Educational Methods & Psychometrics—Vol. 4, Article No. 30
April 2026
https://doi.org/10.65301/emp.2026.264

TRANSPARENCY IN PSYCHOMETRIC REPORTING: A REVIEW OF SCALES FOR WELL-BEING AND QUALITY OF LIFE IN OLDER PERSONS EMPLOYING RASCH ANALYSIS

Marit Preuter

RISE, Research Institutes of Sweden, Division Built Environment, Department System Transition and Service Innovation, Social Sustainability, Gothenburg, Sweden

Marie-Louise Möllerberg

Malmö University, Faculty of Health and Society, Department of Care Science, Malmö, Sweden

Kristofer Årestedt

Linneaus University, Faculty of Health and Life Science, Kalmar, Sweden

Department of Research, Region Kalmar County, Kalmar, Sweden

Jeanette Melin *

Linneaus University, Faculty of Health and Life Science, Kalmar, Sweden

Swedish Defence University, Department of Leadership, Demand and Control, Karlstad, Sweden

Södertörn University, Institute of Social Sciences, Stockholm, Sweden

There is a growing interest in assessing well-being and quality of life (QoL) among older people. To ensure that evidence-based scales are used for this purpose, comprehensive and transparent reporting of measurement properties is an essential first step. This study aimed to evaluate the reporting quality of studies that used Rasch analysis to assess scales designed to measure well-being or QoL in older persons. Articles employing Rasch analysis were identified through a previous broader systematic review of psychometric studies. The findings demonstrate notable deficiencies in the reporting of measurement properties, indicating substantial room for improvement. While detailed reporting alone cannot guarantee that scales possess satisfactory measurement quality, it constitutes a prerequisite for drawing evidence-based conclusions about their measurement quality. Improved reporting practices are essential not only for enhancing the interpretability and replicability of individual studies but also for enabling informed decisions in clinical and policy contexts regarding the use of well-being and QoL scales for older populations.

Keywords: Psychological well-being, Psychometrics, Reporting Quality, Self-assessment, Quality Assessment, Quality Improvement, Quality of Life

Correspondence should be made to Jeanette Melin, Linneaus University, Faculty of Health and Life Science, Kalmar, Sweden. Email: jeanette.melin@lnu.se

1. Introduction

Measurements are a key component when comparing and synthesizing research, and replication is a cornerstone of all forms of research. Yet, comparability between measures is sparsely addressed (Hanisch et al., 2019; Johansson et al., 2023; Nugent, 2011) and questionable measurement practices are a threat to the validity, reliability, and replicability of studies (Flake et al., 2022; Open Science Collaboration, 2015). Importantly, neither rigorous research design, advanced statistical techniques, nor large sample sizes can make an invalid measurement valid afterward (Flake & Fried, 2020).

When aiming to measure a latent trait, such as an individual's well-being and quality of life (QoL), responses to multiple items on an ordinal scale are typically used to measure the latent trait of interest. This implies that a measurement model must handle ordered categories and disentangle the person and item attributes to validate whether the set of items – hereafter referred to as a scale – can generate valid and comparable measures across individuals (Cano et al., 2019). In this regard, the Rasch model (1960) offers a unique perspective, grounded in measurement science and metrology (Andrich, 2004; Pendrill, 2014), providing psychometric properties and interpretability unmatched by other models.

A Rasch analysis – examining how well data fit the model – is useful for evaluating claims about a scale's measurement quality. However, no single indices can determine measurement quality; rather, multiple fit statistics and measurement properties must be assessed and interpreted through an iterative process (Tennant & Conaghan, 2007; Tesio, Caronni, Kumbhare, et al., 2023; Tesio, Caronni, Simone, et al., 2023). To meaningfully evaluate the measurement quality of existing scales, psychometric studies must include comprehensive reporting of sample characteristics, construct definitions, and measurement properties (Dabaghi et al., 2020; Johansson et al., 2023; Mallinson et al., 2022; Van de Winckel et al., 2022). Moreover, it is essential to distinguish between a scale's measurement quality and the reporting quality in a study: while thorough reporting enables valid conclusions about a scale's measurement quality, sparse reporting hinders the assessment of a scale's measurement quality.

Furthermore, since prevailing measurement paradigms tend to privilege statistical evidence, numerical indicators often dominate the evaluation of measurement instruments (Salzberger, 2012). Nevertheless, meaningful measurement of latent traits requires an underlying theory of the construct that guides both the development of item content and expectations regarding the ordering of items along the continuum (c.f., Barbic et al., 2018; Salzberger, 2012). This is typically referred to as a *construct theory* or *substantive theory*. Without such a theoretical foundation, interpretations of measurements of the latent trait risk become more illusory than informative (Stenner et al., 2013). Consequently, Rasch analyses should not only report statistical results but also describe the construct theory informing the measurement and report how the empirical findings relate to that theory. Nevertheless, Rasch analyses may at times be applied more exploratively. For instance, when theory and practice inform each other in a dialectical process, this can contribute to a deeper understanding of the latent trait of interest (Bond et al., 2020).

