Assessment
2017, Vol. 24(3) 371–386

DOI: 10.1177/1073191115609995
journals.sagepub.com/home/asm
SAGE

Positive Mental Well-Being: A Validation of a Rasch-Derived Version of the Warwick-Edinburgh Mental Well-Being Scale

Stephen Houghton[1,2], Lisa Wood[1], Ida Marais[1], Michael Rosenberg[1], Renee Ferguson[1], and Simone Pettigrew[1,3]

Abstract

This study presents a Rasch-derived short form of the Warwick-Edinburgh Mental Well-Being Scale for use as a screening tool in the general population. Data from 2,005 18- to 69-year-olds revealed problematic discrimination at specific thresholds. Estimation of model fit also deviated from Rasch model expectations. Following deletion of 4 items, the 10 remaining items indicated the data fitted the model. No items showed differential item functioning, thereby making comparisons of overall positive mental well-being for the different age, gender, and income groups valid and accurate. Cronbach's alpha and Rasch Person Separation Index indicated a strong degree of reliability. Overall, the 10-item scale challenges researchers and clinicians to reconsider the assessment of positive mental well-being.

Keywords

Warwick-Edinburgh Mental Well-Being Scale, positive mental well-being, Rasch

The construct of positive mental well-being has received increased attention worldwide, not only because of its major implications for health and social outcomes and psychological functioning (Stewart-Brown et al., 2009) but also for its contribution to public health and policy issues (Dolan, Layard, & Metcalfe, 2011). In countries such as the United Kingdom, the promotion of positive mental well-being has become a national priority (Clarke et al., 2011) and this has led to an increased emphasis on mental well-being as a more preventive and population-based complement to the treatment of psychopathology (Perry, Presley-Cantrell, & Dhingra, 2010).

Nevertheless, despite the burgeoning interest in positive mental well-being, population surveys and intervention evaluations still tend to rely on measures of mental illness (e.g., prevalence of depression or anxiety) as their "marker" of mental well-being. A primary reason for this has been the scarcity of appropriate measures for measuring positive mental well-being in individuals and populations (Hu, Stewart-Brown, Twigg, & Weich, 2007). Compounding this, is that where instruments do exist, they tend to have been developed and used within the same populations (Vaingankar et al., 2011) and are therefore unlikely to be valid in other countries where concepts of mental well-being may be unique and culturally specific (Tennant, Hiller, et al., 2007; Vaingankar et al., 2011). Examples of this include the Positive and Negative Affect Scale in the United Kingdom (Watson, Clark, & Tellegen, 1988), the Scales of Psychological Well-Being (see Kafka & Kozma, 2002) and the SF-36 (Ware, Snow, Kosinski, & Gandek, 1993) in the United States, and in Europe, the Eurobarometer (Lehtinen, Sohlman, & Kovess-Masfety, 2005).

Tennant, Hiller, et al. (2007) developed the *Warwick-Edinburgh Mental Well-Being Scale* (WEMWBS) in response to the increasing recognition that obtaining a comprehensive picture of the levels of mental well-being in a population and to understand the factors that influence it requires measures of positive mental well-being which supplement the plethora of other instruments that measure the negative aspects (Deary, Watson, Booth, & Gale, 2013). Consisting of 14 items the WEMWBS covers hedonic (i.e., subjective well-being and frequent positive affect, high life satisfaction, and infrequent negative affect) and eudaimonic

[1]The University of Western Australia, Perth, Western Australia, Australia
[2]University of Strathclyde, Glasgow, Scotland
[3]Curtin University, Perth, Western Australia, Australia

Corresponding Author:
Stephen Houghton, Graduate School of Education, The University of Western Australia, Crawley, Perth, Western Australia 6009, Australia.
Email: stephen.houghton@uwa.edu.au

(i.e., psychological functioning, self-actualization) well-being (Deary et al., 2013), both of which together comprise the broad concept of mental well-being (Clarke et al., 2011). Keyes (2002) has argued that it takes a combination of these to be considered as mentally healthy. For a description of the development of the WEMWBS, including favorable construct validity, discriminatory powers, and test–retest reliability see Clarke et al. (2011), Lloyd and Devine (2012), Tennant, Hiller, et al. (2007), Tennant, Joseph, and Stewart-Brown (2007), and Vaingankar et al. (2011).

With reference to instrument validity (i.e., the validity of test score interpretations) in different populations and countries, there is now evidence showing the suitability of the WEMWBS for measuring positive mental well-being in populations that may differ from those in which the properties (of the WEMWBS) have been assessed previously (see Bartram, Sinclair, & Baldwin, 2013; Clarke et al., 2011; Gremigni & Stewart-Brown, 2011; Lloyd & Devine, 2012; López et al., 2013). Moreover, there is increasing evidence from confirmatory factor analyses supporting a single underlying factor for the 14-item WEMWBS (e.g., Clarke et al., 2011; Lloyd & Devine, 2012; López et al., 2013; Tennant, Hiller, et al., 2007) and for a 7-item shorter version (e.g., Gremigni & Stewart-Brown, 2011). However, even with satisfactory to excellent model fit statistics being reported, the WEMWBS has not been immune to criticism. For example, the issue of item redundancy has been raised (Gremigni & Stewart-Brown, 2011; Tennant, Hiller, et al., 2007). In addition, Clarke et al. (2011) identified issues pertaining to definition and understanding of items, potential for misinterpretation of items, items possibly causing embarrassment, and variable interpretation where items referred to a more holistic reflective approach to oneself.

Studies to date have primarily focused on applying classical test theory (CTT) in testing the validity of the construct and the reliability of the WEMWBS scores. An inherent weakness with this is that conventional analytical techniques based on CTT require linear, interval scale data input (Wright, 1997). However, raw data collected through Likert-type scales are always ordinal because the categories of these scales indicate only ordering without any proportional levels of meaning (see Bond & Fox, 2013). Consequently, misleading conclusions can be drawn.

In an attempt to overcome this, a number of studies have applied the Rasch model (Rasch, 1960, 1980), which focuses on modeling responses at the item level rather than the test level (Cooper et al., 2015). Based on concepts of unidimensionality, invariance, and item difficulty/person ability, the Rasch model assumes that the probability of endorsing an item is a logistic function of the relative difference between item location (difficulty) and person location (ability), which are both measured on a linear scale. It overcomes the inherent weaknesses in CTT by converting ordinal data into continuous interval measures which have a constant interval meaning (Hagquist, Bruce, & Gustavsson, 2009), and in doing so provides objective measurement from ordered category responses (Linacre, 2006). These responses (i.e., raw scores) from different items representing different severity can then be summated, which allows for quantitative comparisons of items and identification of repetitive items or gaps in a scale (Zaporozhets et al., 2015). Misfitting items that correlate with other items of an instrument, but fail to measure the same construct can also be identified by the Rasch model (Zaporozhets et al., 2015). Furthermore, differential item functioning (DIF) can be examined, thereby yielding crucial information about measurement equivalence irrespective of cultural groups and the ages and genders within these groups (see Tennant, McKenna, & Hagell, 2004).

Stewart-Brown et al. (2009) tested the Rasch model on WEMWBS date and found that the fit to the model expectations was poor. After deleting seven items of the WEMWBS, many of which showed considerable bias for gender, a strict unidimensional seven-item scale was obtained, which was named the Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS). Moreover, the polytomous response structure of the new shortened scale worked, with higher scores within an item reflecting greater overall positive mental well-being. In a study involving 500 veterinary professionals (Bartram et al., 2013), the original 14-item WEMWBS deviated significantly from Rasch model expectations. The sequential removal of seven items (the same as in Stewart-Brown et al., 2009) however, to satisfy the strict unidimensionality expectations of the Rasch model, produced a robust interval-level instrument suitable for use as an indicator of population mental well-being. However, as Meijer and Egberink (2012) highlighted, removing items from a scale that violate the assumption of invariant item ordering "requires a delicate balance between different psychometric and content arguments" (p. 603). Therefore, researchers must evaluate whether removing items to achieve invariant item ordering weakens the content validity of the scale and its measurement reliability.

Mokken models, which although based on item response theory are unlike Rasch models in that they are nonparametric and relax some of the strong assumptions about the nonlinear behavior of response probabilities (that are invoked by Rasch models), have been used to retain more items in a scale (Stochl, Jones, & Croudace, 2012). For example, Deary et al. (2013) demonstrated through Mokken scaling that most of the WEMWBS items were suitable for measuring the latent construct of positive mental well-being and that by retaining more items a more authentic assessment (and ordering of items) of the latent construct was offered (in comparison with the Rasch model).

