HEALTH EDUCATION RESEARCH
Theory & Practice

Vol.21 (Supplement 1) 2006
Pages i4–i18
Advance Access publication 3 October 2006

Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling

Mark Wilson*, Diane D. Allen and Jun Corser Li

Abstract

This paper is the first of several papers designed to demonstrate how the application of item response models in the behavioral sciences can be used to enhance the conceptual and technical toolkit of researchers and developers and to understand better the psychometric properties of psychosocial measures. The papers all use baseline data from the Behavior Change Consortium data archive. This paper begins with an introduction to item response models, including both dichotomous and polytomous versions. The concepts of respondent and item location, model interpretation, standard errors and testing model fit are introduced and described. A sample analysis based on data from the self-efficacy scale is used to illustrate the concepts and techniques.

Introduction

The process currently employed to assess the reliability and validity of scales in the behavioral sciences is highly influenced by the theory of a true score [1] or the actual amount of the construct being measured if all sources of error could be eliminated. The idea of a true score provides the theoretical foundation for developing reliability measures such as the widely used Cronbach's alpha [2], and formed the basis of classical test theory (CTT) (also referred to as the classical true score model) developed in the psychological and educational measurement context at a time when norm-referenced testing dominated those areas. 'Grading on a curve' or diagnosing illness based on the top or bottom 10% of the population on a particular measure is an example of norm-referenced testing.

Guttman's [3] scalogram approach (also known as Guttman scaling) initiated the notion that tests should have a meaningful interpretation in terms of the items that comprise the test. Following Guttman, and in contrast to the classical approach, item response modeling (IRM) integrates the items into the measurement model thus narrowing the unit of measure of an instrument to the item level. The idea behind IRM is that people respond to items on tests or surveys based on their ability or attitude and the difficulty or endorsability of the item. In a testing situation, if their ability is high and an item is easy, then they have a high probability of selecting the right answer. Likewise, if people's attitude is strong (e.g. exercise is good) and the item is easily endorsed (e.g. exercise can help me live better) then they have a high probability of endorsing the item strongly. IRM analysis of responses to a test or survey can provide estimates of each person's and each item's location on the construct of interest, along with standard errors for each estimate rather than an aggregated error for the entire test or survey. Hence, IRM is better suited to assessing the reliability and validity of a criterion-referenced test because the properties of the test can be assessed at any selected critical cutoff. The

Graduate School of Education, University of California, Berkeley, CA 94720, USA
***Correspondence to: M. Wilson.**
E-mail: markw@berkeley.edu

doi:10.1093/her/cyl108

use of criterion-referenced procedures is necessary in the behavioral sciences when assessing the efficacy of interventions in changing behavior. However, researchers continue to use classical psychometric procedures that are not well suited to assess the measurement properties of these tests.

Although IRM encompasses many issues of measurement, including item discrimination, guessing parameters, rater effects, facets and item banking, this paper seeks only to provide an introduction to IRM in the context of behavioral measures. We have chosen specifically to focus on one family of models, the Rasch one-parameter models, because they are the simplest and a good starting place. The analyses and interpretations presented are illustrated using baseline data from the self-efficacy (SE) scale collected by the Behavior Change Consortium (BCC) [4], a group of projects at different sites examining people's changes in nutritional, smoking or exercise behaviors in response to intervention. This paper is the first in a series. Others will examine additional IRM measurement issues such as (i) the comparison of classical and IRM perspectives [5]; (ii) the possibility of multidimensionality [6] and (iii) the possibility, advantages and limitations of equating tests [7]. This paper is based on an account of IRM given in Wilson [8]. Our purpose is to provide the reader with a foundation of terms and concepts in IRM with which to evaluate literature and instruments in the behavioral sciences using contemporary measurement tools.

An example: the SE scale for exercise

The BCC [4] proposed to measure not only behavior change upon intervention but also some of the mediating variables such as SE associated with those changes. One of the measures, the SE scale for exercise (SE scale) [9], proved a good choice for demonstrating the use of item response models in analyzing a self-report instrument that canvasses respondent attitudes. Only baseline data were used; no results of intervention were analyzed for the demonstration of IRM.

