Evaluating Item Discrimination Power of WHOQOL-BREF from an Item Response Model Perspectives

Ting Hsiang Lin · Grace Yao

Accepted: 27 May 2008 / Published online: 1 July 2008
© Springer Science+Business Media B.V. 2008

Abstract Quality of life (QOL) has become an important component of health. By using the methodology of psychometric theory, we examine the item properties of the WHOQOL-BRIEF. Samejima's graded response model with natural metrics of the logistic response function was fitted. The results showed items with negative natures were less discriminating. Items written in a specific way were more suitable to assess certain subgroups. The national items showed variation in discriminatory power. Questions measuring objective and specific issues performed worse than items assessing general aspects of the QOL.

Keywords WHOQOL-BREF · Item response theory · Item analysis · Quality of life

1 Introduction

While traditional measures of health have largely been based upon its physical aspects, i.e., mortality and morbidity, in recent years, the concept of “good health” has changed from the mere absence of disease to a complete functioning of both physical and mental activities as well as social well-being (World Health Organisation 1980). In other words, the perception of the overall quality of life (QOL) has become an important component of health. Thus, the healthcare scientists are becoming increasingly concerned with evaluating the QOL, which has been used in many fields as a tool to assess the individual's physical, psychological, and social functions. The World Health Organization (WHO) recognizes the importance of developing a QOL instrument, and has been working towards devising an assessment referred to as WHOQOL-100. This instrument consists of 100 items

T. H. Lin (✉)
Department of Statistics, National Taipei University, 67, Section 3,
Ming-Sheng East Road, Taipei 10433, Taiwan, ROC
e-mail: tinghlin@mail.ntpu.edu.tw

G. Yao
National Taiwan University, Taipei, Taiwan, ROC

representing 24 facets, which can be sub-divided into six major domains, namely, the environment, independence, spirituality, the psychological aspect, physical capacity, and social relationships. An alternative four-domain model is more simplified in that independence is combined with the physical domain, and spirituality is combined with the psychological domain. Due to practical issues such as the administrative burden and lengthy response time, the WHOQOL-100 has later been reduced to a shorter version referred to as WHOQOL-BREF, which comprises 26 items where each item measures a single facet. Since the QOL is conceptualized under the context of culture and value systems, hence, WHOQOL has been developed to be accepted and applied cross-culturally, and it has been translated into many different languages (Yao et al. 2002; Min et al. 2002; DeGirolamo et al. 2000).

When assessing QOL, researchers must be able to rely on instruments with desirable psychometric properties, including reliability, internal and external validity, and discriminatory power, while the items must also provide substantial information regarding the respondents. Both the reliability and validity of WHOQOL-BREF have been studied elsewhere, and satisfactory results have been reported in adapted versions of different languages (Bonomi et al. 2002; Norholm and Bech 2001). However, the psychometric property of the WHOQOL-BREF has only been examined as an entire scale or under several domains or subscales, and no studies have yet investigated the properties of each individual item. In this study, we therefore apply a model-based approach to instrument development referred to as item response theory (IRT) (Lord and Novick 1968; Embretson and Reise 2000; van der Linden and Hambleton 1996).

IRT has been widely used as an item development tool in education, and it has recently been applied in other areas including psychology or medicine. IRT consists of a family of mathematical functions that model a person's response to individual items. The response of a person to an item can be modeled as a function of the characteristics and properties of that item. The response of a person is assumed to depend upon factors called latent traits or attributes (the terms "trait" and "attribute" will be used interchangeably hereafter). A person's performance in regard to an item depends purely on the latent trait, and the relationship between the persons' response to that item and the trait can be modeled by means of a set of monotonic functions. These functions are commonly referred to as option characteristic functions (OCF) and they specify how the probabilities of responses to an item change as the levels of the trait increase or decrease. One advantage of IRT is that IRT possesses some procedures that are used to estimate model parameters for both persons and items, and IRT also performs other statistical tests to investigate how well the hypothetical models fit the data. The general form of IRT model is presented using a logistic function, and most parametric families can be presented as follows:

$$P(X = 1|\theta = t) = c + (1 - c) \frac{\exp[1.7a(\theta - b)]}{1 + \exp[1.7a(\theta - b)]}$$