While these studies using the Rasch model, and in the case of Deary et al. (2013) the Mokken model of double

monotonicity, have been promising, it is acknowledged that the psychometric properties are not inherent to an instrument and therefore it is important to examine these properties in different populations and in different settings. Furthermore, like many health rating scales, the WEMWBS uses the Likert-type summated scales ratings method, which can lead to potential misrepresentation. Specifically, although the total scores of the WEMWBS have a rank order the intervals between each score are not necessarily equal (Bartram et al., 2013; Hobart & Cano, 2009). Therefore, because the intervals between successive scores on the scale does not necessarily reflect equidistant steps in the severity of the underlying construct, it is difficult to make meaningful score comparisons between subgroups or interpret change (i.e., in severity) over time (Hobart & Cano, 2009). Furthermore, concerns have been expressed with the WEMWBS as to whether the categories of an item work as expected, and whether items have been tested for DIF (Bartram et al., 2013; Stewart-Brown et al., 2009). Although not developed using Rasch, item scores in the WEMWBS are typically summed. When doing this, one has to determine that all items measure the same thing, otherwise a sum score is meaningless.

The Rasch model, which relies on stronger statistical assumptions than Mokken models and has the advantage of parsimony, familiarity and a straightforward interpretation of parameters and thus findings (Stochl et al., 2012) was selected for the present study. By considering both psychometric and item content, as well as the theoretical relevance of the items, along with confirming that any DIF is real (and not artificial), the present Rasch model study will produce a more authentic assessment of positive mental well-being that is representative of both hedonic and eudaimonic well-being. Furthermore, DIF is a potential source of bias in person measurement and while associations have been found between mental illness and income (e.g., Kessler et al., 2009; Wray, Dvorak, & Martin, 2013), gender (Kessler et al., 2009; Wray et al., 2013) and age (Jones, 2013), and greater income inequality and worse population health (Wilkinson & Pickett, 2006), the associations between income and mental well-being are not so clear (Kearns, Whitley, Bond, Egan, & Tannahill, 2013). Therefore, the present study sought to extend the work of Stewart-Brown et al. (2009) who highlighted the importance within the framework of Rasch measurement of ensuring no gender and age bias exist in a scale so that effects can be properly understood, by also testing for DIF by income.

The aims of the present study were to (a) evaluate the psychometric properties of the construct and the reliability of the WEMWBS scores in an Australian sample, more specifically item invariance across subgroups of the population, operation of response categories, and identification of redundant items and (b) examine the construct validity of the Rasch-derived interval scale.

Method

Participants

The sample comprised 2,005 respondents aged 18 to 69 years (953 males and 1,052 females), two thirds of who resided in the Metropolitan area of Perth, the capital city of Western Australia, with the remainder living in country areas of Western Australia. Of the participants, 277 were aged 16 to 29 years, 239 aged 30 to 39 years, 426 aged 40 to 49 years, 545 aged 50 to 59 years, and 518 aged 60 to 69 years. Household income was used as a variable representing socioeconomic status: 313 (15.6%) earned less than Aus$40,000, 498 (24.8%) earned between Aus$40,000 and $79,000, and 882 (44.0%) earned more than Aus$80,000 (312 respondents [15.6%] did not supply information). The large majority of respondents (88%) indicated no ethnic affiliation. Among the 12% with an ethnic affiliation, 45% were from Anglo Saxon/European descent with 30% from the Asian region and 22% from a wide range of other countries. Overall, 94% of respondents completing the survey indicated English was spoken fluently in their household.

The sample was weighted to match the age and location distribution of the Western Australian population aged 16 to 69 years according to 2006 census data (Australian Bureau of Statistics, 2006).

Measure

The *WEMWBS* (see Tennant, Hiller, et al., 2007; Tennant, Joseph, et al., 2007) was developed to assess positive mental well-being at a general population level. It comprises 14 positively worded items to which participants respond using a 5-point Likert-type scale (scored 1 "None of the time," 2 "Rarely," 3 "Some of the time," 4 "Often," 5 "All of the time"), thereby providing a total score of 14 to 70. Responses are based on participant's feelings over the previous 2 weeks. Higher levels of positive mental well-being are indicated by higher scores. (For the purposes of a Rasch analysis, scoring has to begin at 0 and therefore in the present study the response options were 0 to 4.)

Rasch Analysis

In the Rasch model (Rasch, 1960, 1980), the construction of stable linear measures is guided by the principle of invariant comparison (i.e., the relative location of two people on the continuum is not dependent on items to which they respond and the relative location of two items does not depend on the people from whom the estimates are made). Determining the extent to which observed rating scale data satisfy the measurement model is therefore an aim of a Rasch analysis (Bartram et al., 2013). When there is a misfit between the data and the model, explanations for the misfit are sought and the instrument adapted accordingly for future administrations (Cano & Hobart, 2011).

In this present study, the polytomous Rasch model (Rasch, 1960, 1980) was used, and in particular the partial credit parameterization of the model (see Andrich, 2009) using the software RUMM2030 (Andrich, Sheridan, & Luo, 2010). In the rating scale parametrization of the polytomous model the thresholds are the same for all items, whereas in the partial credit parameterization the number and values of the thresholds can differ for each item. In RUMM2030, an estimate of the reliability of the scale is available as a person separation index (PSI) of reliability, with values ranging between 0 and 1. This reflects the instrument's capability to differentiate persons on the continuum. Under conditions where there are no floor or ceiling effects, the PSI of reliability is equivalent to Cronbach's alpha (Andrich, 1982). RUMM2030 also has an option for modifying the sample size for the chi-square and analysis of variance fit statistics. To reduce the risk of Type I errors, the alpha level for the item chi-square tests of fit was adjusted using the Bonferroni correction (Bland & Altman, 1995).

Procedure

Permission for the research was granted by the human research ethics committee of the administering institution and the developers of the instrument. Trained interviewers collected the data via the computer-assisted telephone interviewing method for survey administration. Household telephone numbers were randomly selected from an electronic household telephone directory that represented 88% of Western Australian households (Australian Communications and Media Authority, 2012). A random within household sampling method for telephone surveys was utilized so that the chances of selecting a mix of gender and age groups was balanced. In this method, the adult in the house with the next birthday was invited to participate in the survey (Lavrakas, 1993). Up to 10 callbacks were made to each household and this resulted in an affirmative rate of 60%.

Results

We compared whether the data fit the partial credit or the rating scale parameterization best. In the partial credit parameterization, the thresholds are different for each item, whereas each item has the same thresholds in the rating scale parameterization. The data fitted the partial credit parameterization better than the rating scale parameterization of the polytomous model for all of the analyses, as indicated by the likelihood ratio test in RUMM2030 (χ^2 = 403.125, degrees of freedom = 41, $p < .000$).

Response Option Thresholds Ordering

Table 1 shows the WEMWBS items, the response categories, the percentages of responses in each response category for each item, as well as the scoring of each response category. Most of the responses are in the "Often" and "All the time" categories, with the exception of Item 5 (*I have had energy to spare*), which had a considerable number of responses in the other categories. None of the categories had 0 response frequencies. In 9 of the 14 items, the thresholds were disordered. In the remaining 5 items, thresholds are located very close to each other on the continuum. In other words, they are close to being disordered. A threshold is the point on the measurement continuum where a response in either of two adjacent response categories is equally likely. The thresholds partition the measurement continuum into ordered categories for each item. If the response categories work as intended, the thresholds are in the same order as the categories. Thresholds that are not in order, that is, reversed-order thresholds, are evidence supporting the idea that the response categories of the rating scale do not function as intended.

The upper graph in Figure 1, which shows the category probability curves for Item 4, indicates which response categories were problematic. The curves for Item 4 are representative of all the items with disordered thresholds. For data to fit the model, each response category should successively show the highest probability of endorsement. The graph shows that Category 1 does not have a point where it is the most likely response. The threshold between Categories 0 ("None of the time") and 1 ("Rarely") is further to the right on the continuum than the threshold between Categories 1 ("Rarely") and 2 ("Some of the time"). It seems respondents could not readily distinguish between response categories "None of the time" and "Rarely."

Given persons could not distinguish between responses in the first two categories, the responses were rescored as follows: "None of the time" and "Rarely" (both scored 0), "Some of the time" (1), "Often" (2), and "All of the time" (3), for all items. The lower graph in Figure 1 shows the category probability curves for Item 4 *after* rescoring. The curves for this item are representative of all the items after rescoring. Each response category now has a point where it is the most likely response. The thresholds for all items were ordered after rescoring. For future data collection with the WEMWBS in this population, the two categories "None of the time" and "Rarely" should form one category.

Targeting and Reliability

Rasch item estimates are on the same scale as the person estimates in the Rasch model, which means the alignment of persons to items (i.e., targeting) can be assessed. A lack of items at certain points on the measurement continuum can then be identified. The upper graph in Figure 2 shows the frequency distributions of the estimated person and item threshold locations after rescoring. Positive values on the continuum represent higher levels of positive mental well-being. It is clear

Table 1. Percentage of Responses in Each Response Category and Scoring of Responses.