The SE scale was developed out of a social-cognitive approach to behavior change to help explain variations in exercise behaviors. In contrast to an enduring trait such as self-motivation, SE is thought to be a situational belief dependent on current personal attitudes and the particular environment related to the task. SE is defined as 'a specific belief in one's ability to perform a particular behavior' [9 p. 396]. If SE helps mediate behavior change, then those with greater SE should have greater amounts of behavior change in response to interventions than those with lower SE. The SE scale consists of 14 items that express the certainty the respondent has that he or she could exercise under various adverse conditions (see Table I). Items reflect 'potentially conflictual situations' based on 'information gained from previous research with similar populations in which relapse situations had been identified' [9 p. 401]. Respondents rate each item 0, 10, 20, 30 and on up to 100% in 10% increments (resulting in 11 categories of responses): from 0% indicating 'I cannot do it at all' to 100% indicating 'certain that I can do it'. Scoring averages the responses if at least 13 items are completed.

Table I. *SE scale for exercise*

Item number	Items: 'I could exercise ...'
1	... when tired
2	... when feeling anxious
3	... when feeling depressed
4	... during bad weather
5	... during or following a personal crisis
6	... when slightly sore from the last time I exercised
7	... when on vacation
8	... when there are competing interests (like my favorite TV show)
9	... when I have a lot of work to do
10	... when I haven't reached my exercise goals
11	... when I don't receive support from family or friends
12	... following complete recovery from an illness which has caused me to stop exercising for a week or longer
13	... when I have no one to exercise with
14	... when my schedule is hectic

The data on the SE scale used for this paper come from two different BCC projects. The first project, conducted out of Stanford University, focused on 'exercise advice by human or computer' [10]. The Stanford University project sought to increase physical activity among middle-aged and older adults. Researchers compared adoption and maintenance of exercise programs between those who received counseling by a person or a computer, testing whether extrinsic or intrinsic motivation is a more powerful force in behavior change. The second project, conducted out of the University of Tennessee, focused on the 'Health Opportunities with Physical Exercise trial' [11]. The University of Tennessee project evaluated peer versus provider interventionist encouragement and their effect on overcoming barriers to increased physical activity for the urban poor. Several behavioral variables were collected related to physical activity and compared pre- and post-intervention. In both of these projects, the researchers hypothesized that SE was one of the mediating variables for the behavioral changes assessed in these studies. The SE scale was intended to gauge whether more SE was associated with greater success in the intervention.

The data used in the analyses came from respondents whose characteristics are summarized in Table II. Because this data set came from two different studies with different populations of interest, the respondent characteristics had a wide range. The mean age of the respondents was 53 ($\pm$10.7), ranging from 28 to 85 years old. Almost half of the respondents were between 45 and 60 years old, reflecting the Stanford University site's focus on middle-aged and older respondents. The mean income level was \$45 000 ($\pm$28 000), ranging from <\$10 000 to >\$80 000. About 44% of the respondents had an annual income of <\$40 000, reflecting the University of Tennessee site's focus on the urban poor. About 47% indicated that they were non-White. Fifty-two percent were either married or living with a partner. Eighty-eight percent had at least some education past high school. Only 22% were male.

In order to demonstrate dichotomous data analysis and the logic of IRM, the 11 categories of response (0–100% in 10% increments) were first collapsed into two, $\leqslant$50 and >50%. Explicitly, the categories '0, 10, 20, 30, 40, 50%' were coded to '0', and the other categories '60, 70, 80, 90, 100%' were coded as '1'. The dichotomized data will be used to illustrate aspects of item response models. Later, the polytomous data will be analyzed directly.

Table II. *Characteristics of the respondents in this data set*

Respondent characteristics	*n*	%[a]
Site affiliation	504	
Stanford University	221	44
University of Tennessee	283	56
Gender	504	
Male	110	22
Female	394	78
Race	503	
White, not Hispanic	269	53
Black, not Hispanic	202	40
Hispanic	17	3
Other	16	3
Age	502	
28–45	121	24
46–59	247	49
60–85	134	27
Marital status	502	
Not married	72	14
Presently married	243	48
Living with partner	19	4
Divorced	114	23
Separated	23	5
Widowed	31	6
Education level	503	
Less than high school diploma	13	2
High school diploma	49	10
Some college, trade, vocational or technical school	155	31
4 years college/graduated	114	23
Post-graduate work	172	34
Income (per year)	490	
\$0–39 999	214	44
\$40 000–79 999	139	28
\$80 000+	137	28

[a]Percentages may not add to 100 for some characteristics because of rounding.

