An off-target scale limits the utility of SWEMWBS as a measure of well-being in public health surveys

Abstract

Objectives: To assess the utility and measurement properties for the well-being scale *Short Warwick-Edinburgh Mental Well-Being Scale* (SWEMWBS) in a Swedish general population survey.

Study design: A cross-sectional survey-study

Methods: Data were retrieved from the 2018 public health survey in Stockholm County, containing a random sample of 22 856 persons stratified to be representative for the municipalities and districts within the region. The data were analysed according to Rasch Measurement Theory.

Results: Person attribute values are positively skewed (mean 2.32, SD 1.85), with wide gaps in the item threshold attribute values. Overall item fit statistics were acceptable and person measurement separation reliability was 0.83, indicating three statistically distinct ranges in the estimated well-being values.

Conclusion: While the SWEMWBS items indicated acceptable fit to the Rasch measurement model, targeting of items to sample is skewed toward lower levels of well-being and there is a ceiling effect. Thus, we suggest a careful reconsideration of SWEMWBS as a tool for use in general public health surveys, especially for assessing change over time and group differences, as there are large measurement uncertainties for the majority of cases when the population as a whole is sampled. We encourage revisions applying a coherent and comprehensive construct theory for well-being to fill the gaps in the upper end of the SWEMWBS scales' item thresholds. The addition of more challenging items would improve targeting for population-based surveys, increase reliability, and provide more actionable information that could be useful in improving individuals' well-being.

Keywords: well-being, Rasch, measurement, targeting, public health

Authors

Jeanette Melin (PhD, RPT) ^{a,b}, Andreas Lundin ^c, Magnus Johansson (PhD) ^a

^a RISE Research Institutes of Sweden, Safety and Transport, Measurement Science and Technology.

^b The Gothenburg Centre for Person Centred Care (GPCC), University of Gothenburg, Sahlgrenska Academy.

^c Global Public Health, Karolinska Institutet,

Corresponding author

Jeanette Melin

Sven Hultins plats 5, 412 58 Gothenburg, Sweden

+4610-516 66 39

jeanette.melin@ri.se

Introduction

Worldwide and across disciplines, there is a growing interest to monitor and follow the development of population mental health and well-being¹. One of the United Nations sustainability goals (SDG 3.4) states: *By 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being*².

To quote VanderWeele³: *In order to make more substantial improvements to the well-being of our world, it is critical that we have the data to measure it*. However, data is not the sole criterion for success in measurement. A scientific approach to measurement requires empirical data based on sound theories as well as the proper treatment of collected data. For nearly 100 years, since the work of Thurstone in the 1920s, criteria for calibrating instruments measuring in theoretically defined unit quantities have informed research and practice across dozens of fields in psychology and the social sciences. This includes the work by Danish mathematician Georg Rasch in the 1960s, who based his models on the same underlying principles as physical measurements, developed the Rasch Measurement Theory (RMT) to enable separate measures of person and item attributes scaled on the conjoint interval logit scale⁴.

One example of a widely used well-being scale is the *Short Warwick-Edinburgh Mental Well-Being Scale* (SWEMWBS, stemming from the *Warwick-Edinburgh Mental Well-Being Scale* (WEMWBS)⁵) that in previous work has been evaluated to basic measurement criterion according to RMT in public health surveys⁶. The SWEMWBS is used in several national population surveys to estimate well-being and how it develops over time^{7,8}. The SWEMWBS is intended to focus on the positive aspects of well-being. Thus, it should be free of ceiling effects in population samples⁵. The SWEMWBS is included in the Swedish biannual public health survey *Hälsa på lika villkor*⁹, but its measurement properties have not previously been evaluated for this population. The purpose of this study was to investigate the utility and measurement properties of the SWEMWBS according to RMT in a Swedish general population sample.

Methods

Participants and data collection

We retrieved data from the 2018 annual public health survey *Hälsa på lika villkor*⁹ in Region Stockholm, which contains a stratified random sample of 22 856 persons (strata municipalities and districts within the region). The response rate was 39.2 %. The median age was 54 years (range 16-84), 55% (n=12571) were women and 45% (n=10285) were men, and 81% (n=18513) were born in Sweden and 19% (n=4343) were born in other countries.