Item number	Statement	Category percentages				
		None of the time	Rarely	Some of the time	Often	All of the time
1	I have been feeling optimistic about the future.	4	8	28	41	18
2	I have been feeling useful.	2	2	25	41	28
3	I have been feeling relaxed.	3	10	38	37	12
4	I have been feeling interested in other people.	3	5	26	42	24
5	I have had energy to spare.	12	19	38	22	9
6	I have been dealing with problems well.	1	4	22	48	25
7	I have been thinking clearly.	1	2	18	51	28
8	I have been feeling good about myself.	1	4	23	46	25
9	I have been feeling close to other people.	1	4	24	45	25
10	I have been feeling confident.	1	4	25	47	23
11	I have been able to make up my own mind about things.	1	2	12	42	44
12	I have been feeling loved.	2	4	16	37	41
13	I have been interested in new things.	2	7	26	34	30
14	I have been feeling cheerful.	1	4	25	48	23
	Original scoring (Analysis 1)	0	1	2	3	4
	Rescoring (Analysis 2)	0	0	1	2	3

from the graph that the item thresholds aligned well with the persons. The person locations were normally distributed with a mean of 0.706 (*SD* = 1.22). The PSI of reliability and Cronbach's alpha was .88, indicating that the analysis separated the persons sufficiently on the continuum.

Fit of Responses to the Rasch Model

The approach to assessing fit in this research is that of Smith and Plackner (2009), who recommend a

"family approach" to assessing fit in Rasch measurement. Different fit statistics assess different types of misfit and by considering more than one fit statistic for each item a complete picture of misfit can be obtained. If an item misfits according to one fit statistic, the misfit is considered less severe than if the item misfits according to more than one fit statistic. Moreover, fit statistics should be interpreted in context, ordered in the first instance and those items with the worst fit studied first. It is not a matter of assessing statistical fit only against some critical values. Critical values, such as a fit residual greater than +2.5 or less than −2.5 can be helpful, but it should not be the sole basis used for excluding an item from a test. Graphs of item characteristic curves (ICC) should also be considered. Furthermore, every conclusion should be considered against the substantive variable, its purpose and its application. Items should not be deleted on the grounds of statistical misfit only because this is not consistent with sound measurement practice. Explanations of misfit should be sought and these seen as hypotheses for further empirical testing.

Table 2 shows how the 14 items of the WEMWBS fitted the Rasch model as well as the item locations, with their standard errors for all items. (Items in Table 2 are ordered by their location on the continuum.) The fit residual fit statistic is equivalent to the outfit fit statistic. It compares the *sums of individual* item-person residuals (i.e., residuals—the difference between what is expected

according to the model and what is observed) in a similar way to the outfit statistic by summing the squared, standardized residual over the persons to obtain a summary value. A value that is negative and large in magnitude reflects an item with a response pattern whose discrimination tends to be greater than that of the average discrimination of the rest of the items. Likewise, a positive value that is large in magnitude reflects an item with a response pattern whose discrimination tends to be less than that of the average discrimination of the rest of the items. Typically, item fit residuals that lie within the range of approximately −2.5 to +2.5 would be considered fitting the model based on this one criterion. However, we did not apply critical values of ±2.5 as a rule of thumb. We interpret fit statistics relatively and in context.

The easiest item to endorse was Item 11 (*I have been able to make up own mind about things*) and by far the most difficult item to endorse was Item 5 (*I have had energy to spare*). Items 5, 8, and 10 had the largest fit residuals, relative to the other items, and the chi-square probabilities show that the misfit was statistically significant.

Figure 3 shows the ICCs for the three most misfitting items (5, 8, and 10). Item 5 (*I have had energy to spare*) discriminated less than that expected according to the

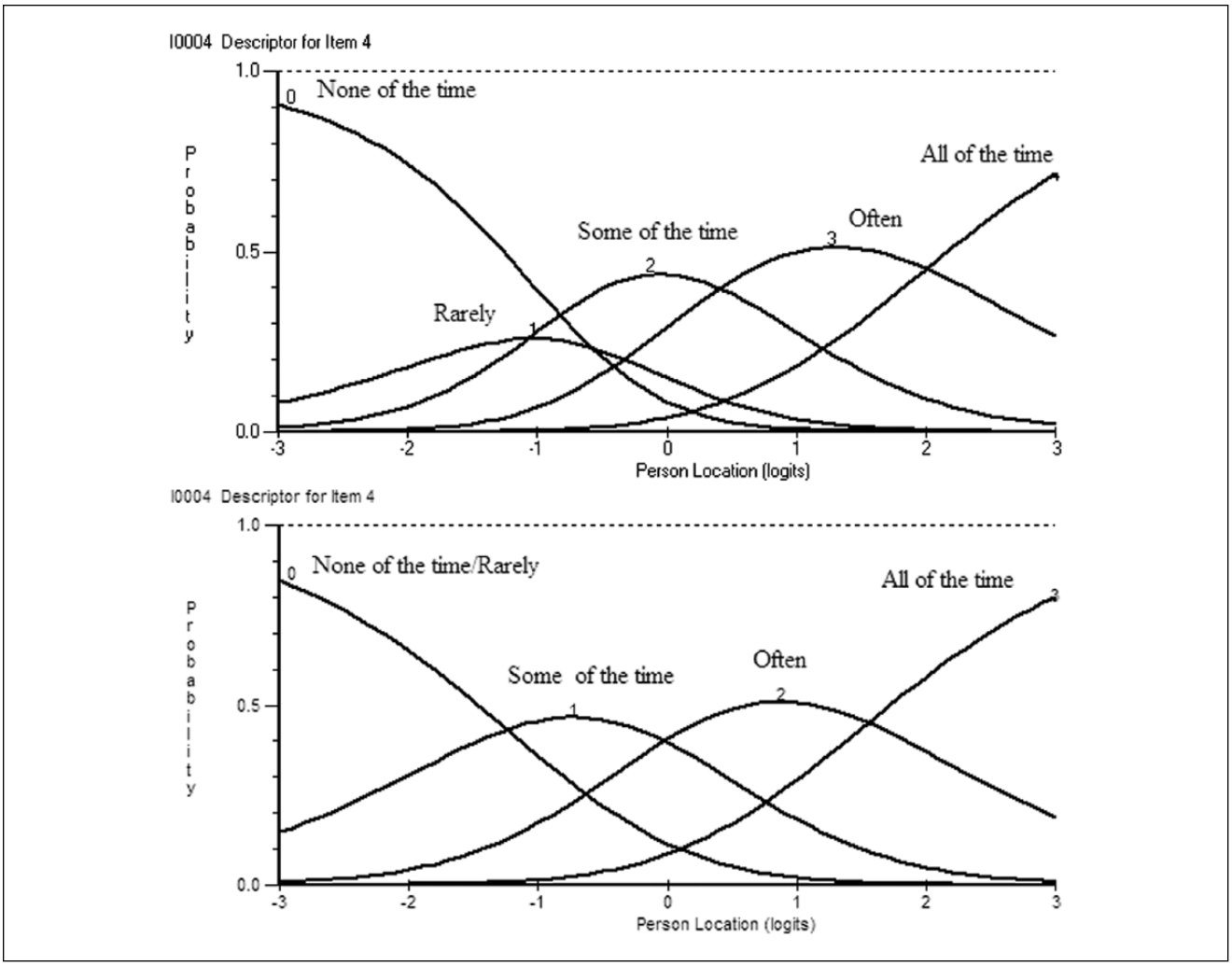

Figure 1. Response category probability curves before rescoring (top) and after rescoring (bottom) for Item 4. After rescoring each response category successively shows the highest probability of endorsement.

model, suggesting that it measured something slightly different from the other items. Items 8 (*I have been feeling good about myself*) and 10 (*I have been feeling confident*) discriminated more than that expected according to the model. There are a number of reasons why items sometimes discriminate more than the average, including response dependency between the items, and that these items "summarize" or capture all aspects measured by the other items.

Table 3 provides summary fit information. The mean of the individual item fit residual statistics was 0.578 (*SD* = 5.65). The mean of the person fit residual statistics was −0.4920 (*SD* = 1.79). If data fitted the model, the means for the fit residuals should be close to 0 and the standard deviation close to 1. The large standard deviation of the item fit residuals (5.65) indicates that the items do not fit the model. Person fit appears to be better than item fit. The total item-trait chi-square statistic (for which the groups are formed on the basis of persons' estimates on the trait measured, in this case mental well-being) is the sum of all the individual item chi-squares. It reflects the property of invariance across the trait and in the present study it was significant, demonstrating that the hierarchical ordering of the items varies across the trait (i.e., the items are ordered differently on difficulty for persons with low values of the trait compared with those with high values on the trait).

