

Applying item response theory modelling for evaluating questionnaire item and scale properties

Bryce B. Reeve and Peter Fayers

Developing a health-related quality of life (HRQoL) questionnaire that is psychometrically sound and captures the full burden of disease or treatment upon a person, requires a thorough questionnaire development process that integrates survey development tools from the qualitative sciences, cognitive aspects of survey methodology, and the field of psychometrics. This chapter describes the role of item response theory (IRT) models in questionnaire evaluation and development. For multi-item scales, IRT models provide a clear picture of the performance of each item (question) in the scale and how the scale functions overall for measuring the construct of interest in the study population. IRT methods can lead to short reliable questionnaires that are tailored to the population of interest.

Item response theory models

IRT refers to a set of mathematical models that describe, in probabilistic terms, the relationship between a person's response to a survey question and his or her level of the 'latent variable' being measured by the scale. This latent variable is usually a hypothetical construct, trait, domain, or ability, which is postulated to exist but cannot be directly measured by a single observable variable or item. Instead, it is indirectly measured using multiple items or questions in a multi-item scale. The underlying latent variable, expressed mathematically by the Greek letter theta (θ), may be any measurable construct, such as mental health, fatigue, or physical functioning. The person's level on this construct is assumed to be the only factor that accounts for their response to each item in a scale. For example, a person with high levels of depression will have a high probability of responding that 'most of the time' they 'felt downhearted and blue'. Someone with hardly any depression is more likely to respond 'little of the time' or 'none of the time'. Provided as a reference guide, Table 1.5.1 summarizes the terminology used in this and other chapters.

Table 1.5.1 Key terms and definitions

Term	Definition
Category response curve (CRC)	See item characteristic curve (ICC).
Classical test theory (CTT)	Traditional psychometric methods such as factor analysis and Cronbach's α , in contrast to IRT.
Dichotomous (binary) response categories	An item having two response categories such as yes/no, true/false, or agree/disagree.
Discrimination parameter (a , α) – slope	IRT model item parameter that indicates the strength of the relationship between an item and the measured construct. Also, the parameter indicates how well an item discriminates between respondents below and above the item threshold parameter, as indicated by the slope of the ICCs.
Information function/curve	Indicates the range over θ for which an item or scale is most useful (reliable) for measuring persons' levels.
Item	A question in a scale.
Item characteristic curve (ICC)	Models the probabilistic relationship between a person's response to each category for an item and their level on the underlying construct (θ). Also called category response curve (CRC).
Item-total correlation	Measure of the relationship between an item and the total score from the set of items within the scale. Higher correlations indicate a stronger relationship between the item and scale score.
Local independence assumption	Once you control for the dominant factor influencing a person's response to an item, there should be no significant association among the item responses.
Polytomous response categories	An item having two or more response categories. For example, a 5-point Likert type scale.
Scale	Consists of multiple items that measure a single domain such as fatigue.
Standard error of measurement (SEM)	Describes an expected observed score fluctuation due to error in the measurement tool. Standard deviation of error about an estimated score.
Theta (θ)	Unobservable construct (or latent variable) being measured by a scale.
Threshold parameter (b , β) – difficulty, location	IRT model item parameter that indicates the severity or difficulty of an item response. The location along the θ -continuum of the item response categories.
Unidimensionality assumption	Assumes that one underlying (or dominant) factor accounts for a person's response to a question within a scale.

Traditional approaches to measurement scales (i.e., classical test theory) are based on *averages* or *simple summation* of the multiple items. In contrast, IRT models are founded on the *probability* that a person will make a particular response according to their level of the underlying latent variable. Thus IRT is analogous to logistic regression, which is widely used in medical statistics. The objective is to model each item by estimating the properties describing its performance. The relationship between a person's response to an item and the latent variable is expressed by 'item characteristic curves' (ICCs), also called category response curves (CRCs).

Figure 1.5.1 presents ICCs for the dichotomous item: 'As a result of any emotional problems, have you accomplished less than you would like?' This question was taken from the role-emotional subscale of the RAND-36/SF-36 instrument (Hays and Morales 2001; Ware and Sherbourne 1992) and combined with the other items in the scale along with items in the vitality, social functioning, and mental health subscales to form a 14-item "mental health" scale. Response data collected from 888 breast cancer survivors approximately 24 months after breast cancer diagnosis were used for the illustrations in this chapter. The latent variable "mental health" is represented by θ along the horizontal x-axis. Individuals with poor mental health are to the left on this axis, while people with good mental health are to the right. Numbers on the θ -axis are expressed in standardized units, and in our examples the population mental health has been standardized to have zero mean and standard deviation of one. Thus a mental health score of $\hat{\theta} = -2.0$ indicates that the person lies two standard deviations below the population mean. The vertical axis indicates the probability that a person will select one of the item's response categories. The two ICCs in Figure 1.5.1 indicate that the

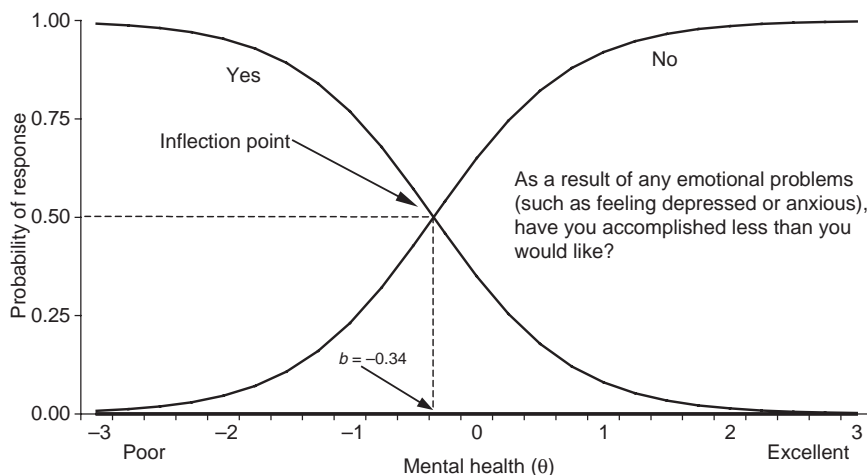


Fig. 1.5.1 IRT item characteristic curves (ICCs) for the mental health item indicated above. The ICC models the probability of a person to endorse one of the response categories (Yes or No) conditional on their mental health level. The difficulty parameter (b) indicates the level of mental health necessary for a person to have a 50% for endorsing the response category.

probability of responding ‘yes’ or ‘no’ to the item asking ‘have you accomplished less than you would like?’ depends on the respondent’s level of mental health. Poor mental health leads to a high probability of selecting ‘yes,’ and high mental health to ‘no’.