Positioning measurements of well-being and QoL within ageing research is both timely and necessary. Given the growing share of older persons (Europäische Kommission, 2019; United Nations, 2024), there is a rising interest in measuring well-being among older people, but at the same time, challenges with definition and heterogeneity among items in scales exist (Halvorsrud & Kalfoss, 2007; Linton et al., 2016; Preuter et al., 2025). This underscores a significant knowledge gap regarding the extent to which existing scales validly can capture the same latent trait, that is well-being and QoL among older persons. Moreover, the ability to validly and reliably measure these latent traits depends not only on theoretical definitions but

also on the psychometric properties of the scales, which in turn require transparent and comprehensive reporting.

2. Methods

Study design

This study examined the quality of measurement reports in scales used to evaluate well-being and QoL among older persons. Articles employing Rasch analysis were extracted from a previous broader systematic review of psychometric studies (Preuter et al., 2025).

Data collection

A systematic review of scales used to measure well-being and quality of life among older persons was conducted between 2024 and 2025 (Preuter et al., 2024, 2025). The systematic review began with literature searches across PubMed, PsycInfo, and CINAHL and was structured into four blocks: i. the population (i.e., older persons), ii. concept (i.e., well-being or quality of life), iii. measurements (i.e., self-reports or questionnaires) and iv. methods (i.e., psychometrics). Further details, including strategies for each database, are presented in the original literature review (Preuter et al., 2025). Of the 120 identified articles, 9 used a Rasch analysis for evaluating the measurement properties and were selected for the present study.

Reporting quality assessment

To assess the reporting quality of studies applying Rasch analysis, Dabaghi and colleagues (2020) developed a checklist based on established criteria for documenting psychometric assessment studies. This checklist was further refined by Johansson and colleagues (2023). In parallel, the Rasch reporting guideline for rehabilitation research, RULER, was introduced (Mallinson et al., 2022; Van de Winckel et al., 2022). However, none of these checklists were found to be fully comprehensive or entirely applicable to the present study, as they included context-specific elements and made preferences regarding statistical analyses and cut-off values. Therefore, we adapted and expanded upon the earlier checklists, together with Rasch methodological textbooks and articles (Andrich & Marais, 2019; Christensen et al., 2013; Hobart & Cano, 2009; Tennant & Conaghan, 2007; Tesio, Caronni, Kumbhare, et al., 2023; Tesio, Caronni, Simone, et al., 2023), and general principles of study quality and transparency (e.g., clear sample description, specification of model assumptions; von Elm et al., 2007). In turn, this resulted in 13 reporting areas, each comprising one to seven quality criteria (Table 1).

Because Rasch analysis rests on explicit model assumptions, transparent reporting is essential for interpretability and replication. Several criteria are conceptually related, and therefore, the checklist was designed as a descriptive inventory of reporting rather than a cumulative quality score. Moreover, it should be acknowledged that the checklist used may benefit from further collaborative development, as broader consensus among a larger group of researchers could help refine and validate the criteria. However, such efforts are beyond the scope of the present study. Importantly, the structured overview provided by the checklist enables readers to assess the extent to which conclusions regarding a scale's measurement quality can be supported by the available reporting.

Table 1.
List of criterion applied for assessing reporting quality.

#	Criterion
1	a) Is a construct theory presented and/or referenced?
	b) Was the construct theory tested?
2	a) Are the sample size and the demographic characteristics of the participants reported?
	b) Is the sample size justified?
	c) If the sample size may be considered too small or too large in relation to the statistical analyses conducted, are the implications discussed?
	d) Is missing data reported?
	e) In case of missing data, is it reported how it was handled?
3	a) Is the software for Rasch model analysis, including software version, stated?
	b) Which software was used?
	c) If R was used, is the package(-s) stated?
	d) If R was used, which package(-s) was used?
	e) Is the Rasch model used stated? (E.g., RSM or PCM)
	f) Which parameter estimation method was used? (E.g., conditional or unconditional estimation)?
4	a) Are tests of item fit to the Rasch model reported?
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)
5	a) Are tests of person fit to the Rasch model reported?
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)
6	a) Are tests of DIF reported?
	b) Which tests of DIF are reported?
7	a) Is an evaluation of LD reported?
	b) Which test/-s of LD were used?
8	a) Is an evaluation of dimensionality reported?
	b) Which test/-s of dimensionality were used?
9	Is an evaluation of the threshold ordering reported? (only applicable for polytomous items)
10	a) Is the proportion or N of participants with max and min total scores (i.e. extremes) reported?
	b) Is the proportion or N of participants that has a location lower than the lowest item threshold and the proportion or N of participants that has a location higher than the highest item threshold reported?
11	a) Is a Wright map or similar figure presented to illustrate sample-to-item targeting?
	b) Are item locations mean and SD reported?
	c) Are person locations mean and SD reported?
12	a) Are person reliability coefficient(s) reported?
	b) Is a test information function figure reported?
	c) Is the proportion (%) of participants located at levels where test information meets the chosen reliability criterion reported?
13	a) Are the item locations reported for each individual item?
	b) Is a transformation table provided that enables raw scores to be converted into linear measurement values?
	c) If a transformation table is provided, is measurement uncertainty (standard error) reported?