Independence of Responses (Response Dependency and Multidimensionality)

Fit indices reported earlier indicate that the items do not fit the model. This section reports the results of specific tests of misfit, namely tests for item response dependency and multidimensionality.

Response Dependency. The correlation between the standardized item residuals for Items 8 (*I have been feeling*

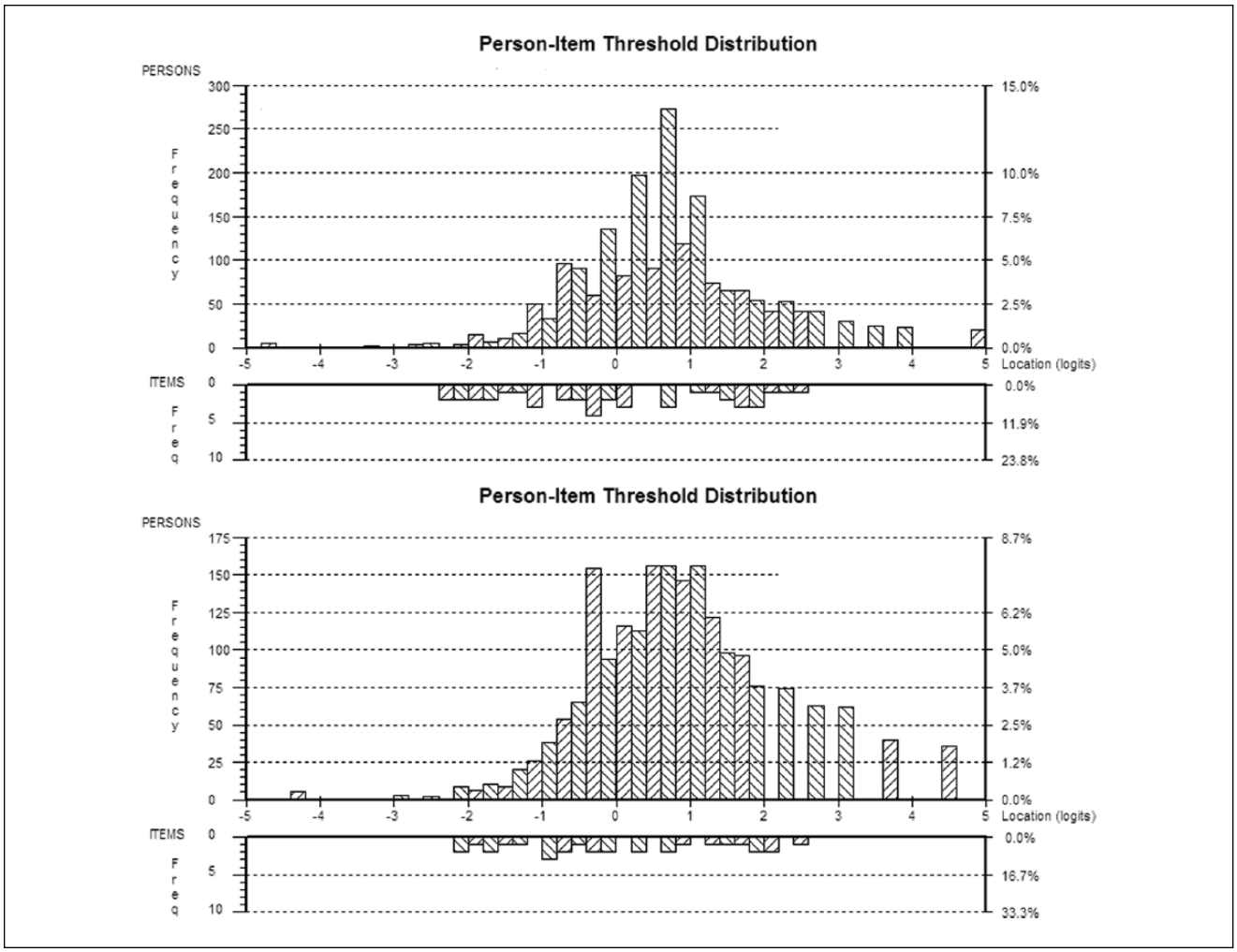

Figure 2. Person threshold distributions for analyses with 14 items (upper) and 10 items (lower), both showing that the items align well with the persons.

good about myself) and 10 (*I have been feeling confident*) was high relative to the other items, .270. This indicates response dependency between these two items. The next highest correlation, .197, was between Items 6 (*I have been dealing with problems well*) and 7 (*I have been thinking clearly*), indicating a smaller amount of response dependency between these two items. The response dependency between Items 8 and 10 was not the reason these items discriminated more than the average, because when either item was deleted, the other still overdiscriminated. The more likely reason for these two items overdiscriminating is that they are "overall well-being" items, capturing all aspects measured by the other items. In studies where factor analysis has been performed on the WEMWBS, Items 8 and 10 have had the highest factor loadings of all items (e.g., Lloyd & Devine, 2012: .854 and .855; Spittlehouse, Vierck, Pearson, & Joyce, 2014: .837 and .846, respectively), thereby suggesting their overly high discrimination results from the fact that they capture variation common to the other items.

Differential Item Functioning

No items showed statistical DIF for income. Item 5 (*I have had energy to spare*) showed statistical DIF for age and Items 4 (*I have been feeling interested in other people*), 8 (*I have been feeling good about myself*), and 10 (*I have been feeling confident*) showed statistical DIF for gender. The statistical DIF was identified through an analysis of variance of person-item residuals, where under the null hypothesis of no DIF, the ratio of the mean square of the residuals between groups and the mean square residuals within groups is not statistically significant. Andrich and Hagquist (2012, 2015) distinguished between real and artificial DIF because an item can show DIF that is not real as a result of real DIF in other items. Briefly, their method involves deleting the item which shows the highest DIF and then investigating the DIF results after deletion. The next item with highest DIF (in the new table produced) is found, deleted, and then the DIF table investigated again after this second

Table 2. All Original 14 Items With Fit Statistics Ordered by Item Location.

Item number	Statement	Location	*SE*	Fit residual	Chi-square	Chi-square probability
11	Make up own mind	−0.851	0.035	1.957	7.312	.605
7	Thinking clearly	−0.482	0.036	−3.330	18.774	.027
12	Feeling loved	−0.418	0.031	3.229	14.226	.115
6	Dealing with problems	−0.188	0.034	−1.958	15.094	.088
8	**Feeling good**	**−0.184**	**0.034**	**−8.124**	**61.888**	**.000**
10	**Feeling confident**	**−0.148**	**0.035**	**−8.030**	**62.443**	**.000**
2	Feeling useful	−0.139	0.032	−1.054	6.148	.725
9	Close to others	−0.094	0.033	−2.239	19.188	.024
14	Feeling cheerful	−0.094	0.034	−4.190	30.775	.000
13	Interested in new things	0.076	0.030	3.278	6.119	.728
4	Interested in others	0.096	0.031	5.388	20.824	.013
1	Feeling optimistic	0.465	0.031	7.278	16.637	.055
3	Feeling relaxed	0.692	0.032	4.967	15.759	.072
5	**Energy to spare**	**1.268**	**0.030**	**10.920**	**107.741**	**.000**

Note. Items in bold have the largest fit residuals relative to the other items. Bonferroni adjustment: $p = .003571$ for 14 items, $p = .05$.

deletion. This process is repeated until no items show DIF. Artificial DIF disappears after all items with real DIF have been deleted. Applying the Andrich and Hagquist (2012) procedure confirmed that the statistical DIF in these items was real.

Figure 4 shows the DIF graphically for each item. Elaborating on the interpretation of the statistical DIF in terms of the interaction between content of the items and the composition of the groups, Figure 4 shows that for Item 5 (*I have had energy to spare*), for persons at the same level of positive mental well-being, those in the youngest age group (<30 years) scored higher than those in the other groups (>30 years). This may be explained in part by age given that more than 50% of respondents were more than 50 years of age and only 13.8% were under 30 years of age. Due to physical ageing, older respondents would be expected to have less energy, despite their level of mental well-being. Item 5 may therefore be measuring an aspect of physical health that is affected by age, rather than a purely mental well-being aspect of health. Thus, based on both the statistical identification and the substantive interpretation of DIF, this item was considered for deletion in subsequent analyses.

Figure 4 also shows that for Item 4 (*I have been feeling interested in other people*), for the same level of mental well-being as estimated using all the items, women scored higher on this item than men. For Items 8 (*I have been feeling good about myself*) and 10 (*I have been feeling confident*) the DIF was in the opposite direction. For the same level of mental well-being, men scored higher on these items than women. These findings may reflect other research evidence showing that at the general population level, women display a greater emotional range, have different methods of interpreting mental health symptoms, and are more likely to seek help from others for these compared with men (Buffel, Van de Velde, & Bracke, 2014;Wilhelm, 2014). These items also misfitted and may be measuring something that the other items, taken together, already capture. Based on both the statistical identification and the substantive interpretation of DIF, Items 4, 8, and 10 were all considered for deletion in subsequent analyses.