The ICCs in Figure 1 are represented by logistic curves that model the probability P that

$$P(X_i = 1 | \theta, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \quad (\text{Equation 1})$$

a person will respond ‘no’ (for the curve labelled ‘no’) to this item. This is a function of the respondent’s level of mental health (θ), the relationship (a) of the item to the measured construct, and the severity or ‘threshold’ (b) of the item in the scale. In IRT, a and b are commonly referred to as item discrimination and item difficulty (or threshold) parameters, respectively. Since P in Equation 1 is the probability of responding ‘no’, the equation for the responding ‘yes’ is just $1 - P$, as illustrated by the curve labelled ‘yes’ in Figure 1.5.1.

The item threshold (difficulty, or severity level), b , is the point on the latent scale where a person has a 50 percent chance of responding ‘no’ to the item. In Figure 1.5.1, the item’s threshold value is $b = -0.34$, which is close to 0.0, indicating that people with mental health levels near the population mean ($\theta = 0$) are equally likely to respond ‘no’ or ‘yes’ to this question. The intersection point of the ICCs, identified by the threshold parameter, locates where an item is optimal for differentiating respondents’ mental health levels around that point. An item with low threshold parameters is optimal for differentiating among persons with poor mental health functioning and vice versa. The threshold parameter varies for each item in a scale, and estimated parameters can be compared to select items to measure different parts of the mental health continuum.

The discrimination or slope parameter (a) in Equation 1 describes the strength of an item’s ability to differentiate among people at different levels along the trait continuum. An item optimally discriminates among respondents who are near the item’s threshold b . The discrimination parameter typically ranges between 0.5 and 2.5 in value. The slope is measured at the steepest point (the inflection point), and is $a = 1.82$ in Figure 1.5.1. The larger the a parameter, the steeper the slope and the more effective the item. Steep slopes, where the ICC increases relatively rapidly, indicate that small changes in the latent variable lead to large changes in item-endorsement probabilities. Items with large a parameters contribute most to determining a person’s $\hat{\theta}$ score. For example, ‘Did you feel tired?’ had a higher discrimination parameter ($a = 2.19$) and item-total correlation ($r = 0.68$) than the item ‘Have you been a very nervous person?’ ($a = 1.05$, $r = 0.48$, respectively). Thus the slope of the ICC for the ‘tired’ item will be steeper than the slope of the ‘nervous’ item, thus ‘tired’ has better discrimination. A slope of zero represents a horizontal ICC, indicating that irrespective of the person’s level there is always a 50:50 chance of responding ‘yes’ or ‘no’ – an entirely uninformative item.

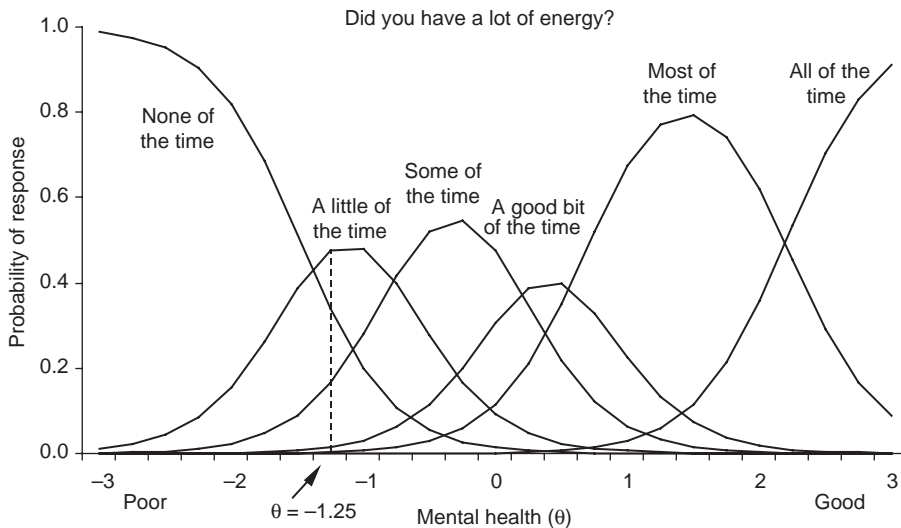


Fig. 1.5.2 IRT category response curves for the mental health item indicated above. There is a curve for each of the item's response options. The vertical dashed line indicates the probability for person 1.25 standard deviations below the population mean to respond to each of the six response categories.

IRT can also model items with more than two response options. The item 'Did you have a lot of energy?' has six response categories, and the ICCs are presented in Figure 1.5.2. Several IRT models are available for 'polytomous' responses, and the Graded Response Model (Samejima 1969) chosen for this data analysis estimated one item discrimination parameter and five threshold parameters (the number of categories minus one) to represent the location along the θ -scale where the probability exceeds 50 per cent that the response is in the associated category or higher category. In Figure 1.5.2, people with very poor mental health (e.g. $\theta < -2$) have a high probability of answering 'none of the time'. A person at $\theta = -1.25$ (indicated by a vertical dashed line) has a 34 per cent probability of endorsing high energy 'none of the time', 48 per cent probability of 'little of the time', 17 per cent probability of 'some of the time', and is unlikely to respond 'a good bit of the time', 'most of the time' or 'all the time'. Moving to the right along the θ -axis, people with better mental health will endorse response categories associated with better health.