DIF = differential item functioning; LD = local dependency; MNSQ = mean square fit statistic; PCM = partial credit model; RSM = rating scale model; SD = standard deviation; SE = standard error; TIF = test information function; ZSTD = standardized fit statistic.

Each criterion was assessed as follows: a value of 1 was assigned if the criterion was reported, and 0 if it was not reported. In addition, “not applicable” could also be used. For instance, if a study did not report item fit statistics (e.g., answered "no" to item 4a: “Are tests of item fit to the Rasch model reported?”), the subsequent item 4b (“Which fit indices are reported?”) was marked as ”not applicable”.

Each of the nine articles was reviewed by two or three authors, with all authors reviewing at least four articles each. In cases of divergent assessments, the authors discussed their reasoning and reached a consensus. All reviewers agreed with criteria 2a, 3c, 3f, 6a, 12c, and 13b (Appendix 1). In contrast, criteria 1a and 13c showed agreement in only one of the nine articles. For criterion 13c, it became apparent that "not applicable" had been interpreted inconsistently. Regarding criterion 1a, clarification was needed on what constituted a minimum standard for considering a criterion as reported. We agreed that fulfilling the criterion required a clear presentation of, or reference to, an underlying construct theory or definition of well-being or QoL that informed the development of the scale. This involved describing how the construct's conceptualisation guided both the selection of item content and the expected ordering of items along the latent continuum. Considering this, we also introduced Criterion 1b to capture whether the authors empirically examined the extent to which the observed item structure aligns with their proposed construct theory.

3. Results

Table 2 presents the assessment of each criterion for the included articles. Two studies (Chachamovich et al., 2008; Forjaz et al., 2012) reported at least half of the applicable criteria. One paper had minimal reporting of Rasch analysis; Lucas-Carrasco and colleagues (2012) reported only sample size (# 2a) and evaluation of dimensionality (# 8a).

Presenting the sample with size and demographics (# 2a, $n = 8$), tests of fit to the Rasch model (# 4a, $n = 7$), DIF (# 6a, $n = 7$), evaluation of threshold ordering (# 9, $n = 7$) and reliability (# 12a, $n = 7$) were most commonly reported while criterion for parameter estimation (# 3f), participants with locations outside threshold range (# 10b), item locations (# 11b), test information function figure (# 12b) test information function values (# 12c) and standard errors in transformations tables (# 13c) were not reported in any article (Table 2).

Table 2.
Reporting of methodological criteria across included.

	Scale	**WHOQOL-OLD**	**WHOQOL-OLD**	**WHOQOL-OLD**	**PWI**	**WHOQOL-OLD**	**WHOQOL-BREF**	**WHO-5**	**CASP-19**	**WHOQOL-OLD**
	Reference	(Chachamovich et al., 2008)	(Conrad et al., 2014)	(Fang et al., 2012)	(Forjaz et al., 2012)	(Gondodiputro et al., 2021)	(Liang et al., 2009)	(Lucas-Carrasco et al., 2012)	(Oluboyede & Smith, 2013)	(Power et al., 2005)
#	Respondents	Brazilians ≥ 60 years	German ≥ 60 years	International ≥ 60 years	Spanish ≥ 60 years	Indonesian ≥ 60 years	Taiwanese ≥ 65 years	Spanish ≥ 65 years	English ≥ 50 years	International ≥ 60 years
1	a) Is a construct theory presented and/or referenced?	0	0	0	0	0	0	0	0	0
	b) Was the construct theory tested?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
2	a) Are the sample size and the demographic characteristics of the participants reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**
	b) Is the sample size justified?	**1**	0	0	0	0	0	0	0	0
	c) If the sample size may be considered too small or too large in relation to the statistical analyses conducted, are the implications discussed?	Not applicable	**1**	0	0	Not applicable	Not applicable	Not applicable	Not applicable	0
	d) Is missing data reported?	**1**	0	0	0	**1**	**1**	0	0	**1**
	e) In case of missing data, is it reported how it was handled?	0	Not applicable	**1**	Not applicable	0	0	Not applicable	Not applicable	**1**
3	a) Is the software for Rasch model analysis, including software version, stated?	**1**	0	**1**	**1**	**1**	**1**	0	0	**1**