Unidimensionality

After extracting the Rasch factor, there should be no further pattern among the residuals. If a principal component analysis (PCA) indicates a meaningful pattern, the scale or test is not unidimensional. The eigenvalue of 1.997 for the first component was somewhat larger than the eigenvalues for the other components. The first principal component explained 14.26% of the total variance among residuals. This suggests some multidimensionality. A Rasch PCA of the residuals indicated that the loadings directed the items into two different sets. When the items that loaded positively on the first component (6, 7, 8, 10, 11, 14) and the items that loaded negatively (1, 2, 3, 4, 5, 9, 12, 13) on the first component were used as separate sets of items, *t* tests for 182 of the 2,005 cases (9.08%) were significant at $p < .05$. This indicated that the person locations estimated from these two sets differed significantly in 9.08% of cases, which is higher than the 5% expected by chance and indicates a violation of unidimensionality.

The Derived 10-Item Scale

Because of DIF, item fit, and possible redundancy, Items 5, 8, and 10 were considered for deletion. However, deletion of items was done iteratively, one item at a time followed by

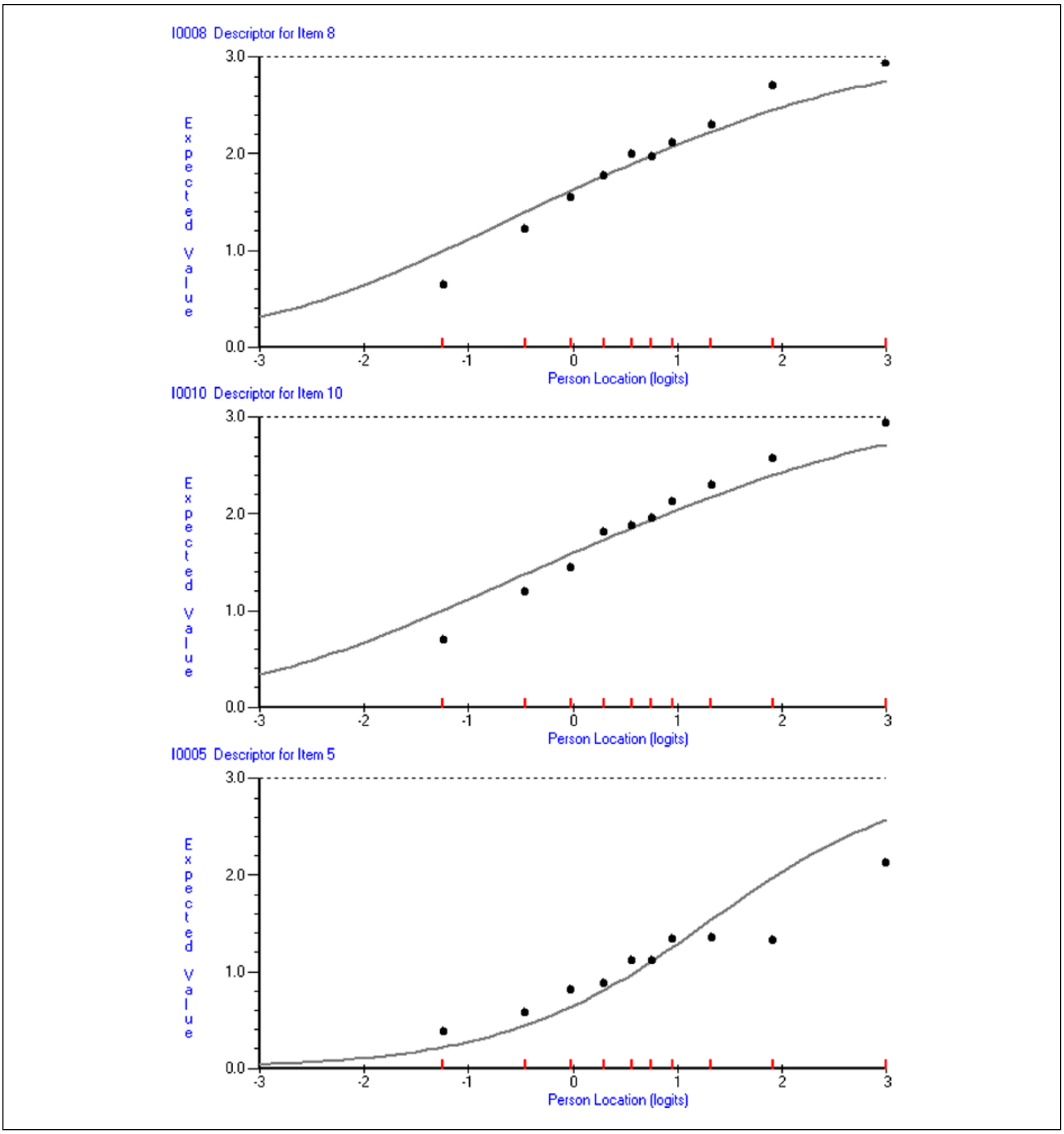

Figure 3. ICCs of the three most misfitting items (total 14 items in analysis) showing Items 8 and 10 discriminated more than expected and Item 5 discriminated less than expected.

a new assessment of misfit after deletion. The item with the greatest misfit, Item 5 (*I have had energy to spare*), was deleted first. This item also showed DIF for age. After that, Item 10 (*I have been feeling confident*) was the item with the poorest fit, in addition to showing DIF for gender. After deletion of Item 10, Item 8 (*I have been feeling good about myself*) was the poorest fitting item and also showed DIF for gender. Following its deletion, no items misfitted according to both the fit residual and chi-square fit statistics. Because Item 4 (*I have been feeling interested in other people*) was the only DIF item remaining and misfitted according to the fit residual statistic, it was also deleted.

Table 3. Fit of Data to the Rasch Model.

Analysis[a]	Item fit residual		Person fit residual		Chi-square value, *df*	*p*	PSI	Unidimensional *t* test	Cronbach alpha
	Mean	*SD*	Mean	*SD*					
1	0.734	5.761	−0.525	1.767	395.048, *df* = 126	.000	.88		.88
2	0.578	5.652	−0.492	1.793	402.927, *df* = 126	.000	.88	9.08%	.88
3	0.845	3.623	−0.5	1.621	109.849, *df* = 90	.080	.83	6.68%	.85
4	0.576	1.621	−0.514	1.655	110.110, *df* = 90	.074	.83	5.00%	.84
5	0.581	1.597	−0.481	1.599	126.746, *df* = 90	.007	.84	5.60%	.85

Note. *df* = degrees of freedom; PSI = person separation index.
[a]Detail of analyses: Analysis 1: 14 items, original scoring; Analysis 2: 14 items, rescored; Analysis 3: 10 items (final scale); Analysis 4: 10 items (final scale) tested on Random Sample 1; Analysis 5: 10 items (final scale) tested on Random Sample 2.

This left a total of 10 remaining items. These items tended to separate into aspects representing the eudaimonic and hedonic elements of positive mental well-being that spanned different regions of the continuum.

It is important to note that response dependency is typically found where an item requests, for example, an overall level of satisfaction following several other such items (e.g., the Course Experience Questionnaire: Wilson, Lizzio, & Ramsden, 1997). Items 8 and 10 in the WEMWBS are response dependent as they function in the same way as this, summarizing the other items on well-being. Because the items do not provide any additional information over and above that provided by the other items, they can be considered redundant and should be considered for deletion. Moreover, response dependency between items results in inflated reliability (Marais & Andrich, 2008). For these reasons, we decided to delete Items 8 and 10.

With reference to the ordering of these 10 items, it is not unlike that reported in the WEMWBS development study using confirmatory factor analyses (Tennant, Hiller, et al., 2007) and in recent studies using Rasch (e.g., Bartram et al., 2013), particularly in terms of Items 11 (*making up mind about things*), 7 (*thinking clearly*), and 6 (*dealing with problems*) being the easiest to endorse (i.e., they spanned the lower range of the continuum). Furthermore, there appears to be an alignment between the ordering of items and the complex construct of mental well-being. Specifically, the ordering tends to reflect the distinct perspectives of mental well-being which focus on psychological functioning and self-realization (i.e., eudaimonic elements) and the subjective experience of happiness and life satisfaction (i.e., hedonic elements; see Clarke et al., 2011; Ryan & Deci, 2001; Tennant, Hiller, et al., 2007). Consistent with definitions of mental well-being (see Ryan & Deci, 2001), the construct represented in the present Rasch analysis covers the majority of eudaimonic and hedonic concepts associated with positive mental well-being that are recognized as having major consequences for mental health and social outcomes (Hupert & Whittington, 2004; Tennant, Hiller, et al., 2007).

