and the Wilson, Agresti-Coull, and Jeffreys methods [39], as implemented in Stata version 13.1 for Mac OS X (StataCorp, College Station, TX, USA). These binomial CIs were calculated for hypothesized observed proportions of 0.06 (6%), 0.08 (8%) and 0.10 (10%) and sample sizes ranging from $n = 100$ to $n = 1000$ in increments of 100, and thereafter in increments of 500 up to $n = 2500$. Lower 95% CIs were inspected regarding their coverage of the 5% (*i.e*., 0.05) threshold.

5. Results

Results are depicted in **Figure 1**. Normal approximation ("Wald") 95% CIs included 5% with sample sizes of $n = 100$ - 2000 and a 6% observed proportion, $n = 100$ - 300 with an 8% observed proportion, and $n = 100$ with a 10% observed proportion (**Figure 1(a)**). "Exact" 95% CIs included 5% with sample sizes of $n = 100$ - 2000 with a 6% observed proportion, $n = 100$ - 200 with an 8% observed proportion and $n = 100$ with a 10% observed proportion (**Figure 1(b)**). The Wilson, Agresti-Coull, and Jeffreys 95% CIs all included 5% with sample sizes of $n = 100$ - 1500 and a 6% observed proportion as well as with sample sizes of $n = 100$ - 200 with an 8% observed proportion, but not for any sample size with a 10% observed proportion (**Figure 1(c)-(e)**).

6. Discussion

The results presented here are fully expected [38]-[41]. However, the lack of these types of considerations in papers presenting results of RM based rating scale analyses suggests that there is a need for reiteration. For