$P(X = 1|\theta = t)$ represents the probability of a positive response, and θ represents the level of the latent trait, e.g., overall satisfaction regarding the QOL. The parameter b is a 'location parameter' and it refers to the inflection point where the item best discriminates the respondent's level of the trait in terms of its being above or below the threshold value. In other words, it refers to the trait level at which an individual has a 50% chance of endorsing the item. The lower-asymptote parameter, c , is referred to as a 'pseudo-guessing' parameter in the educational context. It assumes that there is a non-zero random guessing probability that low-trait respondents will answer an item correctly or endorse an item. In

non-educational settings, this parameter is also conceptually interpretable. For example, in medical studies, the parameter may represent a respondent who is pretending to be healthy or not depressed. The discrimination parameter, a , measures an item's strength in distinguishing respondents whose trait levels are above or below the threshold point. It may be interpreted as how strongly an item is related to the attribute measured by the scale, and it is related to the biserial item-scale correlation (van der Linden and Hambleton 1996). In the IRT models, the constant 1.7 is just a scaling factor in the logistic response function. By imposing some restrictions on the parameters, the model can be transformed into two less complex models (Lord 1980). When the parameter c is set to zero, the model is simplified to a 2-parameter model (2-P model) assuming that there is no guessing component, and all items have different levels of discrimination and item difficulties. When setting both $c = 0$ and $a = 1$ simultaneously, the model is a one parameter model (1-P model) assuming that all items have the same discriminatory power but different thresholds.

The three models mentioned above are more suitable for binary data with dichotomous responses, for example, if one answers questions correctly or agree/disagree with certain issues. For polytomous data, it depends on whether the data are nominal or ordinal. A series of models including partial credit model, rating scale model and graded response model are suitable for ordinal data, and nominal response model are applicable for nominal data (Wu and Zumbo 2007). When analyzing rating scale data like WHOQOL, Samejima's graded response model (Samejima 1967; Hambleton et al. 1991) is more appropriate than a dichotomous response model, where options are ordered along a continuum, as with Likert scales. For multiple-choice items, each category is represented by an option response function. According to the SGR model, the probability of selecting option k or a higher option on item i is:

$$P_{ik}(\theta) = \frac{\exp[a_i(\theta - b_{ik})]}{1 + \exp[a_i(\theta - b_{ik})]} - \frac{\exp[a_i(\theta - b_{i,k+1})]}{1 + \exp[a_i(\theta - b_{i,k+1})]}$$

In the model, k is the particular option chosen by the respondent ($k = 1$ to s , where s is the total number of options for item i). The item discrimination parameter, a_i , is assumed to be the same for all options within item i ; b_{ik} is the location parameter that varies from option to option. $P_{ik}(\theta)$ represents the probability of selecting option k for item i , and θ represents the value of the latent trait.

For those IRT models mentioned above, several assumptions are made. First, the responses can be described by a model with one latent dimension, i.e., only one dimension attribute or latent trait accounts for the responses. This assumption which is referred to as unidimensionality implies that the responses to different items are independent given the fixed level of latent traits. In other words, the underlying latent trait is the only factor affecting the response.

After taking the respondent's trait into account, no relationship exists between subjects' responses to different items. With the unidimensionality assumption, local independence can be obtained. Second, the latent trait θ is defined on a scale of minus infinity to plus infinity $(-\infty, \infty)$. The probability of a particular response to an item is a monotonous and nonlinear function of the trait. Third, the probability of a particular response given the underlying trait might be interpreted as the probability of that response for a randomly selected person from a well-defined population with the given attribute (Holland 1990). The probabilities are bounded between the values of 0 and 1.