	b) Which software was used?	RUMM2020	Not applicable	WINSTEPS 3.697 RUM2020	RUMM2020	WINSTEPS 3.73	WINSTEPS 3.47	Not applicable	Not applicable	RUMM2010
	c) If R was used, is the package(-s) stated?	Not applicable	**1**	Not applicable	Not applicable	Not applicable	Not applicable	99 Not applicable	99 Not applicable	Not applicable
	d) If R was used, which package(-s) was ised?	99	eRm, ltm	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
	e) Is the Rasch model used stated? (E.g., RSM or PCM)	0	**1**	**1**	**1**	0	**1**	0	**1**	**1**
	f) Which parameter estimation method was used? (E.g., conditional or unconditional estimation)?	0	0	0	0	0	0	0	0	0
4	a) Are tests of item fit to the Rasch model reported?	**1**	**1**	**1**	**1**	0	**1**	0	**1**	**1**
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)	Chi2, Residuals	Infit MNSQ	Chi2, Residuals, Infit MNSQ, Outfit MNSQ	Residuals	Not applicable	Infit index	99	Infit MNSQ	Chi2, Residuals
5	a) Are tests of person fit to the Rasch model reported?	0	0	0	**1**	0	0	0	0	0
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)	Not applicable	Not applicable	Not applicable	Residuals	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
6	a) Are tests of DIF reported?	**1**	0	**1**	**1**	**1**	**1**	0	**1**	**1**
	b) Which tests of DIF are reported?	0	Not applicable	0	0	Probability value is below 5%	DIF Size	Not applicable	t-test	Logistic regression
7	a) Is an evaluation of LD reported?	**1**	0	0	**1**	0	0	0	**0**	0
	b) Which test/-s of LD were used?	Residual correlations with absolute cut off	Not applicable	Not applicable	Residual correlations with absolute cut off	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

8	a) Is an evaluation of dimensionality reported?	**1**	0	0	**1**	**1**	0	**1**	**1**	0
	b) Which test/-s of dimensionality were used?	Item trait interaction Chi2	Not applicable	Not applicable	PCA of residuals and t-test	PCA	Not applicable	ICC: discrimination and comparing gender bias	PCA	Not applicable
9	a) Is an evaluation of the threshold ordering reported? (only applicable for polytomous items)	**1**	**1**	**1**	**1**	0	**1**	0	**1**	**1**
10	a) Is the proportion or N of participants with max and min total scores (i.e. extremes) reported?	0	0	0	**1**	0	0	0	0	**1**
	b) Is the proportion or N of participants that has a location lower than the lowest item threshold and the proportion or N of participants that has a location higher than the highest item threshold reported?	0	0	0	0	0	0	0	0	0
11	a) Is a Wright map or similar figure presented to illustrate sample-to-item targeting?	**1**	**1**	0	**1**	**1**	**1**	0	0	0
	b) Are item locations mean and SD reported?	0	0	0	0	0	0	0	0	0
	c) Are person locations mean and SD reported?	**1**	0	0	**1**	0	0	0	0	0
12	a) Are person reliability coefficient(s) reported?	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**
	b) Is a test information function figure reported?	0	0	0	0	0	0	0	0	0
	c) Is the proportion (%) of participants located at levels where test information meets	0	0	0	0	0	0	0	0	0

	the chosen reliability criterion reported?									
13	a) Are the item locations reported for each individual item?	0	0	0	0	1	**1**	0	**1**	**1**
	b) Is a transformation table provided that enables raw scores to be converted into linear measurement values?	0	0	0	**1**	0	0	0	0	0
	c) If a transformation table is provided, is measurement uncertainty (standard error) reported?	Not applicable	Not applicable	Not applicable	0	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable

Criteria that were reported are indicated in green and bold.
DIF = differential item functioning; LD = local dependency; MNSQ = mean square fit statistic; PCM = partial credit model; RSM = rating scale model; SD = standard deviation; SE = standard error; TIF = test information function; ZSTD = standardized fit statistic.