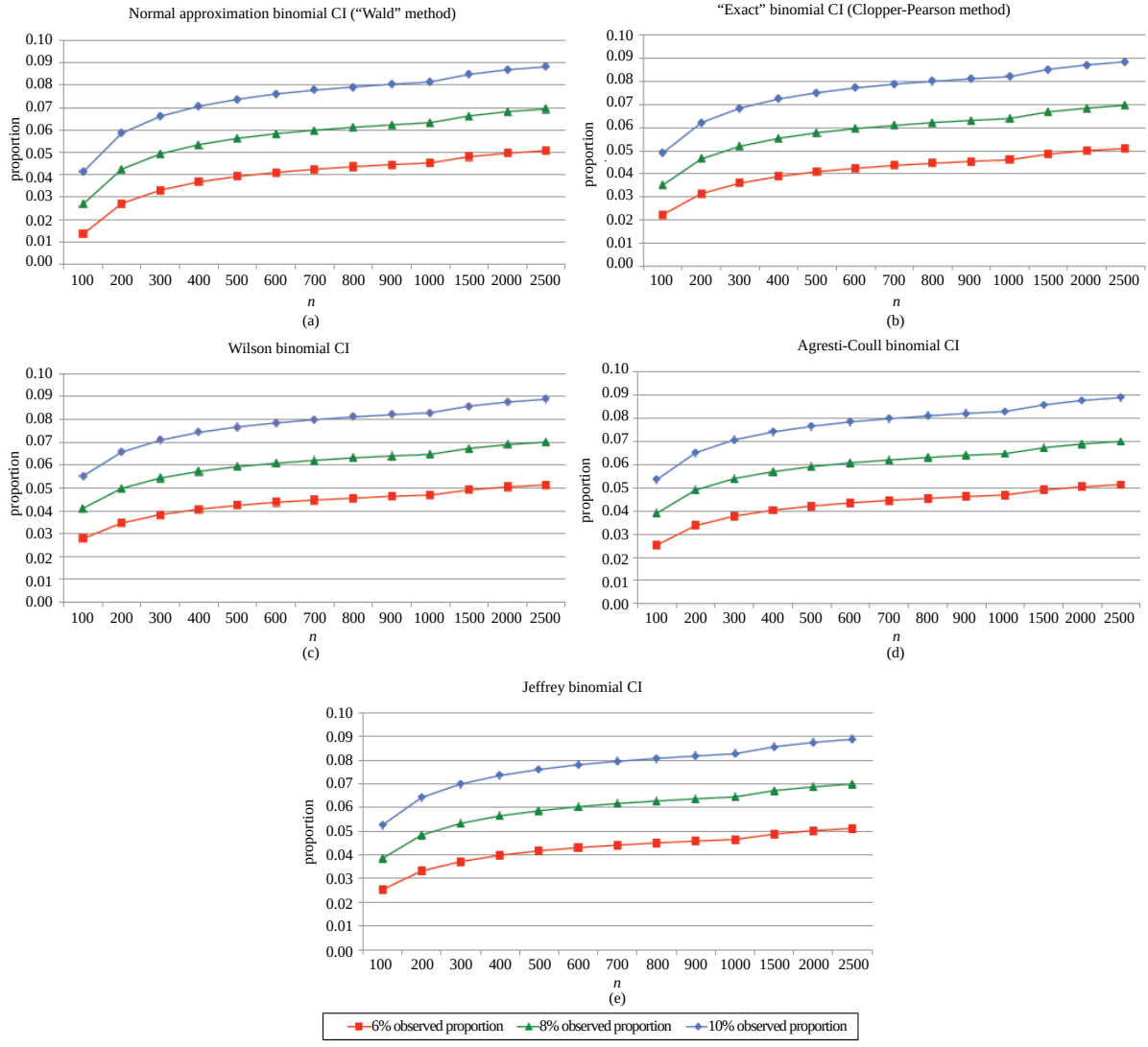

Figure 1. Results from calculations of binomial CIs for hypothesized observed proportions of 0.06 (6%), 0.08 (8%) and 0.10 (10%) and sample sizes ranging from n = 100 to n = 2500. Curves represent the lower 95% CI bounds according to (a) the normal approximation ("Wald"), (b) the "exact" Clopper-Pearson, (c) the Wilson, (d) Agresti-Coull, and (e) Jeffreys methodologies for estimating the 95% binomial CI.

example, Ramp and collaborators [33] used the PCA/*t*-test protocol for testing unidimensionality in the RM with the 20-item physical impact scale of the Multiple Sclerosis Impact Scale (MSIS-29) with a sample of n = 92 people with multiple sclerosis. Results showed that 9.2% of the person measures from two PCA derived item subsets differed and the 95% binomial CI ranged 4% - 14%, which led the authors to infer unidimensionality. Young and coworkers [37] used the same methodology with a 17-item scale purported to measure self-efficacy with a sample of n = 309 people with multiple sclerosis and found 12.2% of the person measures from two PCA derived item subsets to differ (95% binomial CI, 9.8% - 14.7%). The authors reported the scale to exhibit "considerable multidimensionality" (p. 1329). Despite relatively similar observed proportions the two conclusions are in opposite directions due to a more than two-fold larger width of the 95% binomial CI in the former as compared to the latter study, resulting in coverage and noncoverage of the 5% threshold, respectively. However, none of the studies commented on the sample size or method for estimating the 95% binomial CIs in relation to these results. This is not to criticize these or any particular investigators or rating scales, but mentioned here merely to illustrate the problem.

The results presented here illustrate that the PCA/*t*-test protocol for testing unidimensionality in the RM is, as

any other statistical test [40] [41], dependent on sample size. For example, in a study (n = 473) regarding the cross-diagnostic measurement properties of the Nottingham Health Profile index of Distress (NHPD) using the PCA/*t*-test protocol for testing unidimensionality, it was found that the lower binomial normal approximation 95% CI did not overlap 5% [43]. However, if for example two thirds of that sample size had been used instead (with the same proportion of significant individual *t*-tests), the statistical conclusion from this test would instead have supported unidimensionality [43]. Therefore, and despite that the PCA/*t*-test protocol has appeared more useful in detecting multidimensionality than residual based fit indices and factor analytic approaches [7] [26] it must be borne in mind that this procedure, in itself, also is a somewhat arbitrary test. Indeed, inferences are dependent on and, therefore, differ according to sample sizes [40]. Over reliance on statistical tests and their sensitivity to sample size in determining the extent to whether data fit the unidimensional Rasch measurement model was cautioned against already by Rasch [19]:

"On the whole we should not overlook that since a model is never true, but only more or less adequate, deficiencies are bound to show, given sufficient data" (p 92).

One strategy to take the influence of sample size into account when using tests such as the PCA/*t*-test protocol for testing unidimensionality (or other aspects of fit to the RM), would be to conduct and report sensitivity analyses, similar to what is standard practice in, e.g. health economic cost analyses [44]. That is, by keeping all aspects of the data constant and varying only the n in the calculation of CIs (or P-values), it would be possible to consider the influence of sample size in the particular study. As illustrated in this paper, such a procedure is easily adapted to the calculation of binomial CIs, and specialized RM software such as RUMM2030 incorporates this facility for other hypothesis testing procedures in the analysis of fit to the RM [31].