Information functions

Information functions indicate the range over θ where an item or scale is best at discriminating among individuals. Higher information denotes more precision (or reliability). The shape of an *item* information function is defined by the item discrimination, a , and threshold, b , parameters. The threshold determines the location of the information function on the horizontal θ -axis. Thus threshold parameters evaluate how well a particular item matches levels of the construct being studied;

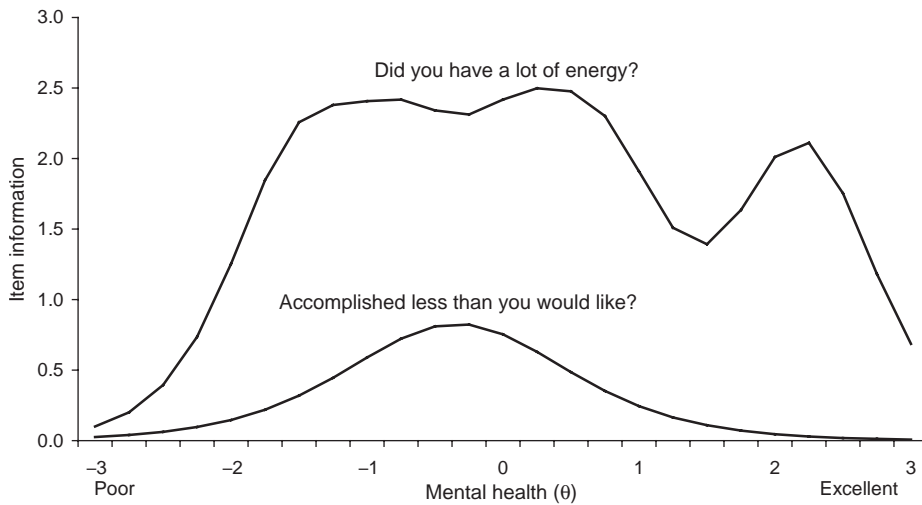


Fig. 1.5.3 IRT item information curves for the two mental health items indicated above. Each curve describes the range over θ for which the item is most useful (precise) for measuring persons' mental health levels.

ideally items should be well spaced across the continuum. Information magnitude is indicated on the vertical axis, and items with high discrimination have the most peaked information function because higher discrimination means the item can better differentiate among individuals who lie near the threshold value. Figure 1.5.3 presents item information functions for the two items of Figures 1.5.1 and 1.5.2. The information function for the 6-category 'energy' item is broader than that for the two-category 'accomplished less' item, because the greater number of response categories allows broader coverage of the continuum. Also, 'energy' is more peaked than 'accomplished less', indicating that the 'energy' item contributes more precision to the measurement of mental health.

Item information functions can identify items that perform well or poorly. Low information for one item may indicate that the item: (1) measures something different from other items in the scale, (2) is poorly worded and needs to be rewritten, (3) is too complex for the respondents, or (4) is placed out of context in the questionnaire.

The individual item information functions can be summed across all items in the scale to form the 'scale information function'. Figure 1.5.4 shows this for the 14-item RAND-36 Mental Health scale. Information magnitude and the associated reliability ($r = 1 - 1/\text{information}$) are shown. The scale is reliable ($r > 0.80$) for measuring mental health across the full continuum. The function is peaked at the lower end of the scale, indicating that poor mental health is measured with most precision. Reliability decreases when measuring excellent mental health.

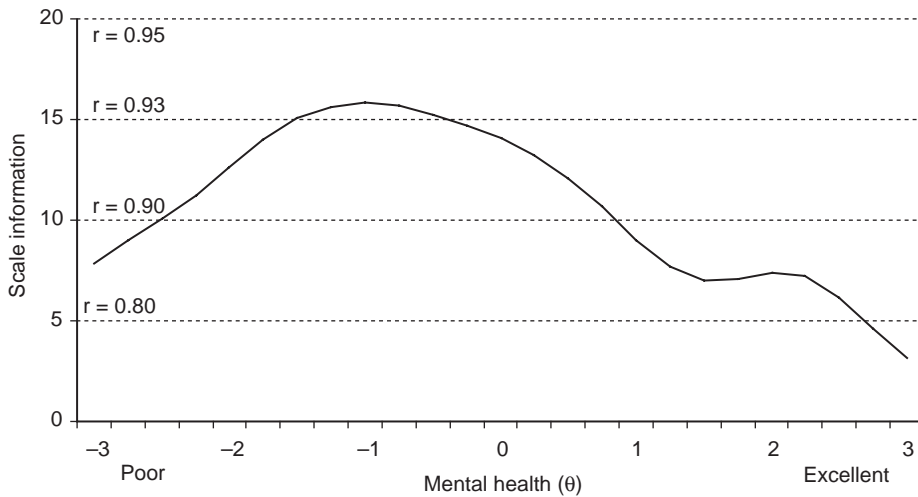


Fig. 1.5.4 IRT scale information curve for the mental health scale. The curve indicates the range over θ for which the scale is most informative (precise) for measuring persons' mental health levels. Horizontal dashed lines indicate the approximate level of reliability associated with different information magnitudes.

The standard error of measurement is $SEM = 1/\sqrt{\text{information}}$, and varies according to θ . For a person at $\theta = 1.5$ (a person 1.5 standard deviations above the mean score) the information is 7, and thus the reliability is 0.86 and the SEM is 0.38.

Common IRT models

IRT models can handle unidimensional or multidimensional data, binary or polytomous responses, and ordered or unordered responses. Table 1.5.2 presents the models most frequently used in HRQoL assessment. Two of these are for dichotomous items (e.g. yes/no or agree/disagree), whereas the remainder are polytomous models suitable for more than two response categories. All these models are unidimensional and designed to measure a single construct.

The so-called Rasch models are identified with asterisks in Table 1.5.1. Non-Rasch models (also called two-parameter models) estimate a discrimination (slope) parameter for each item, suggesting that items be differentially weighted with regard to the underlying construct, whereas Rasch models assume equal discrimination. Equation 1 represents the two-parameter logistic IRT model for dichotomous response data, with discrimination parameter a_i that varies from item to item as indicated by the subscript i . The simple Rasch model is obtained if all a_i are constrained to be equal.