The assumptions must be verified before an appropriate IRT model can be built. Several methods have been proposed for assessing unidimensionality and local independence,

including parametric and nonparametric approach. A commonly used method to assess unidimensionality is linear factor analysis. It can be accomplished by examining screen plots, the number of eigenvalues greater than 1, or comparing the ratio of the first eigenvalue to the second. Other nonlinear factor analysis approaches also have been proposed, for example, normal ogive harmonic analysis Robust method (NOHARM; Fraser 1988; McDonald 1967, 1997, 1999) and TESTFACT (Bock et al. 2003). Among these methods, modified parallel analysis (MPA; Drasgow and Lissak 1983) and DIMTEST (Stout 1987) were specifically developed for IRT models. Secondly, the assumption of local interdependence can be tested by examining if the off-diagonals of the variance covariance matrix after partialling out θ . If the value is close to zero, local independence can be assumed. If the scale is verified to be multidimensional, local independence automatically holds (Hambleton et al. 1991).

The objective of this study is to examine the individual items within the WHOQOL-BREF in terms of their performance based on IRT. A graded response model is applied to evaluate measurement properties across the latent trait continuum that is relevant to the overall satisfaction over the QOL. This result of this study will provide some insights for the future development of the QOL instrument.

2 Methods

The WHOQOL-BREF is one of the five components of the 2001 National Health Interview Survey (NHIS) in Taiwan, which is intended to provide nation-wide estimates for health conditions, health behaviors, and the usage of medical resources. The NHIS is conducted by face-to-face interviews with large cross-sectional sample surveys of the country's population. The sampling scheme used for the NHIS is a stratified multistage systematic sampling design. The 2001 NHIS is composed of five different questionnaires: the household survey, adults' survey, children's survey, the QOL (WHOQOL-BREF Taiwan version) questionnaire, and the youth risk behavior questionnaire. Personal interviews with structured questionnaires were used to collect data by trained interviewers at the subjects' residences.

The WHOQOL-BREF Taiwan version was developed in 1999 in compliance with WHO guidelines (World Health Organization 1993, 1994). A research team translated the WHOQOL items (facets) from English into the local language. The WHOQOL-BREF Taiwan version contains 26 items plus two additional items reflecting cultural importance. These two questions are: "Do you feel respected by others?" and "Are you usually able to get the things you like to eat?" representing two facets of respect and eating. The reliability and validity of the Taiwan version has already been reported by Yao et al. (2002), and the 2 to 4 weeks' test-retest reliability coefficients were found to range from 0.41 to 0.79 at the item/facet level. Content validity coefficients were in the range of 0.53 to 0.78 for item-domain correlations.

The respondents had to be between 20 and 65-years old, and were selected in the National Health Interview Survey before they volunteered to participate in the WHOQOL-BREF. The sample analyzed in this study must have some non-missing data on WHOQOL-BREF. A total of 13,060 individuals were included in this study. The demographic characteristics of this study sample are presented in Table 1. Forty-nine percent of the subjects in the sample were male, and their average age was 38.14 years with a standard deviation of 11.32. Nearly thirty percent of the respondents had received more than a high school education, and the average years of education received was 11.14 years. Some 66% of the subjects were married, 2.96% were divorced, 2.13% were widowed, and 27.44% were single.

Table 1 Demographic characteristics of the subject sample ($N = 13060$)

Variable	Levels	Percents
Gender	Male	50.63
	Female	49.37
Marital status	Married	63.08
	Married, live together sometimes	3.17
	Divorced	2.96
	Separated	0.34
	Widowed	2.13
	Unmarried, live together	0.60
	Single	27.44
	Other	0.28
Education	High school and below	70.56
	Above high school	29.44
	Mean	SD
Age	38.14	11.32
Education (years)	11.14	3.43

The negative items have been reverse coded so that all items in the scale indicate a positive direction, and 1 represents “not satisfied at all” or strongly disagree with the statement, and 5 represents “extremely satisfied” or strongly agree with the description of that item. There are three negatively worded items: item 3 (pain), item 4 (medical treatment), and item 26 (negative feelings). In order to consider the non-response effect in the model, the missing values were not imputed and included as is. A total of 84.74% of the respondents had complete data for all 28 items, 13.24% had one missing item, and 1.29% had two missing items. Less than 1% of the subjects had three or more missing items.