In addition to the influence of sample size, the results presented here also illustrates that the choice of method for estimating the 95% binomial CI influences the results and conclusions from using the PCA/*t*-test protocol for testing unidimensionality in the RM. Specifically, the Wilson, Agresti-Coull, and Jeffreys 95% CIs appear to yield the most stable estimates, followed by the Clopper-Pearson "exact" binomial CI, and with the normal approximation ("Wald") 95% CI appearing to be the most problematic one. These observations are in general agreement with previous studies on the properties of various methods for estimating the 95% binomial CI [38] [39]. For example, Brown et al. found the normal approximation 95% CI (the "Wald" method) to exhibit a highly erratic behaviour in their examination of the actual interval covered, which oscillated considerably in relation both to sample size and the observed proportion, with the actual CI coverage rarely approximating 95% [39]. In contrast, the Wilson and Agresti-Coull 95% CIs behaved much more reliably (particularly for small and large sample sizes, respectively) [39]. This aspect has rarely been considered in RM based studies and it is therefore recommended that authors who choose to apply the PCA/*t*-test protocol for testing unidimensionality in the RM also need to specify the estimation method used for calculating the 95% binomial CI. Furthermore, as this and previous studies illustrate there are good reasons to avoid the normal approximation 95% CI ("Wald" method), which appears to be the default in many software applications, in favour for, e.g. the Agresti-Coull 95% CI.

Other aspects of the PCA/*t*-test protocol for testing unidimensionality in the RM also need to be considered. First, although sometimes considered nonproblematic with sample sizes above 200 [45], methods such as PCA assumes that data are normally distributed, which rarely appears to be considered in the application of the PCA/*t*-test protocol for testing unidimensionality in the RM [32]-[37] [43]. Secondly, the rationale for the suggested loading of 0.3 as a cut-off to define items to be included in the PCA/*t*-test protocol [26] is unclear and other criteria could also be conceivable; additional studies regarding the optimal approach to using this procedure are warranted.

Unidimensionality is not an absolute but a relative matter and there is no single agreed upon method to test for unidimensionality. Therefore, the decision whether a scale is sufficiently unidimensional should ultimately come from outside the data and be driven by the purpose of measurement and clinical/theoretical considerations [10]. As reviewed above, the most important aspect of the dimensionality issue relates to the central role of variable definition, frame of reference, perspective and context of application, and the fact that unidimensionality is not an "either/or" issue but a relative matter of degree [10] [12] [16]. For example, due to inherent psychometric problems with the 39-item Parkinson's Disease Questionnaire (PDQ-39) [46]-[49] we revisited this scale from a conceptual and theoretical perspective according to the World Health Organization's International Classification of Functioning, Disability and Health (ICF) taxonomy [50] using RM analyses, including the PCA/*t*-test protocol for testing unidimensionality [51]. The analyses identified four ICF related item sets including two represent-

ing body functioning, of which one was dismissed as invalid despite meeting standard RM fit criteria and the PCA/*t*-test procedure criterion for unidimensionality. The reason for this conclusion was that the item set did not appear to represent an interpretable underpinning common latent variable. Specifically, the item set covered pain, poor memory, feeling unpleasantly hot or cold, painful cramps or spasms, falling asleep unexpectedly, and distressing dreams or hallucinations (as ordered from more to less impaired body functioning). While each of these items represents body functions (which, as all components of the ICF, in itself is relatively unspecific and broad in nature), it is unclear what common variable would manifest itself in this manner (as expressed by the above item hierarchy) as one moves up and down a scale from less to more. Its clinical meaning is therefore doubtful, regardless of the results from statistical tests.