Measurement

The SWEMWBS⁶ is a 7-item short form of the longer 14-item *Warwick-Edinburgh Mental Well-Being Scale* (WEMWBS)⁵. Item generation for the WEMWBS was based on an expert panel considering literature in the field, qualitative focus groups, and psychometric testing by comparison with a related scale, the *Affectometer 2*. The WEMWBS is purported to cover both hedonic and eudemonic (or psychological) aspects of mental health⁵. Item reduction for SWEMWBS was based on RMT analysis to ensure a unidimensional scale⁶; items mainly representing psychological and eudemonic well-being and less related to hedonic well-being or affect.

The respondents are asked to rate each of the seven items in relation to how often they experienced each statement in the past two weeks. The original response alternatives *none of the time, rarely, some of the time, often* and *all of the time* were in the present study changed to *never, seldom, sometimes, often* and *always*, which are scored 0 through 4. All items use positive wording.

Data analysis

In order to assess the measurement properties of SWEMWBS the Rasch Measurement Theory (RMT) was applied, as “*it is not simply a mathematical or statistical approach, but instead a specifically metrological approach to human based measurement*”¹⁰. RMT was developed based on the same underlying principles as physical measurements, i.e. separate measures of person and item attributes

scaled on the conjoint interval logit scale⁴. In the simplest case with dichotomous responses, it is a logistic regression function:

$$\log \left(\frac{P_{success,i,j}}{1-P_{success,i,j}} \right) = \theta_i - \delta_j \quad (1)$$

where θ is the person, i is the person attribute value and δ is the item, and j is the item attribute value. This model enables estimates of the differences between the person and task attributes. This dichotomous model has subsequently been expanded to polytomous RMT models by incorporating threshold parameters^{11,12}, which we applied to SWEMWBS data, as the tool comprises five response categories. The analyses were conducted using the Rasch Unidimensional Measurement Model 2030 (RUMM) software.

The analysis was structured in line with recommendations by Hobart & Cano¹³. Initially, an examination of the relative distributions of the item threshold δ attribute values to the person attribute values was assessed. In order to understand the implications of the person measures and how those should be used, item threshold-person distributions, Person Separation Index (PSI), and person fit residuals were examined. Specifically, a) the mean person attribute value indicates whether the sample is off-centre relative to the items, b) the PSI is a reliability indicator where 0 implies all error/noise, 1 implies no error, and the coefficient should be over 0.8 (i.e. corresponding to a person separation index of 2.0, and the definition of three statistically different strata¹⁴), and c) the person fit residual should lie within -2.5 to +2.5. In addition to those aspects, individual person measurement uncertainties were assessed for the required change in well-being to become detectable. We applied $k=2$, which approximately corresponds to a 95% confidence interval.

In order to understand the item properties, individually and together, response categories were assessed to see if they worked as intended, and items were studied to see if they exhibited a clinically logical order. Response categories and items were assessed in terms of model fit residuals,

chi-square, item characteristic curve (ICC), residual correlation, unidimensionality, and Differential Item Functioning (DIF) statistics, specifically:

- a) Category thresholds should be sequentially ordered and as the trait level is increasing, the probability of a higher response also increases and be sequentially ordered¹⁵;
- b) Individual item fit residuals should lie between -2.5 and +2.5; the chi-square values should not be statistically significant after applying Bonferroni correction; and the dots of the observed class intervals should follow the expected ICC to support good fit to the model;
- c) Construct definition across samples was examined by comparing the item locations obtained in this study with those obtained in two previously published RMT articles analysing SWEMWBS data^{16,17};
- d) Item residual correlations should not exceed 0.20 above the average correlations¹⁸ to support local independence (potential indication of multidimensionality and/or redundancy due to too similarly phrased items);
- e) Person attribute values estimated from two subsets of items (derived based on item loadings in the first factor in the principal component analysis of the residuals) were compared by an independent t-test where the percentage of persons outside the range of -1.96 to 1.96 should not exceed 5 % to support unidimensionality¹⁹;
- f) For DIF, both main effects and interaction effects were assessed for between gender (men/women) and age (≤ 25 ; 26-45; 46-65; ≥ 66 years) and should be statistically non-significant after Bonferroni correction²⁰. When significant, DIF was followed by stepwise item splits and repeated analyses.