The Rasch model

Item response models differ from the CTT model in several critical ways. First, item response models

are expressed at both the item and test levels, while CTT models are expressed only at the test level, where the focus is on the total score of an instrument or test. Second, item response models focus on the 'probability' of the observed responses rather than the responses themselves. Third, item response models, such as the Rasch model, assume that the probability of selecting a given response to Item *i* is modeled as a function of both the respondent's location on the variable θ (theta) and parameters associated with the item such as the item's location on the variable δ (delta). Note that θ and δ represent certain locations on the underlying construct. The construct is the latent variable that the instrument is intended to measure—it derives its meaning from both the characteristics of the people being measured and the set of items being used for measuring. Fourth, this probability is assumed to be a specific function of θ and δ, generally one that is monotonically increasing, with asymptotes to 0 and 1.0 probability at plus and minus infinity, respectively (i.e. because we assume that we can never be 0% sure, or100% sure, that the respondent will select a given response option no matter how weak or strong his/her propensity is). If knowledge is assessed, as in an educational context, the terms respondent 'ability' and item 'difficulty' make sense. If a psychosocial construct is measured 'respondent attitude' and 'item endorsability' may be more appropriate terms to use. For introductory purposes, however, respondent and item 'location' will be used to emphasize their relationship to the same construct. In the context of the SE scale, 'respondent location' would refer to the amount of SE the person has. In CTT, this would be analogous to the total score on the SE scale. In the Rasch model, 'item location' refers to the amount of SE that a person would have to have to endorse that question. Specifically, the item location would be the point on the scale at which a generic person would need to be in order to have a 0.50 probability of responding positively to that item.

Let us consider three situations:

(i) $\theta = \delta$, when the respondent and item locations are the same, the probability of selecting response '1' (instead of '0') to Item *i* is 0.5 (see Fig. 1). This occurs when the respondent's amount of SE matches the level of SE assessed by the item. For example, if the respondent had a moderate amount of SE for exercising and the endorsability of, say, exercising while on vacation targeted those with a moderate amount of SE, the respondent's probability of selecting response option '1' would be 0.5;

(ii) $\theta > \delta$, when the respondent location is greater than the item location then the probability of selecting 1 to this item is >0.5 (see Fig. 1). For example, if the respondent had a high amount of SE for exercising, but exercising while on vacation targets a behavior that requires a moderate amount of SE, the respondent is expected to have a high probability of selecting response option 1, indicating great certainty that the respondent could do it and

(iii) $\theta < \delta$, when the respondent location is less than the item location then the probability of selecting 1 to this item is <0.5 (see Fig. 1). For example, if the respondent has a low amount of SE for exercising and exercising while on vacation is targeting a behavior that requires a moderate amount of SE, the respondent is expected to have a low probability of selecting response option 1 because the respondent does not think he or she could effectively maintain an exercise program while on vacation.

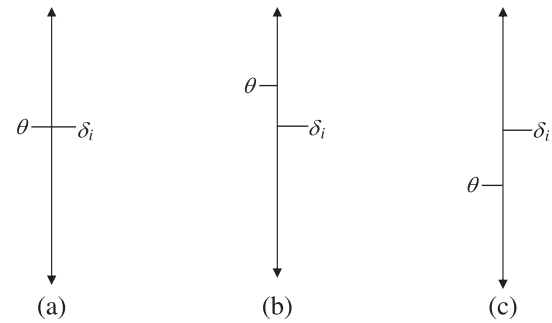

Fig. 1. Representation of three relationships between respondent location and the location of an item.

As indicated above, the probability of selecting response 1 to Item i ($X_i = 1$) is a function, f, of both respondent location (θ) and item location (δ); more specifically it is a function of the 'difference' between respondent location and that item location ($\theta - \delta_i$). This relationship can be expressed by a mathematical representation presented here as Equation (1):

$$P(X_i = 1|\theta, \delta_i) = f(\theta - \delta_i). \quad (1)$$

Graphically, this relationship can be plotted and it is shown in Fig. 2. (Note that the orientation of the graph has been rotated compared with Fig. 1 to match the typical way this graph is shown.) The respondent locations, θ, are plotted on the horizontal axis, and the probability of selecting response '1' to a given item is shown on the vertical axis. This type of figure is customarily called an 'item response function' (IRF) (other common terms are 'item characteristic curve' and 'item response curve') because it describes how a respondent responds to an item. The IRF depends on the respondent's total amount of SE to perform the behavior and the level of SE assessed by the item.