After deletion of these four items, there were no items showing statistical DIF and no items with response dependency. Table 3 shows the summary statistics after deletion of these four items. The PSI dropped from .88 to .83 after removal of the four items. Even though the total item-trait chi-square statistic was not significant and indicated the data fitted the model, the standard deviation of the item fit residual (3.623) was still quite large and not close to 1. This standard deviation reflects three remaining items with large item fit residuals relative to the other items: Item 1 (*I have been feeling optimistic about the future*, fit residual of 6.686); Item 3 (*I have been feeling relaxed*, fit residual of 5.706); and Item 13 (*I have been interested in new things*, fit residual of 3.922). These items did not misfit according to the chi-square fit statistic. Figure 5 shows the ICCs for these items. The misfit was not severe with the dots, showing the observed values in each class interval, being quite close to the curve, which represents the expected value.

The lower graph of Figure 2 shows the person threshold distribution after deletion of the four items. It shows the persons were still reasonably well aligned to the items. The person distribution was not as normally distributed as in the previous analyses and showed a slight ceiling effect. The PSI was .83 and Cronbach's alpha was .85.

The results of the *t* tests for unidimensionality (see Table 3) revealed that 134 of the 2,005 cases (6.68%) had significantly different estimates based on the subsets of items identified through the Rasch PCA of residuals. Even though this is more than expected by chance, it is less than the 9.08% of the previous analysis. It must be acknowledged that with larger sample sizes, there is more power to detect misfit, along with the risk of identifying misfit that is statistically significant, but substantively irrelevant. The large sample size may therefore have played a role in the misfit reported above.

The 10-item scale in this study retains all seven items of the 7-item SWEMWBS (Bartram et al., 2013; Stewart-Brown et al., 2009) and in addition retains three extra items (Items 12, 13, and 14). The 10-item scale has face validity and a balance between hedonic (six items) and eudaimonic

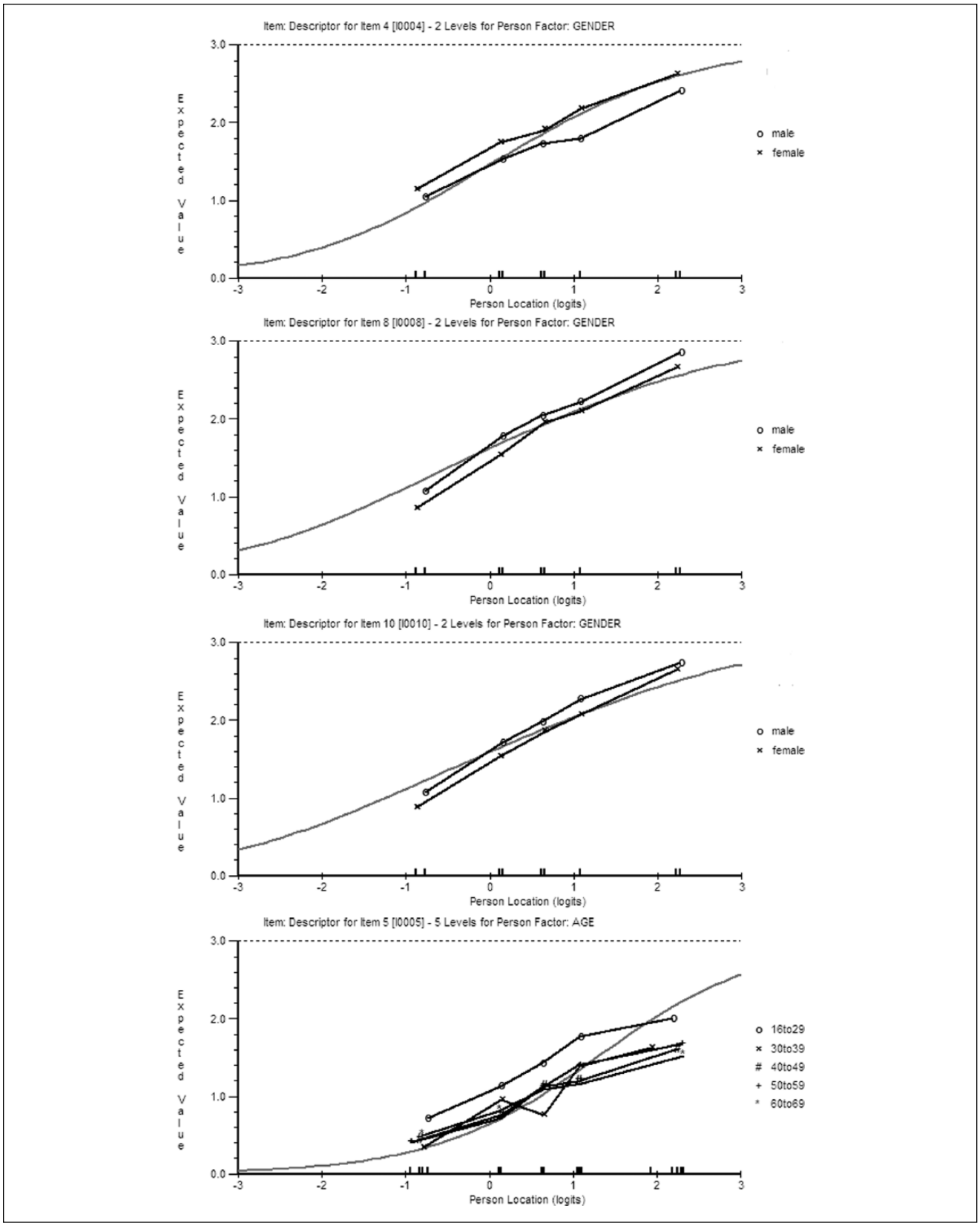

Figure 4. ICCs from the analysis of all 14 items showing DIF for gender (Items 4, 8, 10) and for age (Item 5); for the same overall level of mental well-being, one group scored higher on the item than the other group(s).

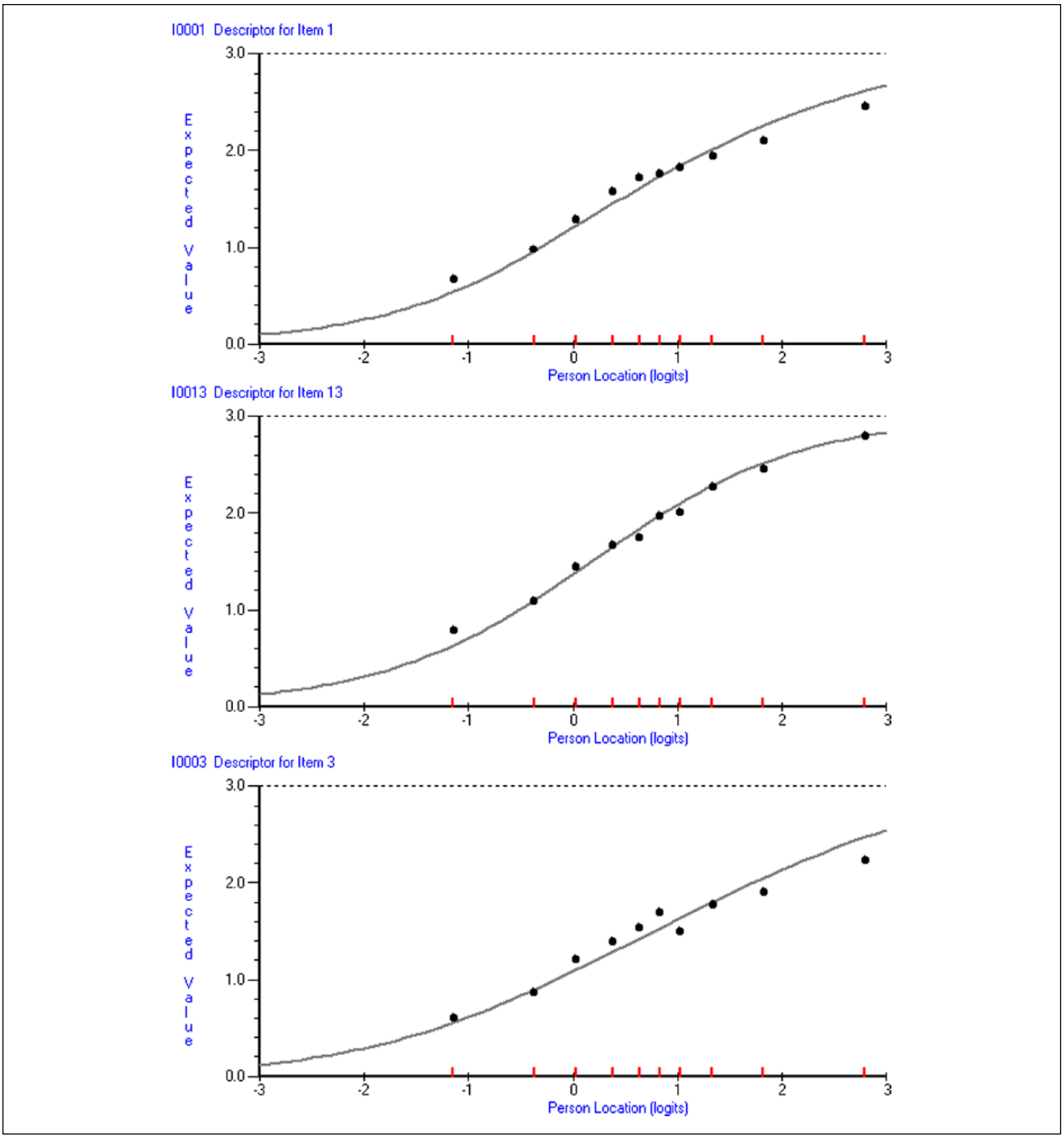

Figure 5. ICCs of the three most misfitting items (total 10 items in analysis) showing that the misfit is not severe.