4. Discussion

This paper evaluated the reporting quality of nine psychometric studies that applied Rasch analysis to assess scales intended to measure well-being or QoL among older persons. Overall, the findings point to substantial variation in how key components of Rasch analysis are reported, with some elements frequently addressed and others consistently absent. Based on the criteria used in this quality assessment, only two of the included studies reported at least half of the criteria. Most studies reported sample size and demographics, tests of fit to the Rasch model, DIF, evaluation of threshold ordering and reliability, while none reported the method used to estimate parameters, participants with locations outside the threshold range, item locations, or test information function figures/values.

The aim of this paper was not to classify the psychometric studies as "good" or "bad," but rather to organize and present the psychometric properties that were reported. Consequently, such a structured overview informs the reader about the extent to which conclusions regarding a scale's measurement quality can be validly supported based on the available reporting. Taking the WHO-5, which meets the fewest number of criteria, as an example, Lucas-Carrasco et al (2012) state that the total score of the WHO-5 satisfies sufficient statistics and is both internally and externally valid. However, because only a limited number of Rasch reporting components were documented, the extent to which one can conclude that the total score met the requirements for sufficient statistics rests on Rasch-based evidence and remains unclear (Kreiner, 2025; Nielsen et al., 2025). This illustrates how limited reporting constrains the interpretability of claims about measurement quality. In contrast, the PWI is concluded to provide valid and reliable linear measures in older persons (Forjaz et al., 2012). The PWI study reported a larger number of Rasch components than most other studies, providing relatively more information on several key aspects of the analysis. Consequently, their conclusions can rest on a more transparent and substantiated foundation. Yet, an important question remains: just because something is reported, do the conclusions necessarily adhere to the recommendations for what constitutes a valid and reliable scale? The aim of this paper was not to provide such answers, but rather to emphasize that reporting is the first essential step. Based on these reports, more or less appropriate decisions can then be made regarding interpretation and action related to the fit statistics, which ultimately form a key component of the basis for making claims about a scale's measurement quality.

In our recent study, evaluating content across well-being and QoL scales psychometrically tested among older persons, we concluded that while there is some thematic consistency, notable heterogeneity persists (Preuter et al., 2025). Taken together, the observed content heterogeneity, along with the present findings of a lack of clearly articulated construct theories, clearly limit the valid use of scales for interchangeable measures of the same latent trait. In practical terms, this implies that interpreting results based on one or several scales requires critical scrutiny of their theoretical foundation, intended latent trait, and item content. Without alignment, comparisons across studies or pooled analyses may be misleading or fail to capture meaningful variation in respondents' well-being or QoL.

A transformation table, that is a table enabling raw scores to be converted into linear measurement values, provides an accessible tool that allows researchers to place their data on a common scale and thereby ensure valid comparisons through metrological traceability. However, if the data do not adequately fit the model and the measurement properties are of insufficient quality, providing such a table cannot be advisable. Consequently, the absence of transformation tables in most of the studies reviewed should not necessarily be regarded as a limitation. Furthermore, it is also important to emphasize that for a transformation table to be

valid for external use, the new sample must be sufficiently similar to the sample from which the table was derived.

Methodological considerations

As we found existing checklists neither fully comprehensive nor entirely applicable to the present study, we developed and employed a customized list of criteria. It is important to note that the few existing checklists are not firmly established; they have been used only sparingly, and the field of reporting quality in psychometrics warrants greater attention. The criteria used in this study were derived from three main sources: (1) previously published Rasch reporting checklists (Dabaghi et al., 2020; Johansson et al., 2023; Mallinson et al., 2022; Van de Winckel et al., 2022), (2) Rasch methodological textbooks and articles (Andrich & Marais, 2019; Christensen et al., 2013; Hobart & Cano, 2009; Tennant & Conaghan, 2007; Tesio, Caronni, Kumbhare, et al., 2023; Tesio, Caronni, Simone, et al., 2023), and (3) general principles of study quality and transparency, such as clear sample description, specification of model assumptions (e.g., von Elm et al., 2007). Moreover, there are some more general standards for psychometric reporting. For instance, the *Standards for Educational and Psychological Testing*, jointly produced by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (Joint Committee on the Standards for Educational and Psychological Testing, 2014), and the *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN; Mokkink et al., 2016). However, generic checklists do not adequately capture the critical indices specific to Rasch analysis. Therefore, we considered it most appropriate to use a checklist with some extensions. The fact that several criteria were not reported at all raises concerns about whether the checklist may have been ”too demanding”. That said, this paper does not aim to judge the included studies as having high or low reporting quality; instead, readers are encouraged to form their own interpretations of which criteria they find most valuable. While the authors are satisfied with how the checklist was applied in the present study and consider it potentially useful for assessing reporting quality in other studies employing Rasch analysis, it should be acknowledged that others may wish to refine it further for their own purposes, or that it could evolve into a broader collaborative effort where a larger group of researchers reach consensus on the criteria.