The software used in this study was PARSCALE (Muraki and Bock 1996). It is flexible for data analysis, and it performs various IRT analyses and scores individuals on the latent trait continuum. PARSCALE implements Samejima’s model for graded categories, and extends it to a rating scale such as that involving Likert-type data in which all items are rated with the same options. In this case, a common set of options boundaries is estimated, and location and discrimination parameters are provided for each item.

3 Results

The coefficient estimates of the 28 WHOOL-BREF items are presented in Table 2. As shown in Table 2, a considerable amount of variation in terms of discriminatory strength exists among all the items. The items with the highest discriminatory power are the general QOL (item 1), energy (item 10), activities of daily living (item 17), work (item 18), self-esteem (item 19), personal relationships (item 20), and social support (item 22). Items with the lowest discriminatory power are pain (item 3), medication (item 4), spirituality (item 6), the physical environment (item 9), leisure activities (item 14), the financial situation (item 19), negative feelings (item 26), and eating (item 28).

Table 2 WHOQOL-BREF items, item parameters, and descriptive statistics

No.	Item content	Facet	Item disc (Rank)	Means (SD)
1.	How would you rate your quality of life?	General	1.094 (2)	3.294 (0.658)
2.	How satisfied are you with your health?	General	0.922 (8)	3.391 (0.739)
3.	To what extent do you feel that physical pain prevents you from doing pain what you need to do?	Pain	0.474 (27)	4.023 (0.882)
4.	How much do you need any medical treatment to function in your daily life?	Medication	0.162 (28)	4.349 (0.905)
5.	How much do you enjoy life?	Positive feeling	0.650 (16)	2.699 (0.858)
6.	To what extent do you feel your life to be meaningful?	Spirit	0.636 (26)	3.519 (0.932)
7.	How well are you able to concentrate?	Thinking	0.812 (14)	3.401 (0.821)
8.	How safe do you feel in your daily life?	Physical safe	0.798 (15)	3.312 (0.853)
9.	How healthy is your physical environment?	Physical environment	0.661 (24)	3.115 (0.894)
10.	Do you have enough energy for everyday life?	Energy	0.986 (5)	3.366 (0.766)
11.	Are you able to accept your bodily appearance?	Body image	0.688 (20)	3.668 (0.830)
12.	Have you enough money to meet your needs?	Financial situation	0.671 (23)	2.813 (0.964)
13.	How available to you is the information that you need in your day-to-day life?	Information availability	0.722 (19)	3.304 (0.869)
14.	To what extent do you have the opportunity for leisure activities?	Leisure activities	0.645 (25)	2.946 (0.937)
15.	How well are you able to get around?	Mobility	0.727 (18)	3.918 (0.783)
16.	How satisfied are you with your sleep?	Sleep	0.738 (17)	3.405 (0.844)
17.	How satisfied are you with your ability	Activities of daily living	1.182 (1)	3.648 (0.674)
18.	How satisfied are you with your capacity for work?	Work	0.987 (6)	3.628 (0.726)
19.	How satisfied are you with yourself?	Self-esteem	1.052 (4)	3.655 (0.721)
20.	How satisfied are you with your personal relationships?	Personal relationship	0.986 (7)	3.588 (0.698)
21.	How satisfied are you with your sex life?	Sex	0.904 (9)	3.523 (0.741)
22.	How satisfied are you with the support you get from your friends?	Social support	1.064 (3)	3.670 (0.624)
23.	How satisfied are you with the conditions of your living place?	Home	0.833 (12)	3.540 (0.753)
24.	How satisfied are you with your access to health services?	Health service	0.818 (13)	3.419 (0.728)
25.	How satisfied are you with your transport?	Transportation	0.886 (11)	3.555 (0.701)
26.	How often do you have negative feelings such as blue mood, despair, anxiety, depression?	Negative feeling	0.679 (21)	3.458 (0.833)
27.	Do you feel respected by other?	Respect	0.892 (10)	3.316 (0.755)
28.	Are you usually able to get the things you like to eat?	Eating	0.677 (22)	3.711 (0.789)