(four items) elements of mental well-being. This is similar to that found in the 7-item SWEMWBS (4 hedonic and 3 eudaimonic items) and the 14-item WEMWBS (9 hedonic and 5 eudaimonic items), both of which also include more items measuring hedonic elements of mental well-being. However, the underlying construct in the current 10-item WEMWBS has less of a focus on elements of an individual's feelings of happiness, satisfaction, and interest in life (i.e., hedonic), compared with the 14-item WEMWBS.

When evaluated with a reasonable sample size of 500, the scale was found to be unidimensional. As shown in Figure 5, the misfit of the remaining items was not severe. Furthermore, no items in the scale showed DIF, thereby making comparisons of overall mental well-being for the

different age, gender, and income groups valid and accurate. The mean locations of the two gender groups on the scale were males ($M = 0.817$, $SD = 1.24$) and females ($M = 0.800$, $SD = 1.24$). The mean locations of the different age groups were 16 to 29 years ($M = 0.541$, $SD = 1.17$), 30 to 39 years ($M = 0.739$, $SD = 1.03$), 40 to 49 years ($M = 0.746$,$SD = 1.16$), 50 to 59 years ($M = 0.826$, $SD = 1.32$), and 60 to 69 years ($M = 1.014$, $SD = 1.31$).

Discussion

The primary aim of this research was to evaluate the psychometric properties of the 14-item WEMWBS among Western Australians through the application of the Rasch model. The results provide evidence to support the interval scale measurement properties of a 10-item WEMWBS that covers both eudemonic and hedonic mental well-being and which also meets the Rasch model criteria. This is supportive of the broad consensus that mental well-being is a complex construct that covers hedonic and eudaimonic well-being (Clarke et al., 2011; Ryan & Deci, 2001; Tennant, Hiller, et al., 2007).

Our Rasch analysis suggests respondents experienced difficulties in discriminating between the response options "None of the time" and "Rarely" and therefore these categories were amalgamated. Subsequently, participants were able to discriminate. In future data collections using the WEMWBS in Australian contexts in this age of population, these two categories should form one category.

Initially, the WEMWBS was not compatible with interval scaling for use with the participating population because the estimation of model fit revealed deviations from the Rasch model expectations. Consequently, by carefully examining the item fit statistics and DIF, the misfitting items were sequentially removed (iteratively, one item at a time followed by a new assessment of misfit after deletion), and this resulted in a 10-item scale. While the content validity of a scale can be compromised by removing items (Bohlig, Fisher, Masters, & Bond, 1998), the aim of this study was to produce a reliable, concise, and more authentic assessment instrument and this was achieved. However, as recommended by Meijer and Egberink (2012), we did not rely solely on statistically defined procedures as the basis for eliminating items. Rather, the item content and its theoretical relevance to the construct were carefully examined prior to any elimination. In doing so, a more authentic assessment of the latent construct was retained.

The final 10-item Rasch-derived scale had minimal item bias (DIF), minimal response dependency, and its polytomous response structure functioned as intended (i.e., higher scores within an item reflected greater overall positive mental well-being). The PSI was acceptable (.83) as was the Cronbach's alpha ($\alpha = .85$), which exceeds the minimum value (.7) for group-level measurement (Reeve et al., 2007). Thus, the measure can discriminate among groups of respondents whose levels of positive mental well-being are different.

Of the four items iteratively deleted, Item 5 (*I have had energy to spare*) showed DIF for age; Items 10 (*I have been feeling confident*), 8 (*I have been feeling good about myself*), and 4 (*I have been feeling interested in other people*) showed DIF for gender. These same items were deleted in the Stewart-Brown et al. (2009) and Bartram et al. (2013) Rasch analyses because they also showed DIF.

Each of these analyses resulted in an identical 7-item WEMWBS; in the present study the Rasch analysis led to a 10-item WEMWBS. However, the items (i.e., 1, 2, 3, 6, 7, 9, and 11) that were retained in the present study, were also retained in the Bartram et al. (2013) and Stewart-Brown et al. (2009) studies. Stewart-Brown et al. (2009) raised the point that most of the items in their seven-item scale represented aspects of eudemonic well-being, with few covering hedonic well-being. Given that the concept of positive mental well-being is regarded as covering both of these aspects (Clarke et al., 2011), the deletion of fewer items in this current study resulted in a construct that presents a more balanced view of positive mental well-being than in the seven-item scale derived by Stewart-Brown et al. (2009).

It is worth noting that there were different reasons for items being deleted from each of the three studies cited, particularly with reference to DIF. An advantage of the present study over Bartram et al. (2013) and Stewart-Brown et al. (2009) is that it distinguished between real and artificial statistical DIF (for a review of the implications of artificial DIF, see Andrich & Hagquist, 2012, 2015) and this may be why more items were retained. According to Andrich and Hagquist (2012, 2015) understanding the nature of artificial DIF is critical. Furthermore, it responds to Dorans and Holland (1993, p. 66), who argued that "future research should focus on trying to uncover testable, verifiable, robust explanations for why DIF occurs when it does."

The retention of more items in the present study compared with the other Rasch studies may have been the result of differences between samples (i.e., Australian general population; Scottish general population: Stewart-Brown et al. [2009]; and U.K. veterinary surgeons: Bartram et al. [2013]) and ages of these samples (bigger percentage in older age group, no data on age; and bigger percentage in younger age group) and differences in data collection methods used which may have particularly affected the way in which the response categories functioned. In the present study, the participants were interviewed via computer-assisted telephone, whereas in the other studies, WEMWBS administration was via face-to-face interview (Lloyd & Devine, 2012; Stewart-Brown et al., 2009) or questionnaire survey (Bartram et al., 2013). The use of computer-assisted telephone interviewing is not the standard method for administering the WEMWBS

and therefore it is possible that the present findings may not generalize when other alternative methods of administration are used. Furthermore, in the original Scottish survey on which the original WEMWBS validation was based, data were collected via computer-assisted interview, thereby preserving an element of confidentiality. The telephone interview in the present study may not have presented such a degree of confidentiality to the participants. In an attempt to address this issue, interviewers highlighted to respondents that information provided would be coded such that it would not allow the identification of individuals. Nevertheless, the participants' responses may still have been affected. Finally, that telephone respondents needed to hold a question and its response options in working memory in order to answer accurately may also have affected accuracy of responses.

It must be acknowledged that there was no demonstration of convergent and discriminant validity in the present study. This must be a strong focus of future research if the 10-item version of the WEMWBS is to be used as an indicator of population mental well-being, for monitoring population-level changes in mental well-being, and as an outcome measure in interventions. Furthermore, the response rate in the present study was 60% and it is possible that those who chose to respond might somehow be different on the variable of interest compared with those who chose not to respond.

Related to this, it is known that some people only accept calls they are expecting, while others do not accept calls from unknown numbers. Furthermore, those with mental health issues are less likely to participate in surveys (see Kessler et al., 2007). While estimated lifetime prevalence of mental disorders varies widely across countries (Kessler et al., 2007), no comparable measures for positive mental well-being appear to exist, meaning that comparisons with the present data cannot be undertaken. However, Lloyd and Devine (2012, p. 262) reported that "levels of mental wellbeing in Northern Ireland 2009/10 were similar to those in Scotland in 2006," irrespective of intervening events that are potentially detrimental to the mental well-being of a population such as a major recession. Given that Perth has not experienced such major events, it is possible that levels of mental well-being may not be any different to those reported by Lloyd and Devine (2012).