Equation (1) specified that the probability of selecting a given response option is a function of the difference between respondent location (θ) and item location (δ). For the Rasch model, the actual function for the probability of selecting a given response option to Item i, where e indicates the natural log base of 2.718, is

$$P(X_i = 1|\theta, \delta_i) = \frac{e^{(\theta - \delta_i)}}{1 + e^{(\theta - \delta_i)}}. \quad (2)$$

The Rasch model equation is the simplest model, and hence a good starting point for understanding item response models. Although the expression on the right-hand side looks somewhat complex, it is a function of $\theta - \delta_i$, as in Equation (1). Respondent location (θ) and item location (δ) are both graphically presented on the construct map (a graph representing the whole construct or idea being measured) (Fig. 3). This difference governs the probability that the respondent will select the positive response option. These relationships are displayed in a 'Wright map' as shown in Fig. 3, with the respondents' locations and items' locations presented on either side of a vertical line.

There are a number of features in the Wright map that are worth pointing out. The dashed vertical line expresses the construct or latent variable in logits, which relate the latent variable to the probability of response. (The relationship of a logit, technically the 'log of the odds', to Equation (2) is shown in a later section.) The logits provide the units of the construct, and are specified to the left of the vertical line. The raw score units of the SE scale are also presented to the left of the logits. The Wright map shows that a logit of 0, which corresponds to having a moderate amount of SE, is similar to a raw SE scale score of 7. This provides a translation between the metric of the item response map (the logits) and the more familiar raw score metric (as used in the classical approach). The results from an item response analysis can be transformed into the raw score metric using this (non-linear) relationship.

On the left-hand side of the vertical line, under 'Respondents', the locations of the respondents on the logits scale are indicated by X's (each 'X' may indicate more than one respondent). These form a histogram on the Wright map showing the shape of the respondent distribution. Figure 3 shows a fairly flat (not a normal) distribution for the respondents' locations, indicating that respondents had a wide range of SE in overcoming barriers

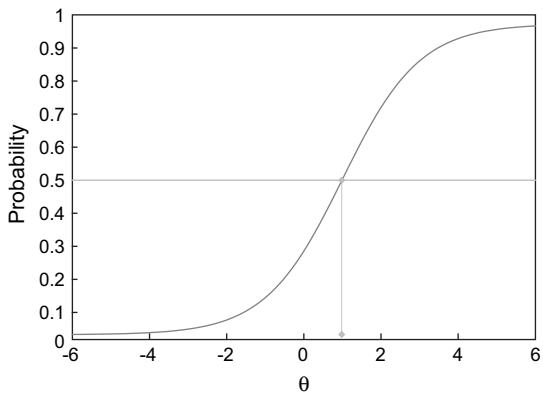

Fig. 2. Relationship between respondent location (θ) and probability of a response of '1' for an Item i with endorsability (δ) of 1.0.