As outlined in the Methods section, the criterion regarding construct theory required consideration to ensure consistency among raters. One possible explanation is that 'construct theory' is not clearly defined in the literature. In this study, we adopted a minimum threshold for what qualifies as a construct theory. However, we acknowledge that a more comprehensive approach – such as that proposed by Wilson (2005) – is preferable. His framework includes the development of a construct map, which defines the construct and outlines what it means to be located at various points along the continuum (Wilson, 2005), thereby supporting stronger construct validity.

5. Conclusion

There are deficiencies in the reporting of measurement properties in studies that apply Rasch analysis to scales of well-being and QoL for older persons. Thus, the results reveal clear room for improvement in reporting. Although more comprehensive and transparent reporting alone cannot ensure that scales have satisfactory measurement properties, it is an essential prerequisite for evidence-based conclusions on the scales’ measurement quality. By strengthening reporting practices, the field can move toward more consistent and substantiated

claims regarding measurement quality, thereby enabling better-informed decisions in both research and practice.

Funding

This work was supported by the Kamprad Family Foundation under Grant 20233141.

Disclosure statement

The authors report there is no competing interest to declare.

Preregistration

The study was preregistered at PROSPERO International prospective register of systematic reviews, CRD42024531400.

Authors' contributions

The manuscript was designed by J.M. and M.P. M.P. and J.M. conducted the literature search. M.P., M-L.M. and J.M. screened articles and M.P., M-L.M. J.M. and K.Å evaluated the reporting quality. J.M. drafted the manuscript. All authors edited and revised the manuscript. All authors read and approved the final manuscript.

How to Cite

Preuter, M., Möllerberg, M.-L., Årestedt, K., & Melin, J. (2026). Transparency in psychometric reporting: A review of scales for well-being and quality of life in older persons employing Rasch analysis. *Educational Methods & Psychometrics, 4*: 30. https://doi.org/10.65301/emp.2026.264

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(Supplement), I–7. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health Sciences*. Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Barbic, S. P., Cano, S. J., & Mathias, S. (2018). The problem of patient-centred outcome measurement in psychiatry: Why metrology hasn't mattered and why it should. *Journal of Physics: Conference Series*, *1044*, 012069. https://doi.org/10.1088/1742-6596/1044/1/012069

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (Fourth edition). Routledge/Taylor & Francis Group.

Cano, S. J., Pendrill, L. R., Melin, J., & Fisher, W. P. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Chachamovich, E., Fleck, M. P., Trentini, C., & Power, M. (2008). Brazilian WHOQOL-OLD module version: A Rasch analysis of a new instrument. *Revista de Saúde Pública*, *42*(2), 308–316. https://doi.org/10.1590/S0034-89102008000200017

Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds). (2013). *Rasch models in health*. ISTE ; John Wiley & Sons.

Conrad, I., Matschinger, H., Riedel-Heller, S., Von Gottberg, C., & Kilian, R. (2014). The psychometric properties of the German version of the WHOQOL-OLD in the German population aged 60 and older. *Health and Quality of Life Outcomes*, *12*(1), 105. https://doi.org/10.1186/s12955-014-0105-4

Dabaghi, S., Esmaielzadeh, F., & Rohani, C. (2020). Application of Rasch analysis for development and psychometric properties of adolescents' quality of life instruments: A systematic review. *Adolescent Health, Medicine and Therapeutics*, *11*, 173–197. https://doi.org/10.2147/AHMT.S265413

Europäische Kommission (Ed.). (2019). *Ageing Europe: Looking at the lives of older people in the EU*. Publications Office of the European Union.

Fang, J., Power, M., Lin, Y., Zhang, J., Hao, Y., & Chatterji, S. (2012). Development of short versions for the WHOQOL-OLD module. *The Gerontologist*, *52*(1), 66–78. https://doi.org/10.1093/geront/gnr085

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). *Construct validity and the validity of replication studies: A systematic review*. PsyArXiv. https://doi.org/10.31234/osf.io/369qj

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920952393

Forjaz, M. J., Ayala, A., Rodriguez-Blazquez, C., Prieto-Flores, M.-E., Fernandez-Mayoralas, G., Rojo-Perez, F., & Martinez-Martin, P. (2012). Rasch analysis of the International Wellbeing Index in older adults. *International Psychogeriatrics*, *24*(2), 324–332. https://doi.org/10.1017/S104161021100158X

Gondodiputro, S., Wiwaha, G., Lionthina, M., & Sunjaya, D. K. (2021). Reliability and validity of the Indonesian version of the World Health Organization quality of life-old (WHOQOL-OLD): A Rasch modeling. *Medical Journal of Indonesia*, *30*(2), 143–151. https://doi.org/10.13181/mji.oa.215065