The ORF is presented in Fig. 1. The ORF displays the response probability to each option separately as a function of the latent attribute. Ideally, an item with high discriminatory power will have different option response functions given a fixed level of theta. For

Fig. 1 Option characteristic function for individual WHOQOL-BREF items

example, for those subjects with higher theta values such as where theta equals 2 in item 1, the probabilities of observing response options 3, 4, and 5 are 0.05, 0.85, and 0.1, respectively. Conversely, the probabilities of observing response options 1 and 2 are quite low. The slope of the highest option (e.g., extremely satisfied, most healthy) increases monotonically with the latent trait, i.e., the more the subjects possess the attribute, the more likely it is that they will choose the response option. On the other hand, the lowest option (e.g., not satisfied at all, most unhealthy) is more likely to be selected by individuals without the attribute. Therefore, the curve of the low option decreases monotonically with the latent trait. In both cases, the curves of these two extreme options should be approximately S-shaped (or inverse S-shaped), and the reflection points should be around parameter b . For the middle options, the OCFs lie between the two extreme options and the distributions are somewhat bell-shaped or skewed bell-shaped.

Fig. 1 continued

4 Discussion

This study provides a demonstration of the applicability of IRT to the item selection/reduction in the QOL instrument. We are particularly interested in studying the psychometric properties of the WHOQOL-BREF and seeking subsets of optimal items that best reflect the underlying latent trait. The discrimination coefficients derived from IRT provide a direct comparison for evaluating the performance of the individual items.

There are several findings in these studies that are useful for future instrument development, including item selection or reduction. First, questions with reverse coding have poorer discriminatory power than un-recoded items. These include items stated in either negative wordings or phrased towards unhealthy or unhappy directions. For example, item 3 (pain) and 4 (medication) are ranked as 27 and 28th out of 28 items, and item 26 (negative feelings) is ranked as 21st. One possible explanation for this is that reverse coded

Fig. 1 continued

items provide different validity even if they have been recorded. The effect of wording has been examined by several studies. Zumbo et al. (1997) used item-total correlation and IRT models to test Roskam's (1985) conjecture and found inconsistent effects of wording and item. Another study conducted by Herche and Engelland (1996) also found that the use of reversed-polarity items led to some problems. A mixed use of positively and negatively worded items in the same scale can adversely affect measurement consistency. The reason for this is because of the degradation of scale dimensionality resulting from bias. Therefore, they recommended against using negatively worded items.

The second point is somewhat related to the first finding mentioned above. The descriptions of some items were so specific that they were more suitable for assessing a particular subgroup rather than a general population, and as a consequence these items also performed poorly. For example, both item 4 (medication) and item 3 (pain) provided little

Fig. 1 continued

information when assessing the QOL of a relatively healthy population in this case. Item 26 (negative feelings) is another example. This can be easily verified by examining the distributions of the response options. A total of 7.19% of the subjects answered that great or total pain hindered them from doing what they needed to do, and 5.54% reported that they had a great or extreme need for medical treatment. Given the skewed distribution of the responses, these items performed poorly in terms of distinguishing low versus high trait groups. We can draw the same conclusion by examining the OCF. It can be seen that the probabilities of selecting the pre-recoded “middle” options 2 through 4 (recoded as 4, 3, and 2) of item 4 are almost the same regardless of the level of the latent trait, and they are almost parallel to each other. Likewise, recoded options 1, 2 and 3 of item 3 are less distinguishable among themselves.