To avoid misfit inflated by the large sample size, item chi-squares and DIF were assessed using the whole sample as well as with adjusted sample sizes for $n=200$ and $n=500$ ²¹.

Results

As shown in Figure 1 and Table 1, item threshold attribute values are covered by the person attribute values, but not the other way around. Especially, person θ attribute values are positively skewed and there are gaps (indicated by the blue arrows in Figure 1) in the item threshold attribute values. The PSI was 0.83, indicating three statistically distinct ranges in the estimated well-being values.

Figure 1. Person-item threshold histograms SWEBWBS. Pink upper bars show person attribute values distributions and blue lower bars item threshold attribute values distributions scaled on a conjoint logit scale. Low values indicate lower well-being and high values indicate higher well-being. The green curve show where most information about the persons are provided and are inverse functions of the measurement standard errors (SE). Horizontal error bars correspond to 2 SE. Blue arrows indicate gaps in the item threshold attribute values compared to the person attribute values.

Table 1. Summary of measurement properties for SWEMWBS. Total person $n=22\,856$, extremes $n=1574$

Measurement property	Values
Person attribute values mean (SD)	2.32 (1.85)
Person attribute mean 2SE	1.47
Person attribute values range with extremes	-5.17 to 6.20
Person attribute values range without extremes	-4.33 to 5.28
Person attribute PSI with extremes	0.83
Person attribute PSI without extremes	0.85
Item attribute values range	-0.57 to 0.53
Item thresholds attribute values range	- 3.37 to 4.51

In total, 3011 (14.2%) persons had fit residuals ≤ -2.5 , i.e. having low variation in their response pattern and 103 (0.5%) persons had fit residuals ≥ 2.5 , i.e. having irrational response pattern. In total, 1574 (6.9%) persons responded *all of the time* (i.e. highest response option giving extremes) on all seven items.

None of the items had disordered thresholds, indicating that the response options worked as intended, i.e. progressively more often used as overall wellbeing increased (Figure 2). Table 2 reports item mean attribute values, associated measurement uncertainties, fit residuals, and ChiSq. All except items 2 (*I've been feeling useful*) and 3 (*I've been feeling relaxed*) had fit residuals outside the desired range ± 2.5 and significant ChiSq. However, by assessing adjusted sample sizes and visual inspection of the dots of the class intervals in relation to the ICC, it was evident that the misfit (fit residuals and ChiSq) was inflated by the large sample size.

Figure 2. Item threshold map for SWEMWBS where 0 = *never*, 1 = *seldom*, 2 = *sometimes*, 3 = *often*, and 4 = *always*.

Table 2. Item fit statistics for SWEMWBS.

Item		Location	2SE	Fit Residuals	ChiSq	n=21158 Prob	n=200 Prob	n=500 Prob
7	I've been able to make up my own mind about things	-0.57	0.02	4.95	270.20	0.00	0.98	0.70
5	I've been thinking clearly	-0.47	0.03	-17.37	274.10	0.00	0.98	0.69
4	I've been dealing with problems well	-0.02	0.03	-17.47	400.49	0.00	0.92	0.40
6	I've been feeling close to other people	0.14	0.02	3.96	109.38	0.00	1.00	0.98
1	I've been feeling optimistic about the future	0.19	0.02	-7.94	59.80	0.00	1.00	1.00
2	I've been feeling useful	0.21	0.02	-0.59	30.14	0.00	1.00	1.00
3	I've been feeling relaxed	0.53	0.02	-1.78	187.52	0.00	0.99	0.88

Bolded numbers indicate misfit; Fit Residuals ± 2.5 or significant ChiSq.