In conclusion, positive mental well-being has the potential to improve quality of life, prevent mental and physical illness, and the use of health services (Keyes, 2007; Keyes, Dhingra, & Simoes, 2010). As such there is a need for validated measures of positive mental well-being that can be used in population-level studies. This present study, which is the first non-Northern hemisphere study, adds Australian evidence to the growing body of data demonstrating that the 10-item WEMWBS is a concise robust measure for monitoring positive mental well-being.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Western Australian Health Promotion Foundation (Healthway). The Warwick-Edinburgh Mental Well-Being Scale was funded by the Scottish Executive National Programme, commissioned by NHS Health Scotland, developed by the University of Warwick and the University of Edinburgh, and is jointly owned by NHS Health Scotland, the University of Warwick and the University of Edinburgh.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95-104.

Andrich, D. (2009). Educational measurement: Rasch models. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International encyclopedia of education* (pp. 307-342) (3rd ed.). Amsterdam, Netherlands: Elsevier.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*, 387-416.

Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement*, *75*, 185-207.

Andrich, D., Sheridan, B. E., & Luo, G. (2010). RUMM2030 [A Windows program for Rasch Unidimensional Models for Measurement]. Perth, Western Australia, Australia: RUMM Laboratory.

Australian Bureau of Statistics. (2006). *National Health Survey: Summary of results* (CDATA01). Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/mf/4364.0

Australian Communications and Media Authority. (2012). *Convergence and communications Report 1: Australian household consumers' take-up and use of voice communications services*. Retrieved from http://www.acma.gov.au/theACMA/Library/researchacma/Research-reports/research-report-convergence-and-communications

Bartram, D. J., Sinclair, J. M., & Baldwin, D. S. (2013). Further validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in the UK veterinary profession: Rasch analysis. *Quality of Life Research*, *22*, 379-391.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal*, *310*, 170.

Bohlig, M., Fisher, W. P., Jr., Masters, G. N., & Bond, T. (1998) Content validity and misfitting items. *Rasch Measurement Transactions*, *12*(1), 607.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Psychology Press.

Buffel, V., Van de Velde, S., & Bracke, P. (2014). Professional care seeking for mental health problems among women and men in Europe: The role of socioeconomic, family-related

and mental health status factors in explaining gender differences. *Social Psychiatry and Psychiatric Epidemiology*, *49*, 1641-1653.

Cano, S. J., & Hobart, J. C. (2011). The problem with health measurement. *Patient Preference and Adherence*, *5*, 279-290.

Clarke, A., Friede, T., Putz, R., Ashdown, J., Martin, S., Blake, A., . . . Stewart-Brown, S. (2011). Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Validated for teenage school students in England and Scotland. A mixed methods assessment. *BMC Public Health*, *11*(1), 487.

Cooper, Z., Doll, H., Bailey-Straebler, S., Kluczniok, D., Murphy, R., O'Connor, M. E., & Fairburn, C. G. (2015). The development of an online measure of therapist competence. *Behaviour Research and Therapy*, *64*, 43-48.

Deary, I. J., Watson, R., Booth, T., & Gale, C. R. (2013). Does cognitive ability influence responses to the Warwick-Edinburgh Mental Well-Being Scale? *Psychological Assessment*, *25*, 313-318.

Dolan, P., Layard, R., & Metcalfe, R. (2011). *Measuring subjective well-being for public policy* (Office for National Statistics). Retrieved from http://eprints.lse.ac.uk/35420/1/measuring-subjective-wellbeing-for-public-policy.pdf

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Gremigni, P., & Stewart-Brown, S. L. (2011). Measuring mental well-being: Italian validation of the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS). *Giornale Italiano di Psicologia*, *2*, 485-508.

Hagquist, C., Bruce, M., & Gustavsson, P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, *46*, 380-393.

Hobart, J. C., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment*, *13*(12), 1-177. doi:10.3310/hta13120

Hu, Y., Stewart-Brown, S., Twigg, L., & Weich, S. (2007). Can the 12-item General Health Questionnaire be used to measure positive mental health? *Psychological Medicine*, *37*, 1005-1013.

Hupert, F. A., & Whittington, J. E. (2004). Positive mental health in individuals and populations. In F. A. Hupert & N. Baylis (Eds.), *The science of well-being* (pp. 307-342). Oxford, England: Oxford University Press.

Jones, P. B. (2013). Adult mental health disorders and their age at onset. *British Journal of Psychiatry*, *202*(s54), S5-S10. doi:10.1192/bjp.bp.112.119164

Kafka, G., & Kozma, A. (2002). The construct validity of Ryff's scales of psychological well-being (SPWB) and their relationship to measures of subjective well-being. *Social Indicators Research*, *57*, 171-190.

Kearns, A., Whitley, E., Bond, L., Egan, M., & Tannahill, C. (2013). The psychosocial pathway to mental well-being at the local level: Investigating the effects of perceived relative position in a deprived area context. *Journal of Epidemiological Community Health*, *67*, 87-94.

Kessler, R., Angermeyer, M., Anthony, J., Graf, R., Demyttenaere, K., Gasquet, I., . . . Ustan, T. (2007). Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, *6*, 168-176.

Kessler, R., Heeringa, S., Lakoma, M. D., Petukhova, M., Rupp, A. E., & Schoenbaum, M. (2009). The individual-level and societal-level effects of mental disorders on earnings in the United States: Results from the National Comorbidity Survey Replication. *American Journal of Psychiatry*, *165*, 703-711.

Keyes, C. L. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, *43*, 207-222.

Keyes, C. L. (2007). Promoting and protecting mental health as flourishing: A complimentary strategy for improving national mental health. *American Psychologist*, *62*, 95-108.

Keyes, C. L., Dhingra, S. S., & Simoes, E. J. (2010). Change in levels of positive mental health as a predictor of future risk of mental illness. *American Journal of Public Health*, *100*, 2366-2371.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Lehtinen, V., Sohlman, B., & Kovess-Masfety, V. (2005). Level of positive mental health in the European Union: Results from the Eurobarometer 2002 survey. *Clinical Practice and Epidemiology in Mental Health*, *1*(1), 9. doi:10.1186/1745-0179-1-9

Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.

Lloyd, K., & Devine, P. (2012). Psychometric properties of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in Northern Ireland. *Journal of Mental Health*, *21*, 257-263.

López, M., Gabilondo, A., Codony, M., García-Forero, C., Vilagut, G., Castellví, P., . . . Alonso, J. (2013). Adaptation into Spanish of the Warwick Edinburgh Mental Wellbeing Scale (WEMWBS) and preliminary validation in a student sample. *Quality of Life Research*, *22*, 1099-1104.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 200-215.

Meijer, R. R., & Egberink, I. J. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, *72*, 589-607.

Perry, G., Presley-Cantrell, L., & Dhingra, S. (2010). Addressing mental health promotion in chronic disease prevention and health promotion. *American Journal of Public Health*, *100*, 2337-2339.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago, IL: University of Chicago Press.

Reeve, B., Hays, R., Bjorner, J., Cook, K. F., Crane, P., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Suppl. 1), S22-S31.

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52*, 141-166.

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, *10*, 424-437.

Spittlehouse, J. K., Vierck, E., Pearson, J. F., & Joyce, P. R. (2014). Temperament and character as determinants of well-being. *Comprehensive Psychiatry*, *55*, 1679-1687.

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & 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. doi:10.1186/1477-7525-7-15

Stochl, P., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied researchers. *BMC Medical Research Methodology*, *12*, 74. doi:10.1186/1477-7525-7-15

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., . . . Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, *5*, 63. doi:10.1186/1477-7525-5-63

Tennant, R., Joseph, S., & Stewart-Brown, S. (2007). The Affectometer 2: A measure of positive mental health in UK populations. *Quality Life Research*, *16*, 687-695.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(Suppl. 1), S22-S26.

Vaingankar, J., Subramaniam, M., Chong, S. A., Abdin, E., Edelen, M., Picco, L., . . . Sherbourne, C. (2011). The positive mental health instrument: Development and validation of a culturally relevant scale in a multi-ethnic Asian population. *Health and Quality of Life Outcomes*, *9*, 92. doi:10.1186/1477-7525-9-92

Ware, J. E., Snow, K. K., Kosinski, B., & Gandek, B. (1993). *SF-36 Health survey: Manual and interpretation guide*. Boston, MA: New England Medical Center.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.

Wilhelm, K. (2014). Gender and mental health. *Australian & New Zealand Journal of Psychiatry*, *48*, 603-605.

Wilkinson, R. G., & Pickett, K. E. (2006). Income inequality and population health: A review and explanation of the evidence. *Social Science & Medicine, 62*(7), 1768-1784.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, *22*, 33-53.

Wray, T. B., Dvorak, R. D., & Martin, S. L. (2013). Demographic and economic predictors of mental health problems and contact with treatment resources among adults in a low-income primary care setting. *Psychology, Health & Medicine*, *18*, 213-222.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33-45.

Zaporozhets, O., Fox, C. M., Beltyukova, S. A., Laux, J. M., Piazza, N. J., & Salyers, K. (2015). Refining change measure with the Rasch model. *Measurement and Evaluation in Counseling and Development*, *48*, 59-74.