This is not surprising since any quantitative analysis merely is based on numbers that may or may not work together in a particular way regardless of what they represent and whether they represent matters that are meaningful and interpretable or not. This line of reasoning is similar to that underpinning the inappropriateness of coefficient alpha as an index of dimensionality [52], and illustrates the fundamental problem of any data driven approach to rating scale development. Instead, assessment of dimensionality requires a good deal of judgment and the importance of prioritizing a theory driven approach over a data driven one in order to achieve interpretable and useful rating scale derived measures cannot be under emphasized [13]. That is, rating scales and their items should ultimately be developed and selected based on explicit definitions of the variables that they are intended to measure. Basically, the results of an RM analysis should uncover a pattern that is coherent with theory, according to which the hierarchical ordering of items should represent a meaningful story about what it means to move up and down the scale for the variable of interest [20]. For "established" rating scales that may or may not have been developed based on a clear variable definition, item sets still need to be hierarchically and substantively meaningful and interpretable as representatives of a common latent variable. RM analysis provides an integrated framework for analyzing the extent to which this has been achieved, and a means to detect and diagnose anomalies that can be used to refine the scale, its theoretical underpinnings, or both, in view of the perspective, context and purpose of measurement.

7. Conclusion

In conclusion, use and interpretation of results from the PCA/*t*-test protocol for testing unidimensionality in the RM must be made with the same considerations as with any hypothesis testing procedure and is dependent on sample size as well as choice of estimation method for the 95% binomial CI. The PCA/*t*-test procedure should not be viewed as a "definite" test for unidimensionality and does not replace an integrated quantitative/qualitative interpretation based on an explicit variable definition in view of the perspective, context and purpose of measurement. Statistical procedures and reliance on P-values and CIs cannot compensate for conceptual and theoretical considerations. It is recommended that when the PCA/*t*-test protocol is used for testing unidimensionality in the RM, it should be accompanied by sensitivity analyses (or similar considerations) with respect to the influence of sample size and the type of binomial CI used should be specified (avoiding the normal approximation CI). However, this and other data driven statistical procedures should only be applied under and following careful theoretical consideration of the rating scale at hand and its underpinning target latent variable.

Acknowledgements

The author wants to thank Albert Westergren for valuable discussions. This work was supported by Kristianstad University, Kristianstad, Sweden and was in part conducted within the Basal Ganglia Disorders Linnaeus Consortium (BAGADILICO) at Lund University, Lund, Sweden.

References

[1] Hobart, J.C., Cano, S.J., Zajicek, J.P. and Thompson, A.J. (2007) Rating Scales as Outcome Measures for Clinical Trials in Neurology: Problems, Solutions, and Recommendations. *Lancet Neurology*, **6**, 1094-1105. http://dx.doi.org/10.1016/S1474-4422(07)70290-9

[2] Bollen, K. and Lennox, R. (1991) Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, **110**, 305-314. http://dx.doi.org/10.1037/0033-2909.110.2.305

[3] Fayers, P.M. and Hand, D.J. (2002) Causal Variables, Indicator Variables and Measurement Scales: An Example from Quality of Life. *Journal of the Royal Statistical Society*, **165**, 233-266. http://dx.doi.org/10.1111/1467-985X.02020

[4] Stenner, A.J., Stone, M.H. and Burdick, D.S. (2009) Indexing vs. Measuring. *Rasch Measurement Transactions*, **22**, 1176-1177.

[5] Nunnally, J.C. and Bernstein, I.H. (1994) Psychometric theory. McGraw-Hill, Inc., New York.

[6] Thurstone, L.L. (1931) The Measurement of Social Attitudes. *The Journal of Abnormal and Social Psychology*, **26**, 249-269. http://dx.doi.org/10.1037/h0070363

[7] Smith Jr., E.V. (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.

[8] Stout, W. (1987) A Nonparametric Approach for Assessing Latent Trait Unidimensionality. *Psychometrika*, **52**, 589-617. http://dx.doi.org/10.1007/BF02294821

[9] Food and Drug Administration (2009) Patient-Reported Outcome Measures: Use in Medicinal Product Development to Support Labelling Claims. Food and Drug Administration, Washington DC.

[10] Andrich, D. (1988) Rasch Models for Measurement. Sage Publications, Inc., Beverly Hills.

[11] Andrich, D. (2002) Implications and Applications of Modern Test Theory in the Context of Outcomes Based Education. *Studies in Educational Evaluation*, **28**, 103-121.http://dx.doi.org/10.1016/S0191-491X(02)00015-9

[12] Cano, S.J., Barrett, L.E., Zajicek, J.P. and Hobart, J.C. (2011) Dimensionality Is a Relative Concept. *Multiple Sclerosis*, **17**, 893-894. http://dx.doi.org/10.1177/1352458511406910

[13] Yo, C.H., Osborn Popp, S., DiGangi, S. and Jannasch-Pennell, A. (2007) Assessing Unidimensionality: A Comparison of Rasch Modeling, Parallel Analysis, and TETRAD. *Practical Assessment Research & Evaluation*, **12**, 19 p.