Third, the discriminatory strength of the two national items is quite different. The question as regards measuring being respected by others has above average discriminatory power (coefficient = 0.892, ranked as 10th) while the one assessing eating provides little information (coefficient = 0.677, ranked as 22nd). A study conducted by Skevington et al. (1999) examined the performance of 144 new WHOQOL national items. One of the facets in regard to eating and appetite was proposed in the Hong Kong version and five items were tested. These items were, “1. How is your appetite?” “2. How often are you able to eat the food you desire?” “3. Is the food you desire easily available to you?” “4. Do you enjoy the food you eat?” and “5. Are you able to eat any kind of food?”. The results showed that two items regarding food availability (items 2 and 3) performed worse than the other three and both items exhibited some weakness. In the same study, four items were proposed to measure a new facet of “being respected and accepted”. The item that asked if the respondent was satisfied with the respect received from others performed the best among the four. This is somewhat consistent with our findings, and it suggests that new items should be cross-validated from various perspectives.

Fourth, it turned out to be a surprise that general questions overall performed better than object-specific items. These specific items included assessing the spirituality

(item 6), the physical environment (item 9), body image (item 11), the financial situation (item 12), information availability (item 13), and leisure activities (item 14). On the other hand, more general questions such as item 1 (How would you rate your QOL?) or item 17 (How satisfied are you with your ability?) are more informative. We have not found similar results from previous studies. One possible explanation is that the respondents are more concerned with the physical, psychological and social domains than the environmental one since all the items fall into that category except items 6 and 11. However, this is just our speculation, and perhaps future research will help explain this issue.

There are some caveats that the reader should be aware of when interpreting these results. First, the performance of each item using the WHOQOL-BREF is not necessarily completely replicable in the other WHOQOL of different languages. Due to the lack of data, we are unable to make comparisons between the Taiwan and English versions or other languages, and we are uncertain how compatible the results will be. For example, a study (Allalouf et al. 1999) discussed the inconsistent results when translating a text from Hebrew into Russian, and several sources of the inconsistency were found. Kristjansson and colleagues also highlighted the importance of conceptual, semantic, syntactic, and experiential equivalences in interpreting scales or questionnaires of different languages. Extra caution should be used not only in translation but also the culture difference and intention of the items (Kristjansson et al. 2003). These include changes in context, word difficulty and item format as well as differences in cultural relevance. We can only conclude that the performance of an item is stable for different languages only under two conditions. First, the translation is compatible and precise enough so that there is no discrepancy between two languages. Second, the discriminatory power of the item is so strong that it is quite robust under all cultures.

Second, this study's population generally can be classified into several subgroups (e.g., by gender or by race), and inferences should be equally valid for members of all subgroups. If there exists differential validity of a given level of latent trait for some subgroup, there might exist some bias in the scale (Cole and Moss 1989). There are various sources of bias, and one possibility is the presence of biased items. In terms of IRT, the term differential item functioning (DIF) is used for item bias and there are many methods that have been proposed to detect DIF (Camilli 1992; Kim et al. 1995; Holland and Thayer 1988; Zumbo 2007). We will not provide the details here since the topic is beyond the scope of this paper. However, it is important to further look into whether the WHOQOL instrument contains item bias.

The IRT model provides many advantages and strengths for instrument development and item selection. These features of IRT analysis allow researchers to evaluate the quality and properties of individual items, to assess the measurement scale, and to determine a respondent's latent trait level. It is particularly useful when assessing some hard to observe or abstract constructs of individuals, such as perception of the QOL. To sum up, a more precise and valid measure of the QOL can be constructed using the IRT approach. It is important not only to develop a comprehensive but also short and easily administered QOL instrument, which will have a significant impact for clinical or research purposes.

Acknowledgements The authors thank Bureau of Health Promotion, Department of Health and National Health Research Institute in Taiwan, for providing the data.