The simplicity of the Rasch model confers several advantages. Any two items' threshold parameters can be compared, independent of the group of subjects being surveyed (specific objectivity), and any two persons' scores can be compared irrespective of the particular subset of items being administered. Also, fitting Rasch models is possible

Table 1.5.2 Commonly-used item response theory (IRT) models in HRQoL assessment

IRT Model	Item Response Format	Model Characteristics
Rasch Model*/One Parameter Logistic Model	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
Two Parameter Logistic Model	Dichotomous	Discrimination and threshold parameters vary across items.
Graded Response Model	Polytomous	Ordered responses. Discrimination varies across items.
Nominal Model	Polytomous	No prespecified item response order. Discrimination varies across items.
Partial Credit Model*	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale Model*	Polytomous	Discrimination equal across items. Distance between item threshold steps equal across items.
Generalized Partial Credit Model	Polytomous	Generalization of the Partial Credit Model that allows discrimination to vary across items.

* Models belonging to the family of Rasch models.

with smaller sample sizes. However, these properties only hold if the model fits the data adequately, and frequently the simple Rasch model is unrealistic.

IRT model assumptions

The parametric, unidimensional IRT models described above make three key assumptions about the data: (1) unidimensionality, (2) local independence, and (3) that the IRT model fits the data. It is important that these assumptions be evaluated. However, IRT models are robust to minor violations and no real data ever meet the assumptions perfectly.

Unidimensionality posits that the set of items measure a single continuous latent construct θ . In other words, a person's level on this single construct gives rise to that person's item responses. This assumption does not preclude a set of items from having a number of minor dimensions (subscales), but does assume that one dominant dimension suffices to explain the underlying structure. Scale dimensionality can be evaluated by factor analysis of the item responses. If multidimensionality is indicated by factor analysis and supported by clinical theory, it may be appropriate to divide the scale into subscales. Multidimensional IRT models do exist, but are complex.

Local independence means that if θ is held constant there should be no association among the item responses. Violation of this assumption may result in biased parameter

estimates, leading to erroneous decisions when selecting items for scale construction. Local independence can be evaluated by examining the residual correlation matrices to look for systematic error among item clusters that may indicate a violation of the assumption. The impact of local dependence can be explored by observing how the IRT item parameters and person scores change when one of the locally dependent items are dropped.

Model fit can be examined at both the item and person level, to determine whether the estimated item and person parameters can reproduce the observed item responses (Reise in press). Graphical and empirical approaches can evaluate item fit (e.g., Hambleton *et al.* 2000; Orlando and Thissen 2000) and person fit (Embretson and Reise 2000). Currently, no standard set of fit indices is universally accepted, and a number of different indices are reported by IRT software. Since IRT models are probabilistic models, most fit indices measure deviations between the predicted and observed response-frequencies. The citations above describe many types of residual analyses that can be used to evaluate model fit.

Applying IRT to evaluate item and scale properties and to suggest scale improvement

Item threshold levels and discrimination can be plotted as ICCs and information curves, to show the contribution of each question to the scale. This enables evaluation of how well an item performs in terms of its relevance or contribution for measuring the underlying construct, the level of the underlying construct targeted by the question, the possible redundancy of the item relative to other items in the scale, and the appropriateness of the response categories. Item-level information can be combined to provide pictures of how well the scale performs in terms of breadth and depth of coverage across the construct. IRT models provide a powerful tool for development of scales that are short, reliable, and targeted towards their study population.

Relevance and difficulty of the item content

Multi-item scales may measure latent constructs such as mental health, fatigue or patient satisfaction with medical care. In such scales not all items will be equally strongly associated with the underlying factor. The stronger the relationship, the better that item is for estimating a person's score. There are a number of ways to evaluate the relevance of items. Applying classical test theory (CTT), one would look at an item-score's correlation with the total scale score. In IRT, an item's relationship with the underlying construct is reflected in the discrimination parameter.

Table 1.5.3 presents CTT statistics and IRT parameter estimates for the 14 items making up the Mental Health Summary Scale. The column labelled 'item-total correlation' shows the different item-weights on the mental health construct. Questions such as 'Did you have a lot of energy?' and 'Have you felt downhearted and blue?' have high

Table 1.5.3 Mental health item properties

Mental Health Questions	CTT Statistics		IRT Model Item Properties						
	Mean (SD)	Item-total correlation	Alpha if item deleted	a	b1	b2	b3	b4	b5
Did you feel full of pep?	3.50 (1.32)	0.71	0.89	2.67	-1.64	-0.89	0.06	0.63	2.31
Did you have a lot of energy?	3.41 (1.35)	0.73	0.89	2.90	-1.48	-0.74	0.11	0.70	2.20
Did you feel worn out?	4.29 (1.24)	0.67	0.89	2.04	-2.62	-1.58	-1.01	0.04	1.35
Did you feel tired?	3.92 (1.20)	0.68	0.89	2.19	-2.29	-1.35	-0.65	0.49	2.03
What extent have your physical/emotional problems interfered with social activities?	4.25 (1.03)	0.64	0.89	2.11	-2.8	-1.79	-1.03	-0.22	
How much time have physical/emotional problems interfered with social activities?	4.18 (1.04)	0.65	0.89	2.04	-2.85	-1.90	-0.85	-0.09	
Cut down the amount of time you spent on work or other activities?	1.80 (0.40)	0.53	0.90	1.96	-1.08				
Accomplished less than you would like?	1.60 (0.49)	0.55	0.90	1.82	-0.34				
Didn't do work or other activities as carefully as usual?	1.76 (0.43)	0.50	0.90	1.65	-0.98				
Have you been a very nervous person?	4.97 (1.13)	0.48	0.90	1.05	-4.89	-3.31	-2.53	-1.03	0.48
Have you felt so down in the dumps that nothing could cheer you up?	5.40 (1.00)	0.61	0.90	1.71	-3.43	-2.87	-2.46	-1.41	-0.49
Have you felt calm and peaceful?	3.93 (1.25)	0.63	0.89	1.72	-2.56	-1.49	-0.52	0.24	2.51
Have you felt downhearted and blue?	4.97 (1.10)	0.67	0.89	1.73	-3.31	-2.51	-1.91	-0.78	0.43
Have you been a happy person?	4.48 (1.10)	0.59	0.90	1.62	-3.55	-2.33	-1.17	-0.44	1.69

Note: Some of the item wordings have been cut to minimize table space. IRT item parameter estimates (a = discrimination parameter; b = threshold parameter) were generated using Samejima's (1969) Graded Response Model and using the IRT software MULTILOG (Thissen 1991). All items have been scored (or reverse-scored) so higher scores reflect better mental health.

item-total correlations ($r = 0.73$ and $r = 0.67$, respectively), and are helpful in defining the underlying construct. The question, 'Have you been a very nervous person?' has the lowest item-total correlation ($r = 0.48$). This same pattern of relationships can be observed in Table 1.5.3 when looking at the IRT discrimination (a) parameters estimated by Samejima's (1969) Graded Response Model. These relationships appear intuitively plausible given the phrasing of the items.