[14] Kelley, T.L. (1942) The Reliability Coefficient. *Psychometrika*, **7**, 75-83.http://dx.doi.org/10.1007/BF02288068

[15] Hattie, J. (1985) Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, **9**, 139-164. http://dx.doi.org/10.1177/014662168500900204

[16] Hobart, J. and Cano, S. (2009) Improving the Evaluation of Therapeutic Interventions in Multiple Sclerosis: The Role of New Psychometric Methods. *Health Technology Assessment*, **13**, 1-177.

[17] Horton, M., Marais, I. and Christensen, K.B. (2013) Dimensionality. In: Christensen, K.B., Kreiner, S. and Mesbah, M., Eds., *Rasch Models in Health*, John Wiley & Sons, Inc., Croydon, Surrey, 137-158.

[18] Andrich, D. (2009) Interpreting RUMM2030 Part IV: Multidimensionality and Subtests in RUMM. RUMM Laboratory Pty Ltd., Perth.

[19] Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks Paedagogiske Institut, Copenhagen.

[20] Stenner, A.J., Fisher Jr., W.P., Stone, M.H. and Burdick, D.S. (2013) Causal Rasch Models. *Frontiers in Psychology*, **4**, 536. http://dx.doi.org/10.3389/fpsyg.2013.00536

[21] Guttman, L. (1944) A Basis for Scaling Qualitative Data. *American Sociological Review*, **9**, 139-150. http://dx.doi.org/10.2307/2086306

[22] Andrich, D. (1985) An Elaboration of Guttman Scaling with Rasch Models for Measurement. In: Brandon-Tuma, N., Ed., *Sociological Methodology*, Jossey-Bass, San Francisco, 33-80.

[23] Wright, B.D. and Bell, S.R. (1984) Item Banks: What, Why, How. *Journal of Educational Measurement*, **21**, 331-345. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01038.x

[24] Andrich, D., Sheridan, B. and Luo, G. (2009) Interpreting RUMM2030. RUMM Laboratory Pty Ltd., Perth.

[25] Smith, R.M. (1996) A Comparison of Methods for Determining Dimensionality in Rasch Measurement. *Structural Equation Modeling*, **3**, 25-40. http://dx.doi.org/10.1080/10705519609540027

[26] Tennant, A. and Pallant, J. (2006) Unidimensionality Matters. *Rasch Measurement Transactions*, **20**, 1048-1051.

[27] Linacre, J.M. (1998) Detecting Multidimensionality: Which Residual Data-Type Works Best? *Journal of Outcome Measurement*, **2**, 266-283.

[28] Tennant, A. and Conaghan, P.G. (2007) 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 Care & Research*, **57**, 1358-1362. http://dx.doi.org/10.1002/art.23108

[29] Horton, M. and Tennant, A. (2010) Assessing Unidimensionality Using Smith's (2002) Approach in RUMM 2030. Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Copenhagen.

[30] Cowles, M. and Davis, C. (1982) On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, **37**, 553-558. http://dx.doi.org/10.1037/0003-066X.37.5.553

[31] Andrich, D., Sheridan, B. and Luo, G. (1997-2012) RUMM2030: Rasch Unidimensional Models for Measurement. RUMM Laboratory, Perth.

[32] Forjaz, M.J., Martinez-Martin, P., Dujardin, K., Marsh, L., Richard, I.H., Starkstein, S.E. and Leentjens, A.F. (2013) Rasch Analysis of Anxiety Scales in Parkinson's Disease. *Journal of Psychosomatic Research*, **74**, 414-419. http://dx.doi.org/10.1016/j.jpsychores.2013.02.009

[33] Ramp, M., Khan, F., Misajon, R.A. and Pallant, J.F. (2009) Rasch Analysis of the Multiple Sclerosis Impact Scale MSIS-29. *Health and Quality of Life Outcomes*, **7**, 58. http://dx.doi.org/10.1186/1477-7525-7-58

[34] Riazi, A., Aspden, T. and Jones, F. (2014) Stroke Self-Efficacy Questionnaire: A Rasch-Refined Measure of Confidence Post Stroke. *Journal of Rehabilitation Medicine*, **46**, 406-412. http://dx.doi.org/10.2340/16501977-1789