Halvorsrud, L., & Kalfoss, M. (2007). The conceptualization and measurement of quality of life in older adults: A review of empirical studies published during 1994–2006. *European Journal of Ageing*, *4*(4), 229–246. https://doi.org/10.1007/s10433-007-0063-3

Hanisch, R. J., Gilmore, I. S., & Plant, A. L. (2019). Improving reproducibility in research: The role of measurement science. *Journal of Research of the National Institute of Standards and Technology*, *124*, 1–13. https://doi.org/10.6028/jres.124.024

Hobart, J. C., & Cano, S. J. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment*, *13*(12). https://doi.org/10.3310/hta13120

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and reliable? Basic and expanded recommendations for psychometric reporting and quality assessment.* OSF Preprints. https://doi.org/10.31219/osf.io/3htzc

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Kreiner, S. (2025). On specific objectivity and measurement by Rasch models: A statistical viewpoint. *Educational Methods & Psychometrics*, *3*. https://doi.org/10.61186/emp.2025.7

Liang, W.-M., Chang, C.-H., Yeh, Y.-C., Shy, H.-Y., Chen, H.-W., & Lin, M.-R. (2009). Psychometric evaluation of the WHOQOL-BREF in community-dwelling older people in Taiwan using Rasch analysis. *Quality of Life Research*, *18*(5), 605–618. https://doi.org/10.1007/s11136-009-9471-5

Linton, M.-J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open*, *6*(7), e010641. https://doi.org/10.1136/bmjopen-2015-010641

Lucas-Carrasco, R., Allerup, P., & Bech, P. (2012). The validity of the WHO-5 as an early screening for apathy in an elderly population. *Current Gerontology and Geriatrics Research*, *2012*, e171857. https://doi.org/10.1155/2012/171857

Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Van de Winckel, A. (2022). Rasch reporting guideline for rehabilitation research (RULER): The RULER statement. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013

Mokkink, L. B., Prinsen, C. A. C., Bouter, L. M., Vet, H. C. W. de, & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Brazilian Journal of Physical Therapy*, *20*(2), 105–113. https://doi.org/10.1590/bjpt-rbf.2014.0143

Nielsen, T., Fellinghauer, C., Strobl, C., Kronthaler, D., & Kreiner, S. (2025). Statistical anxiety and attitudes towards statistics in psychology students: A measurement comparison study of the German and Danish language versions of the HFS-R. *Educational Methods and Psychometrics*, *3*, 1–43. https://doi.org/10.61186/emp.2025.6

Nugent, W. R. (2011). The (non)comparability of the correlation effect size across different measurement procedures: A challenge to meta-analysis as a tool for identifying "evidence based practices". *Journal of Evidence-Based Social Work*, *8*(3), 253–274. https://doi.org/10.1080/15433710903346574

Oluboyede, Y., & Smith, A. B. (2013). Evidence for a unidimensional 15-item version of the CASP-19 using a Rasch model approach. *Quality of Life Research*, *22*(9), 2429–2433. https://doi.org/10.1007/s11136-013-0367-z

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35. https://doi.org/10.1080/19315775.2014.11721702

Power, M., Quinn, K., Schmidt, S., & WHOQOL-OLD Group. (2005). Development of the WHOQOL-Old Module. *Quality of Life Research*, *14*(10), 2197–2214. https://doi.org/10.1007/s11136-005-7380-9

Preuter, M., Möllerberg, M.-L., Årestedt, K., & Melin, J. (2024). *Measurement properties, construct definition and availability of existing questionnaires for measuring older persons' well-being: A systematic review.*. (No. CRD42024531400). Prospero. https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=531400

Preuter, M., Möllerberg, M.-L., Årestedt, K., & Melin, J. (2025). *What's in the scales? A systematic review of content overlap in well-being and quality of life scales for older persons* (2h5xj_v1). SocArXiv. https://doi.org/10.31235/osf.io/2h5xj_v1

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Salzberger, T. (2012). Reporting a Rasch analysis. In *Rasch Models in Health* (pp. 347–362). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118574454.ch19

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00536

Tennant, A., & 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*(8), 1358–1362. https://doi.org/10.1002/art.23108

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation*, 1–13. https://doi.org/10.1080/09638288.2023.2169771

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 1–14. https://doi.org/10.1080/09638288.2023.2169772

United Nations. (2024). *World population ageing 2023: Challenges and opportunities of population ageing in the least developed countries* (1st ed). United Nations Research Institute for Social Development.