Not only is the relevance or discrimination of the item important, but also its *difficulty* or *location*. 'Difficulty' is borrowed from educational assessment, where the goal is to match item difficulty with student ability. For example, nothing is learned about six-grade students' math ability if they are given an easy math question that all can answer, or a difficult math problem that none can solve. Likewise, one learns little about a healthy person's mental state with a question like 'Do you have suicidal thoughts?' because most people would answer 'no'. However, that question becomes informative for people with high levels of depression.

In CTT, item difficulty is measured by mean scores. Table 1.5.3 presents mean scores for the 14 mental health items. These indicate item 'severity', with low means, indicating poor mental health. The question 'Have you felt so down in the dumps that nothing could cheer you up?' has a high mean score (5.40) corresponding to the response categories 'none-' or 'a little' of the time, which indicates the sample is mentally healthy. In IRT, item difficulty is reflected by the threshold parameters (b); this item has six response levels and thus five difficulty-parameters b_1 to b_5 . The threshold parameters for this item are all negative, from -3.43 to -0.49 , indicating that the item functions best when measuring people with poor mental health. Figure 1.5.5 summarizes this clearly – the dashed information curve shows 'down in the dumps' to be informative when measuring people with poor mental health, but uninformative for others.

Both item relevance (discrimination) and location (difficulty) are important features in determining the best items for a particular study population. Figure 1.5.5 provides two other item-information curves. To measure a person with poor mental health, the 'down in the dumps' question is most informative. To measure someone with average levels of mental health, the question, 'Have you accomplished less than you would like?' is the optimal choice. For someone with good to excellent mental health, the first two items are not appropriate but the question 'Have you been a happy person?' is the best choice. IRT modelling facilitates the task of picking questions that match items to the study population.

Evaluating appropriateness of response categories

One decision when developing a questionnaire is the choice of the item response options, from a simple dichotomous 'yes' or 'no' to a seven-point or more response scale. If pilot or other prior data are available, CTT and IRT can provide helpful information. In CTT, one can check response frequencies for under or over utilized categories, although this may lead to erroneous decisions. For example, the response categories and associated response frequencies for the question, 'Did you feel worn out?' were: 'all the time' 24; 'most of the time' 74; 'a good bit of the time' 89; 'some of the time' 266; 'a little of the time' 304; and

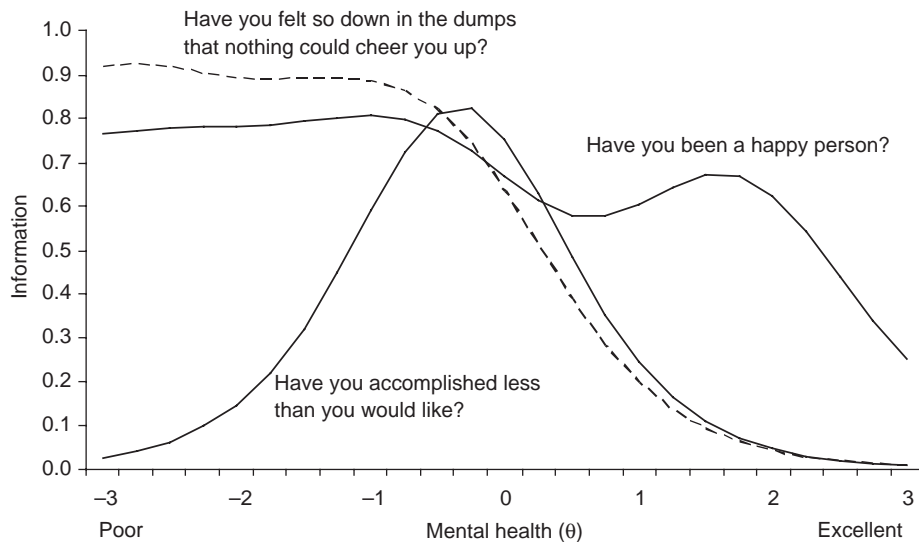


Fig. 1.5.5 IRT information curves for three items in the mental health scale. The ‘accomplished less’ question has two response options (note the narrow curve), and the other two items have six response categories, thus a broader curve.

‘none of the time’ 131. One might conclude that ‘all of the time’ and ‘most of the time’ were the less informative categories, but let’s look at the IRT results.

In contrast to Figure 1.5.2 which showed an ideal picture of a well functioning item with spread out response categories, Figure 1.5.6 presents the ICCs for ‘feeling worn out’. The category ‘a good bit of the time’ is overshadowed by ‘most of the time’ and ‘some of the time’; nowhere along the mental health continuum was this response category more likely to be chosen than other options. Thus ‘a good bit of the time’ could be dropped. The first two response categories should be retained, despite the results from CTT analysis, because they measure poor mental health functioning.

Finally, ICCs for the question, ‘Have you felt so down in the dumps that nothing could cheer you up?’ are presented in Figure 1.5.7. The ICCs show that only ‘some of the time’, ‘a little of the time’ and ‘none of the time’ are utilized by this sample. When an ICC shows one response category covering a large area of the θ -continuum, additional categories may be desirable.

Evaluating item redundancy

CTT attempts to increase reliability by lengthening multi-item scales. Often this results in questionnaires with items that are redundant in content but different in phrasing. For example, the first four items in Table 1.5.3 belong to the Vitality subscale. Not surprisingly, the items for ‘pep’, ‘energy’, (not) ‘worn out’ and (not) ‘tired’ are highly correlated and have very high internal consistency (α -reliability = 0.89). However, short questionnaires are strongly preferred for health outcomes populations.