As shown in Table 3, mean item attributes hierarchy slightly deviated from the two previous Rasch studies on SWEMWBS; item 2 (*I've been feeling useful*) had a relatively higher item attribute value in the present Swedish cohort compared to a UK cohort ¹⁶ and an Australian cohort ¹⁷.

Table 3. Item δ attributes values for the Swedish cohort compared with a UK cohort ¹⁶ and an Australian cohort ¹⁷. The UK study included the same 7 items while the Australian study included the same 7 items as well as 3 additional items from WEBWBS.

Item		Region Stockholm 2018		Bartram et al 2013		Houghthon et al 2017	
		Location	2SE	Location	2SE	Location	2SE
7	I've been able to make up my own mind about things	-0.57	0.02	-1.07	0.08	-0.85	0.07
5	I've been thinking clearly	-0.47	0.03	-0.66	0.08	-0.48	0.07
4	I've been dealing with problems well	-0.02	0.03	0.08	0.08	-0.19	0.07
6	I've been feeling close to other people	0.14	0.02	0.11	0.06	-0.09	0.07
1	I've been feeling optimistic about the future	0.19	0.02	0.48	0.06	0.47	0.06
2	I've been feeling useful	0.21	0.02	0.00	0.08	-0.14	0.06
3	I've been feeling relaxed	0.53	0.02	1.06	0.06	0.69	0.06

The relative cut-off, i.e. 0.2 above the average correlation of -0.16, for item residuals was 0.04. A closer inspection of the residual matrix showed that one item pair correlated above the cut-off; item 4 (*I've been dealing with problems well*) and item 5 (*I've been thinking clearly*) residuals correlated to 0.13. Thus, items were overall locally independent from each other.

Person attribute values derived from the positively and negatively loading items in the first factor in the principal component analysis of the fit residuals were compared by an independent t-test. 7% of the person attribute values were outside the desirable ± 1.96 , i.e. slightly above the criterion of 5% but within the confidence interval, thus support unidimensionality.

Uniform DIF was present between men and women as well as between age groups. It was significant for all comparisons on all items except item 4 (*I've been dealing with problems well*) among different age groups. However, the number of significant DIF were reduced when assessing adjusted sample

sizes (n=200 no significant DIF, n=500 item 3 for gender and item 2 and 3 for age) and item splits (i.e. item divided for person factors) could have resolved the DIF without significantly impacting person attribute value. Person attribute values with and without item splits showed correlation coefficients above 0.99, supporting little overall effect of these items being perceived differently between the groups.

Discussion

This study has demonstrated an acceptable fit of SWEMWBS to the basic Rasch measurement model. However, the findings raise several concerns about the utility of SWEMWBS as a measure of well-being in public health surveys. As shown in Figure 1, the best measurement region is indicated by the green information curve that overlaps with the low well-being end of the spectrum. Moreover, the highly positive skewed person attribute values imply large measurement uncertainties for most of the population. As such, this requires large changes in well-being to be detectable and not biased due to measurement noise. Moreover, there were many extremes on the positive side, with 1574 (6.9%) persons responding *always* (i.e. highest response option) on all seven items and it is obvious that we know very little about their well-being, and it would be impossible for them to show improvement on the scale. The SWEMWBS has previously been contended as being free of ceiling effects in general population samples⁵ but our analysis does not support this claim. We suggest careful consideration in the use and interpretation of SWEMWBS in public health surveys, especially for assessing change, due to the large mistargeting between sample and items and the ceiling effect observed. Although, it might be used to screen for low well-being or with groups with lower levels of well-being.