References

- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36, 185–198. doi:[10.1111/j.1745-3984.1999.tb00553.x](https://doi.org/10.1111/j.1745-3984.1999.tb00553.x).
- Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4.0 [Computer software and manual]. Lincolnwood, IL: Scientific Software International.
- Bonomi, A. E., Patrick, D. L., Bushnell, D. M., & Martin, M. (2000). Validation of the United States' version of the World Health Organization Quality of Life (WHOQOL) instrument. *Journal of Clinical Epidemiology*, 53, 1–12. doi:[10.1016/S0895-4356\(99\)00123-7](https://doi.org/10.1016/S0895-4356(99)00123-7).
- Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, 16, 129–147. doi:[10.1177/014662169201600203](https://doi.org/10.1177/014662169201600203).
- Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education.
- DeGirolamo, G., Rucci, P., Scocco, P., Becchi, A., Coppa, F., D'Addario, A., et al. (2000). Quality of life assessment: Validation of the Italian version of the WHOQOL-Brief. *Epidemiologia e Psichiatria Sociale*, 9, 45–55.
- Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *The Journal of Applied Psychology*, 68, 363–373. doi:[10.1037/0021-9010.68.3.363](https://doi.org/10.1037/0021-9010.68.3.363).
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science*, 24, 366–374. doi:[10.1177/0092070396244007](https://doi.org/10.1177/0092070396244007).
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577–601. doi:[10.1007/BF02294609](https://doi.org/10.1007/BF02294609).
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.
- Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261–276. doi:[10.1111/j.1745-3984.1995.tb00466.x](https://doi.org/10.1111/j.1745-3984.1995.tb00466.x).
- Kristjansson, E. A., Desrochers, A., & Zumbo, B. D. (2003). Translating and adapting measurement instruments for cross-cultural research: A guide for practitioners. *Canadian Journal of Nursing Research*, 35, 127–142.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum: Hillsdale.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores (with contributions by A. Birnbaum)*. Reading, MA: Addison-Wesley.
- McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric monographs, no. 15.
- McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Min, S. K., Kim, K. I., Lee, C. I., Jung, Y. C., Suh, S. Y., & Kim, D. K. (2002). Development of the Korean versions of WHO quality of life scale and WHOQOL-BREF. *Quality of Life Research*, 11, 593–600. doi:[10.1023/A:1016351406336](https://doi.org/10.1023/A:1016351406336).
- Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Version 3*. Chicago: Scientific Software International Inc.
- Norholm, V., & Bech, P. (2001). The WHO Quality of Life (WHOQOL) questionnaire: Danish validation study. *Nordic Journal of Psychiatry*, 55, 229–235. doi:[10.1080/080394801681019075](https://doi.org/10.1080/080394801681019075).
- Roskam, E. E. (1985). Current issues in item response theory. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3–19). Amsterdam: North Holland.
- Samejima, F. (1967). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 34(Supplement), 100–114.
- Skevington, S. M., Bradshaw, J., & Saxenab, S. (1999). Selecting national items for the WHOQOL: Conceptual and psychometric considerations. *Social Science & Medicine*, 48, 473–487. doi:[10.1016/S0277-9536\(98\)00355-4](https://doi.org/10.1016/S0277-9536(98)00355-4).

- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589–617. doi:[10.1007/BF02294821](https://doi.org/10.1007/BF02294821).
- van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer.
- World Health Organization. (1980). *International classification of impairments, disabilities and handicaps*. Geneva: WHO.
- World Health Organization. (1993). WHOQOL study protocol. Geneva: WHO (MNH/PSF/93.9).
- World Health Organization. (1994). WHOQOL protocol for new centers. Geneva: WHO (MNH/PSF/94.4).
- Wu, A. D., & Zumbo, B. D. (2007). Thinking about item response theory from a logistic regression perspective: A focus on polytomous models. In S. S. Sawilowsky (Ed.), *Real data analysis. AERA, Educational Statisticians Book Series* (pp. 241–269). Greenwich, CT: Information Age Publishing, Inc.
- Yao, G., Chung, C. W., Yu, C. F., & Wang, J. D. (2002). Development and verification of validity and reliability of the WHOQOL-BREF Taiwan version. *Journal of the Formosan Medical Association*, 101, 342–351.
- Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.
- Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement*, 57, 963–969.