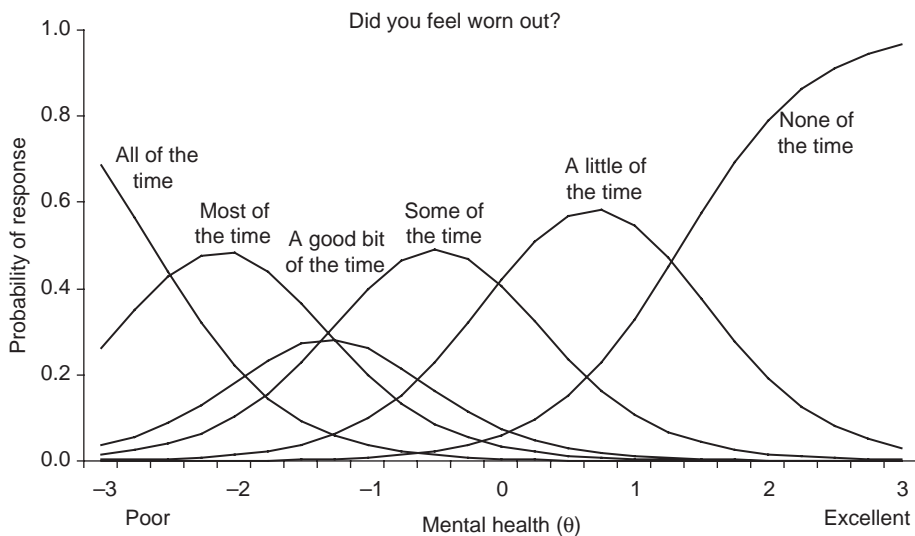


Fig. 1.5.6 IRT category response curves for the mental health item indicated above. The 'a good bit of the time' response is overshadowed by its neighbor categories suggesting this option may need to be considered to be dropped in future revisions of the scale.

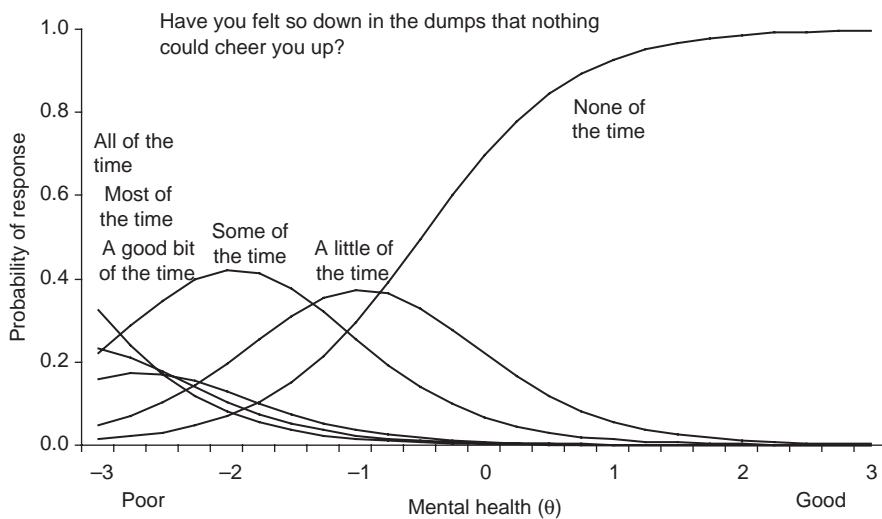


Fig. 1.5.7 IRT category response curves for the mental health item indicated above. Most respondents endorse 'none of the time' for this 'down in the dumps' question. Responses to the other five response categories suggest poor mental health.

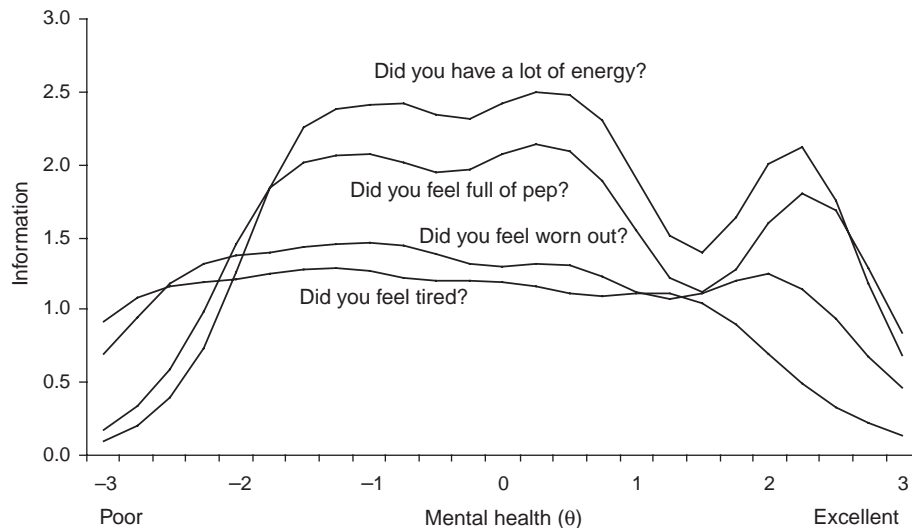


Fig. 1.5.8 IRT item information curves for four vitality questions in the mental health scale.

The item information curves, in Figure 1.5.8, indicate content redundancy for these two item pairs. The pair of curves for the ‘energy’ and ‘pep’ have identical patterns, as do those for the ‘tired’ and ‘worn out’; they provide redundant information. Little content information is lost if the questionnaire is shortened by deleting the ‘pep’ and ‘worn out’ items. While the redundancy among these items is also apparent from reading the item content, there are times when two or more seemingly very different items can still occupy the same informational space, and the developer may wish to remove one of them to shorten the questionnaire.