Early work on SWEMWBS took a step in the right direction to a scientific approach for measurement when addressing the measurement properties according to RMT⁶ rather than classical test theory⁵. However, further development to a scientific measurement approach has to a great extent been absent for the SWEMWBS as well as well-being measures in general. There are current

recommendations for measuring well-being in public surveys claimed to be based on the state of knowledge²², that lack a scientific measurement approach. The recommendations include having four or fewer generic questions. Suggestions of few items are probably related to the lack of insight into the issues with measurement uncertainties since ordinal responses are typically statistically treated as having interval scale properties. However, measurement uncertainties reflect the lack of exact knowledge of the value of the measurand (e.g. person attribute values and item attribute values), and characterizes the dispersion of the values that reasonably could be attributed to it²³. Few items and/or mistargeted items imply low measurement precision²⁴. More items mean that we get more information about each respondent and especially if the number of well-targeted items increases the measurement uncertainties for estimates of person attribute values are reduced. Although the measurement uncertainties will always be greater at the upper (and lower) end of a scale as the model anticipates infinity in both directions¹⁴, limiting reliable decisions on general well-being in public health surveys for people with moderate to high well-being.

In order to accommodate the mistargeting and to further advance possibilities for an optimal well-being scale to be used in public health surveys, we would first argue for addressing the issue of a coherent ordinal construct theory²⁵ for well-being. First, drawing from Figure 1, the person-item thresholds maps the well-being terrain by showing the person's overall well-being relative to the item properties and should be used to extend the understanding of what characterizes low, medium, and high levels of well-being. Second, as there is a gap in item thresholds as well as lacking items/item thresholds at the upper end of the scale, we suggest that the SWEMWBS is extended with more challenging items to improve its targeting for population-based surveys. Adding items with higher levels of difficulty should ideally be driven by the construct theory and the target population's understanding of what characterizes higher levels of well-being.

In addition to our findings, the previous studies on WEMWBS and SWEMWBS could also be used to extend the construct theory. When comparing the result in this study with those, there are at last

two critical issues that need further cross-country evaluation. First, as shown in Table 3, item 2 (*I've been feeling useful*) had a relatively higher item attribute value in the present Swedish cohort compared to the UK and Australian cohorts^{16,17}. It raises questions whether feeling useful has a different meaning in different cultures and/or there are translation issues. Second, the Swedish cohort has more extremes and shows greater mistargeting between person and items compared to the other studies. This could be explained by different sample characteristics, cultural differences or a different response scale.

Apart from the suggestion of adding (or replacing) items, this also needs to be considered in the context of the fact that the SWEMWBS might represent a few items out of hundreds of items included in public health surveys. A concern apparent in our study of SWEMWBS, but less discussed in the literature, is the large number of persons (n=3011, 14.2%) having low variation in their response pattern. This raises questions such as; how valid are person responses when they tick the same response option for all seven items? Did the person even read the item or is this a matter of acquiescence, i.e. agreement regardless of item content? Or is this a consequence of the response burden in lengthy public health surveys?

A major strength of this study is that the data used stem from a public health survey with respondents stratified to be representative of the municipalities and districts within the Stockholm region. However, it is not representative of the whole Swedish population. Moreover, at this stage, it was not possible to control for other person characteristics such as health, family situation, living conditions, employment, or income, but it is required in further research.

Conclusions

Despite an acceptable fit of SWEMWBS to the basic Rasch model, due to the large mistargeting between sample and items and the ceiling effects, there are limitations with the utility of SWEMWBS as a measure of well-being in public health surveys. Thus, we suggest careful consideration if SWEMWBS is used in general public health surveys, especially for assessing change

over time and group differences as there are large measurement uncertainties for the majority of respondents. We encourage addressing a coherent construct theory for well-being as well as filling the gaps and upper end of the SWEMWBS scales' item thresholds with more challenging items to improve its targeting for population-based surveys.

Acknowledgements

The authors would like to give a particular thanks to Fredrik Lindencrona Region Stockholm/ Swedish Association of Local Authorities and Regions for encouraging this work and prof. William P. Fisher, Jr., Living Capital Metrics LLC, Sausalito CA USA/BEAR Center, Grad. School Education, UC Berkeley/RISE, Research Institutes of Sweden, Gothenburg, Sweden for valuable and supportive discussion about measurement sciences.

Ethical approval

Ethical approval was obtained from the Swedish Ethical Review Authority, Dnr 2021-01646.

Funding

Part of this work was supported by the innovation partnership between Region Stockholm and RISE, and part by the RISE platform *Centre for Categorically Based Measures*.