Van de Winckel, A., Kozlowski, A. J., Johnston, M. V., Weaver, J., Grampurohit, N., Terhorst, L., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Mallinson, T. (2022). Reporting guideline for RULER: Rasch reporting guideline for rehabilitation research: Explanation and elaboration. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1487–1498. https://doi.org/10.1016/j.apmr.2022.03.019

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ : British Medical Journal*, *335*(7624), 806–808. https://doi.org/10.1136/bmj.39335.541782.AD

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Manuscript Received: 17 DEC 2025
Final Version Received: 23 MAR 2026
Published Online Date: 10 APR 2026

Appendix 1

#	Criterion	Chachamovich et al 2008	Conrad et al 2014	Fang et al 2012	Forjaz et al 2012	Gondodipurto et al 2021	Liang et al 2009	Lucas-Carrasco et al 2012	Oluboyede & Smith 2013	Power et al 2005	Number 1 = Agree	Number 0 = Disagree	% Agree
1	a) Is a construct theory presented and/or referenced?	0	0	0	0	0	0	0	0	0	0	9	0 %
	b) Was the construct theory tested?	-	-	-	-	-	-	-	-	-			
2	a) Are the sample size and the demographic characteristics of the participants reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
	b) Is the sample size justified?	**1**	**1**	**1**	**1**	0	0	0	**1**	**1**	6	3	67 %
	c) If the sample size may be considered too small or too large in relation to the statisti-cal analyses conducted, are the implications discussed?	**1**	0	**1**	0	**1**	0	0	0	0	3	6	33 %
	d) Are missing data reported?	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	7	2	78 %
	e) In case of missing data, was how it was handled reported?	**1**	**1**	0	**1**	**1**	0	0	0	0	4	5	44 %
3	a) Is the software for Rasch model analysis, including software version, stated?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	8	1	89 %
	b) Which software was used?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
	c) If R was used, is the package(-s) stated?	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	0	7	2	78 %
	d) If R was used, which package(-s) was/were used?	**1**	**1**	**1**	**1**	**1**	1	**1**	**1**	**1**	9	0	100 %
	e) Is the Rasch model used stated? (E.g., RSM or PCM)	**1**	**1**	**1**	**1**	0	**1**	0	**1**	**1**	7	2	78 %

	f) Which parameter estimation method was used? (E.g., conditional or unconditional estimation)?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
4	a) Are tests of item fit to the Rasch model reported?	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	8	1	89 %
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)	0	**1**	0	**1**	**1**	0	**1**	**1**	**1**	6	3	67 %
5	a) Are tests of person fit to the Rasch model reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	7	2	78 %
	b) Which fit indices are reported? (E.g., Residuals, ZSTD, MNSQ)	**1**	**1**	**1**	0	0	**1**	**1**	**1**	0	6	3	67 %
6	a) Are tests of DIF reported?	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0	6	3	67 %
	b) Which tests of DIF are reported?	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	8	1	89 %
7	a) Is an evaluation of LD reported?	**1**	0	**1**	**1**	**1**	**1**	0	**1**	**1**	7	2	78 %
	b) Which test/-s of LD were used?	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	7	2	78 %
8	a) Is an evaluation of dimensionality reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	8	1	89 %
	b) Which test/-s of dimensionality were used?	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	0	7	2	78 %
9	Is an evaluation of the threshold ordering reported? (only applicable for polytomous items)	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
10	a) Is the proportion or N of participants with max and min total scores (i.e. extremes) reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	7	2	78 %
	b) Is the proportion or N of participants that has a location lower than the lowest item threshold and the	**1**	**1**	**1**	**1**	0	0	**1**	0	1	6	3	67 %

	proportion or N of participants that has a location higher than the highest item threshold reported?												
11	a) Is a Wright map or similar figure presented to illustrate sample-to-item targeting?	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	8	1	89 %
	b) Are item locations mean and SD reported?	**1**	0	**1**	**1**	**1**	**1**	0	**1**	**1**	7	2	78 %
	c) Are person locations mean and SD reported?	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	7	2	78 %
12	a) Are person reliability coefficient(s) reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	8	1	89 %
	b) Is a test information function figure reported?	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	0	7	2	78 %
	c) Is the proportion (%) of participants located at levels where test information meets the chosen reliability criterion reported?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
13	a) Are the item locations reported for each individual item?	**1**	**1**	**1**	**1**	0	0	**1**	0	**1**	6	3	67 %
	b) Is a transformation table provided that enables raw scores to be converted into linear measurement values?	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	9	0	100 %
	c) If a transformation table is provided, is measurement uncertainty (standard error) reported?	0	0	0	**1**	0	0	0	0	0	1	8	11 %

Criterion 1b was added after consensus discussions and subsequent clarifications; accordingly, no agreement was assessed.
Criteria with agreement =1, indicated in green and bold. 100% agreement is indicated in green and less than 50% agreement in orange.