Evaluating content equivalence – differential item functioning

Questions should be equally applicable to all targeted populations. Thus a lot of care is taken when instruments are translated to other languages. However, despite linguistic equivalence, populations may give culturally different responses. In a depression questionnaire, Azocar *et al.* (2001) found that a Latino population endorsed ‘I feel like crying’ more than an Anglo population, because Latinos regard crying as socially acceptable behaviour. This resulted in Latinos receiving a higher average depression score than Anglos. This is known as differential item functioning (DIF).

DIF occurs whenever one group consistently responds differently to an item than another group. In other words, respondents with similar levels of θ have different probability of responding to an item according to their population membership. Scales containing such items have reduced validity for between-group comparisons because their scores are influenced by a variety of attributes other than those intended.

IRT provides an attractive framework for identifying items with DIF. Item characteristic curves (ICCs) can be estimated separately for each group. Differences between the ICCs indicate that the probabilities of item endorsement vary according to group membership.

DIF detection and interpretation is discussed in greater detail in the chapter by Groenvold and Petersen. DIF analysis has been used to detect item equivalence across racial, gender, cultural, and treatment groups; and between two administration modes (e.g. telephone versus self-administered) and two translated language versions (Azocar *et al.* 2001; Morales *et al.* 2000; Teresi 2001; Fleishman *et al.* 2002; Orlando and Marshall 2002).

Scale analysis using IRT models

IRT scale analysis consists of evaluating scale information/reliability and the standard error of measurement (SEM), as in Figure 1.5.4. In contrast, CTT estimates a single reliability coefficient for all score values (usually Cronbach's α). Cronbach's α -reliability for the 14-item MH scale was 0.90, which implies that the scale is adequate. It is more likely, however, that the reliability varies, depending on who is being measured. Figure 1.5.4 shows that reliability is very high ($r > 0.90$) for individuals with low to middle levels of mental health (i.e., $-2.5 < \theta < 1$), and less precise, although still adequate, outside this range. The SEM for moderate to low mental health scores can be calculated to be approximately $1/\sqrt{11} = 0.3$, whereas for higher scores it rises to $1/\sqrt{3} = 0.58$. In CTT, the SEM for all score levels was 0.31.

The IRT scale information, and equivalently the SEM curve, evaluates the performance of an instrument. A developer wishing to shorten the instrument can delete an item and recalculate the new scale information curve to see the consequences. Information curves can also indicate areas for improvement. In our example, the curve suggests adding more items that discriminate among people who have good or excellent mental health.

Key methodological considerations for IRT modelling

Should I use CTT or IRT methods?

Applying IRT models does not imply abandoning CTT. Rather, IRT complements CTT to provide thorough analysis of an instrument. A researcher skilled in CTT could have found similar psychometric issues in the MH Summary scale to those we demonstrated. However, the ability of IRT models to describe item functioning along a continuum cannot be easily achieved with CTT. Major benefits of IRT are its comprehensive representation of content, and its ability to determine the optimal number of response categories for individual items.

Other sound and appealing features of IRT modelling are discussed in Embretson and Reise (2000) and Reeve (2003). One essential feature is that item properties are

invariant with respect to the sample of respondents. This means that the threshold and discrimination parameters remain stable even when the items are administered to different groups of people. It is this item-parameter invariance that makes IRT ideal for evaluating DIF across groups. It also enables 'test-equating', which is the linking onto a common metric of two or more questionnaires that measure the same domain. Furthermore, person scores are also invariant with respect to the set of questions used in an IRT-based scale. This provides the basis for computerized adaptive testing, where respondents receive different sets of questions but their scores can be compared.

Which IRT model should I use?

This is a difficult question and there is no single answer to it; we offer one perspective. Several key factors are involved in deciding which model to use: (1) the number of item response categories; (2) the construct being measured; (3) the purpose of the study; and (4) the sample size (discussed in the next section). First, the number of response categories limits the choice of IRT models, as described in Table 1.5.2.

Second, the nature of the construct being measured will affect the choice of model. Models in Table 1.5.2 assume that a single domain is the only factor affecting a person's responses to the items in the scale. However, complete unidimensionality, especially in HRQoL assessment, is rare, and the researcher must decide how much this assumption can be relaxed. While IRT models are robust to minor violations of the unidimensionality assumption, the greater the departure from this, the less satisfactory the model. Non-Rasch two-parameter models, which allow discrimination to vary from item to item, are more robust to departures from unidimensionality than Rasch models which assume equal loading of each item on the underlying construct. Rasch models may be a good choice for measuring constructs like mobility, where questions such as 'Can you walk a short distance?' and 'Can you run a mile?' have an obvious hierarchical ordering of difficulty. Many HRQoL domains are multifaceted and items in such a scale will have different relationships (correlations) with the underlying construct. Non-Rasch models may then be the more appropriate choice for reflecting item properties. However, there is a divide between those psychometricians who will only use Rasch models because of their strong statistical properties, and those who will use any IRT model (including Rasch) to find the best fit and the best interpretation.

The choice of model relates also to the purpose of the study. Non-Rasch models may be better for capturing the nature of the questions and the respondents' behaviour to the questions, and may be better for evaluating the psychometric properties of a questionnaire if data have been collected on an existing instrument. However, Rasch models may be appropriate for defining behaviour according to well-understood mathematical rules. It may also be possible to revise a questionnaire, selecting only items that meet the strict assumptions of this model, especially when choosing items from a large pool.

Other models exist. Multi-dimensional IRT models (Wilson in press) can improve the measurement of each domain by using information from other correlated domains. The added information improves reliability, especially for scales with few items. Non-parametric IRT models (Ramsay 1997) are helpful when sampling distributions are skewed, when respondent behaviour to items does not form a monotonically increasing item characteristic curve across the θ continuum, or when one needs an initial estimate for an exploratory look at the response data.

For a more complete discussion of IRT models, see Embretson and Reise (2000), Thissen and Wainer (2001), and van der Linden and Hambleton (1997).

What sample size does one need for IRT analysis?