[35] Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J. and Weich, S. (2009) Internal Construct Validity of the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): A Rasch Analysis Using Data from the Scottish Health Education Population Survey. *Health and Quality of Life Outcomes*, **7**, 15. http://dx.doi.org/10.1186/1477-7525-7-15

[36] Tur, B.S., Kucukdeveci, A.A., Kutlay, S., Yavuzer, G., Elhan, A.H. and Tennant, A. (2009) Psychometric Properties of the WeeFIM in Children with Cerebral Palsy in Turkey. *Developmental Medicine and Child Neurology*, **51**, 732-738. http://dx.doi.org/10.1111/j.1469-8749.2008.03255.x

[37] Young, C.A., Mills, R.J., Woolmore, J., Hawkins, C.P. and Tennant, A. (2012) The Unidimensional Self-Efficacy Scale for MS (USE-MS): Developing a Patient Based and Patient Reported Outcome. *Multiple Sclerosis Journal*, **18**, 1326-1333. http://dx.doi.org/10.1177/1352458512436592

[38] Newcombe, R.G. (1998) Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*, **17**, 857-872. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

[39] Brown, L.D., Cai, T.T. and DasGupta, A. (2001) Interval Estimation for a Binomial Proportion. *Statistical Science*, **16**, 101-133. http://dx.doi.org/10.1214/ss/1009213286

[40] Feinstein, A.R. (1998) P-Values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin. *Journal of Clinical Epidemiology*, **51**, 355-360. http://dx.doi.org/10.1016/S0895-4356(97)00295-3

[41] McCormack, J., Vandermeer, B. and Allan, G.M. (2013) How Confidence Intervals Become Confusion Intervals. *BMC Medical Research Methodology*, **13**, 134. http://dx.doi.org/10.1186/1471-2288-13-134

[42] Clopper, C.J. and Pearson, E.S. (1934) The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, **26**, 404-413. http://dx.doi.org/10.1093/biomet/26.4.404

[43] Wann-Hansson, C., Klevsgard, R. and Hagell, P. (2008) Cross-Diagnostic Validity of the Nottingham Health Profile Index of Distress (NHPD). *Health and Quality of Life Outcomes*, **6**, 47. http://dx.doi.org/10.1186/1477-7525-6-47

[44] Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J. and Stoddart, G.L. (2005) Methods for the Economic Evaluation of Health Care Programmes. 3rd Edition, Oxford University Press, New York.

[45] Hair, J.F., Black, B., Babin, B., Anderson, R.E. and Tatham, R.L. (2006) Multivariate Data Analysis. 6th Edition, Prentice Hall, Upper Saddle River, NJ.

[46] Hagell, P. (2005) Feasibility and Linguistic Validity of the Swedish Version of the PDQ-39. *Expert Review of Pharmacoeconomics & Outcomes Research*, **5**, 131-136. http://dx.doi.org/10.1586/14737167.5.2.131

[47] Hagell, P. and Nilsson, M.H. (2009) The 39-Item Parkinson's Disease Questionnaire (PDQ-39): Is It a Unidimensional Construct? *Therapeutic Advances in Neurological Disorders*, **2**, 205-214. http://dx.doi.org/10.1177/1756285609103726

[48] Hagell, P. and Nygren, C. (2007) The 39 Item Parkinson's Disease Questionnaire (PDQ-39) Revisited: Implications for Evidence Based Medicine. *Journal of Neurology, Neurosurgery & Psychiatry*, **78**, 1191-1198. http://dx.doi.org/10.1136/jnnp.2006.111161

[49] Hagell, P., Whalley, D., McKenna, S.P. and Lindvall, O. (2003) Health Status Measurement in Parkinson's Disease: Validity of the PDQ-39 and Nottingham Health Profile. *Movement Disorders*, **18**, 773-783. http://dx.doi.org/10.1002/mds.10438

[50] World Health Organization (2001) International Classification of Functioning, Disability and Health: ICF. World Health Organization, Geneva.

[51] Nilsson, M.H., Westergren, A., Carlsson, G. and Hagell, P. (2010) Uncovering Indicators of the International Classification of Functioning, Disability, and Health from the 39-Item Parkinson's Disease Questionnaire. *Parkinson's Disease*, **2010**, Article ID: 984673. http://dx.doi.org/10.4061/2010/984673

[52] Cortina, J.M. (1993) What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, **78**, 98-104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.