References

1. OECD Better Life Initiative. Measuring well-being and progress. Paris: OECD; 2020 [cited 2021 Jan 15]. (How's Life?). Available from: <http://www.oecd.org/statistics/better-life-initiative.htm>
2. United Nations. The 17 Goals. Department of Economic and Social Affairs - Sustainable Development. [cited 2020 Dec 21]. Available from: <https://sdgs.un.org/goals>
3. VanderWeele TJ. The Global Flourishing Study. Psychology Today. 2020 [cited 2020 Dec 21]. Available from: <https://www.psychologytoday.com/intl/blog/human-flourishing/202005/the-global-flourishing-study>
4. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks Paedagogiske Institut; 1960. 184 p.
5. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes. 2007 Nov 27;5(1):63.
6. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. 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 Qual Life Outcomes. 2009 Dec;7(1):15.
7. Danish Regions. The Danish National Health Survey. 2017 [cited 2021 Jan 15]. Available from: <https://www.danishhealthdata.com/find-health-data/Danskernes-sundhed>
8. NHS Digital. Health Survey for England. 2019 [cited 2021 Jan 15]. Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england>

9. Public Health Agency of Sweden. Nationella folkhälsoenkäten – Hälsa på lika villkor. 2020 [cited 2021 Jan 15]. Available from: <http://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/om-vara-datainsamlingar/nationella-folkhalsoenkaten/>
10. Pendrill L. Man as a Measurement Instrument. NCSLI Meas. 2014 Dec;9(4):24–35.
11. Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978 Dec 1;43(4):561–73.
12. Wright BD, Masters GN. Rating scale analysis. Chicago: Mesa Press; 1982. 206 p.
13. Hobart JC, Cano SJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess [Internet]. 2009 Feb [cited 2020 Jan 13];13(12). Available from: <https://www.journalslibrary.nihr.ac.uk/hta/hta13120/>
14. Standard Errors and Reliabilities: Rasch and Raw Score. [cited 2020 Dec 21]. Available from: <https://www.rasch.org/rmt/rmt204f.htm>
15. Tennant A, Conaghan PG. 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 Res. 2007 Dec 15;57(8):1358–62.
16. Bartram DJ, Sinclair JM, Baldwin DS. Further validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in the UK veterinary profession: Rasch analysis. Qual Life Res. 2013 Mar 1;22(2):379–91.
17. Houghton S, Wood L, Marais I, Rosenberg M, Ferguson R, Pettigrew S. Positive Mental Well-Being: A Validation of a Rasch-Derived Version of the Warwick-Edinburgh Mental Well-Being Scale. Assessment. 2017 Apr;24(3):371–86.

18. Marais I, Andrich D. Formalising dimension and response violations of local independence in the unidimensional Rasch model. :27.
19. Smith RM. A comparison of methods for determining dimensionality in Rasch measurement. *Struct Equ Model Multidiscip J*. 1996 Jan 1;3(1):25–40.
20. Hagquist C. Explaining differential item functioning focusing on the crucial role of external information – an example from the measurement of adolescent mental health. *BMC Med Res Methodol*. 2019 Sep 5;19(1):185.
21. Hagell P, Westergren A. Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement. *J Appl Meas*. 2016;17(4):416–31.
22. VanderWeele TJ, Trudel-Fitzgerald C, Allin P, Farrelly C, Fletcher G, Frederick DE, et al. Current recommendations on the selection of measures for well-being. *Prev Med*. 2020 Apr;133:106004.
23. JCGM 200:2012. International vocabulary of metrology – Basic and general concepts and associated terms (VIM). [cited 2020 Dec 18]. Available from:
https://www.bipm.org/utis/common/documents/jcgm/JCGM_200_2012.pdf
24. Sarstedt M, Diamantopoulos A, Salzberger T. Should we use single items? Better not. *J Bus Res*. 2016 Aug;69(8):3199–203.
25. Wilson M. Constructing measures: an item response modeling approach. Mahwah, N.J: Lawrence Erlbaum Associates; 2005. 228 p.