There are many issues involved, and no definitive answer. We briefly address the main issues. First, the choice of IRT model affects the required sample sizes: Rasch models estimate fewest parameters, and thus smaller sample sizes are adequate for stable parameter estimates – perhaps as few as 100 (Linacre 1994, suggests 50 for the simplest Rasch model). This makes Rasch models attractive to health outcomes researchers, as large samples are often unavailable. For non-Rasch models, it has been shown that the Graded Response IRT Model can be estimated with 250 respondents, but around 500 are recommended for accurate parameter estimates (Embretson and Reise 2000). Non-Rasch software typically employs maximum-likelihood routines to estimate IRT parameters, and requires larger sample sizes for estimation.

Study purpose can affect the sample size. To evaluate questionnaire properties, one does not need large sample sizes for a clear picture of response behaviour, although it is important to have a heterogeneous sample that accurately reflects the range of population characteristics. But if the purpose is to generate accurate IRT scores for a questionnaire, or to calibrate items in an item bank for CAT, sample sizes over 500 are required.

Another important consideration is the sampling distribution of the patients. Ideally, patients should be spread fairly uniformly over the range of interest. Items at extreme ends of the construct will have higher standard errors associated with their estimated parameters if fewer people are located there. The size of standard error that is acceptable will depend on the researcher's measurement goals.

The better the item response data meet the IRT assumptions of unidimensionality, conditional independence, and hierarchical ordering by difficulty, the smaller the sample size need be. Also, the relationship between the items and the measured construct is important, as poorly related items may require larger sample sizes (Thissen 2003). Increasing the number of response categories also increases the need for larger samples, as more item parameters must be estimated. The ideal is to have respondents in each cell of all possible response patterns for a set of items; however, this is rarely achieved. At the least, it is important to have some people respond to each of the categories for every item to allow the IRT model to be fully estimated.

Conclusion

The main goal of this introductory IRT chapter was to demonstrate the wealth of information that can be gained by using these models. IRT is invaluable for evaluating the psychometric properties of both new and existing scales, and for revising questionnaires based on these findings. We have shown how item-characteristic curves can evaluate the response categories for items, and how this helps questionnaire developers determine whether more or fewer response categories are needed. We have illustrated how information curves enable developers to evaluate item and scale functioning over the range of the underlying construct, allowing developers to tailor their instrument for maximum precision when measuring study populations.

The real attraction of IRT models to the health outcomes community is the application of computer adaptive tests (CAT), which integrate the powerful features of IRT with the advances in computer technology. CAT delivers dynamic measures that tailor each questionnaire for each individual, based on information provided by their responses to previous questions. Therefore each person receives a different set of questions, yet their scores can be combined or compared with others because the items have been calibrated by IRT models. CAT offers shorter yet more reliable questionnaires that can for example be administered over Internet or personal handheld devices. The chapter by Bjorner and Ware provides details of CAT.

If IRT provides so many useful tools for evaluating and developing instruments, why is its use not more widespread? Several obstacles limit the use of IRT. First, most researchers have been trained in CTT statistics, are comfortable interpreting these statistics, and can easily generate from readily available software the familiar summary statistics such as Cronbach's α . In contrast, IRT models require advanced knowledge of measurement theory to understand their mathematical complexity, to check that assumptions are met, and to choose the appropriate model. In addition, supporting software and literature are not well adapted for researchers outside the field of educational measurement.

Despite the conceptual and computational challenges, the many potential advantages of IRT models should not be ignored. Knowledge of IRT is spreading within the academic disciplines of psychology, education, and public health. More books and tutorials are being written on the subject, and user-friendly software is being developed. Research that applies IRT models is appearing more frequently in health outcomes literature. A better understanding of the models and applications of IRT is emerging, and this will result in health outcomes instruments that are shorter, more reliable, and better targeted toward the populations of interest.

References

- Azocar, F., Arean, P., Miranda, J., and Munoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology*, 57 (3), 355–365.

- Embretson, S. E. and Reise, S. P.** (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fleishman, J. A., Spector, W. D., and Altman, B. M.** (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences*, **57B** (5), S275–S284.
- Hambleton, R. K., Robin, F., and Xing, D.** (2000). Item response models for the analysis of educational and psychological test data. In H. Tinsley and S. Brown S. (eds), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, pp. 553–585. San Diego, CA: Academic Press.
- Hays, R. D. and Morales, L. S.** (2001). The RAND-36 measure of health-related quality of life. *Annals of Medicine*, **33**, 350–357.
- Linacre, J. M.** (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, **7**, 4, 328.
- Morales, L. S., Reise, S. P., and Hays, R. D.** (2000). Evaluating the equivalence of health care ratings by whites and Hispanics. *Medical Care*, **38** (5), 517–527.
- Orlando, M. and Marshall, G. N.** (2002). Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychological Assessment*, **14** (1), 50–59.
- Orlando, M. and Thissen, D.** (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, **24**, 50–64.
- Ramsay, J. O.** (1997). A functional approach to modeling test data. In W. J. van der Linder and R. K. Hambleton (eds) *Handbook of Modern Item Response Theory*, pp. 381–394. New York: Springer.
- Reeve, B. B.** (2003). Item response theory modeling in health outcomes measurement. *Expert Review of Pharmacoeconomics and Outcomes Research*, **3** (2), 131–145.
- Reise, S. P.** (in press). Item response theory and its applications for cancer outcomes measurement. In J. Lipscomb, C. Gotay and C. Snyder (eds), *Outcomes Assessment in Cancer*. Cambridge: Cambridge University Press.
- Samejima, F.** (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, **34** (4) Pt. 2, Whole No. 17.
- Teresi, J. A.** (2001). Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. *Journal of Mental Health and Aging*, **7** (1), 31–40.
- Thissen, D.** (1991). *MULTILOG User's Guide, Version 6.3*. Chicago, IL: Scientific Software.
- Thissen, D.** (2003). Estimation in Multilog. In M. du Toit (ed.) *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood, IL: Scientific Software International.
- Thissen, D. and Wainer, H.** (eds) (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.
- van der Linden, W. J., and Hambleton, R. K.** (eds) (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.
- Ware J.E. and Sherbourne, C. D.** (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, **30**, 473–83.
- Wilson, M.** (in press). Subscales and summary scales: issues in health-related outcomes. In J. Lipscomb, C. Gotay, and C. Snyder (eds) *Outcomes Assessment in Cancer*. Cambridge: Cambridge University Press.