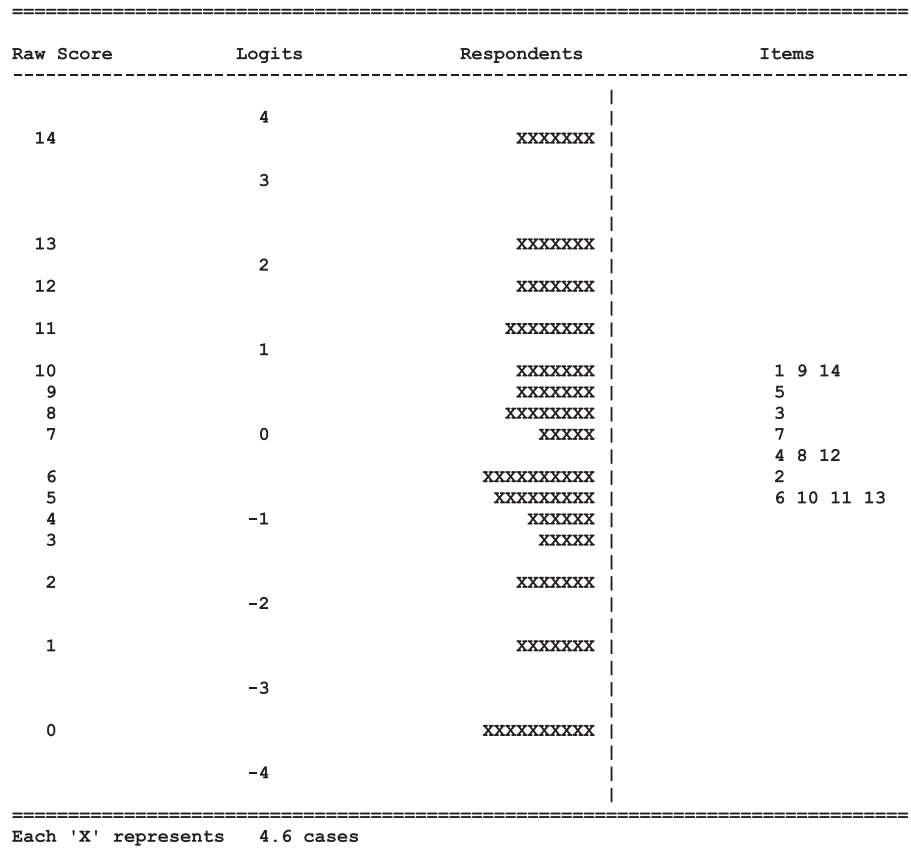

Fig. 3. Wright map of SE scale based on dichotomous items.

toward physical activity. Given that these are baseline data, we would have expected to see a somewhat skewed distribution with more respondents having lower levels of SE (i.e. more respondents with negative logits or <0 logit on the SE scale).

Finally, on the right-hand side of the vertical line, under 'Items', the locations of the items are shown. Each item location indicates the amount of SE a generic person must have if there is a 0.50 probability of that person giving a positive response to these items. The Wright map shows that the 14 items are located between 1.0 and −1.0 logits. Because of the nature of the IRF for each item, these item locations indicate the location on the SE continuum at which each item can provide the most information. For example, Item 7, located at −0.01 logits, is expected to provide more information for respondents who possess 0.0 logit of SE (i.e. respondents that have a moderate amount of SE) than respondents with much higher or lower amounts. Ideally, the location of the set of items would have the same range as the location of the set of respondents, unless the scale is designed to provide more precision (i.e. lower standard error of measurement) only at a certain point along the SE continuum.

Armed with Equation (2), the relationship between the logits and the probability of response can be made clear. For the respondent at 0.0 logit in Fig. 3, the probability of responding '1' on Item 7 should be 0.50 (because the respondent

and the item are at the same location on the Wright map). To check this note that

$$P(X_i = 1 | 0.0, 0.0) = \frac{e^{0.0-0.0}}{1+e^{0.0-0.0}} = \frac{e^{0.0}}{1+e^{0.0}} = \frac{1}{1+1} = 0.50.$$

Similarly, for a respondent located at 1.0 logit, the probability of a '1' on Item 7 will be >0.50, because the respondent is higher than the item. To be exact, the probability will be

$$P(X_i = 1 | 1.0, 0.0) = \frac{e^{1.0-0.0}}{1+e^{1.0-0.0}} = \frac{e^{1.0}}{1+e^{1.0}} = \frac{2.718}{1+2.718} = 0.73.$$

Similarly, for the respondents at −1.0 logit, the probability of a '1' will be <0.50, because the respondent is lower than the item. To be exact, the probability will be $e^{-1}/(1 + e^{-1}) = 0.27$. Thus, in the Wright map (vertical) distances relate to probability.

Where do the estimates of the respondent and item locations come from? The equations given above for the Rasch model are not directly solvable for the θs and δs. Therefore, they are estimated using one of several statistical estimation approaches. Although several software packages can perform these estimations, the software used for this paper is called 'ConQuest' [12]; it performed all the statistical calculations needed in the following sections. Discussion of estimation is beyond the scope of this paper. Interested readers should consult Adams and Wilson [13]; another useful source on the Rasch model is Fischer and Molenaar [14].

The Wright map shown in Fig. 3 is not just a sketch of the 'idea' of the construct of SE, it is an empirical map, based on respondents' self-reports that can be used to interpret the measure, both qualitatively and quantitatively. The respondents range from those at the top who are 'more confident' that they can be effective at continuing an exercise program to those that are 'less confident' at the bottom. Table III shows the actual equivalence between the raw scores and the logit estimates.

Table III. *Raw score to logit estimate equivalences, dichotomous data*

Raw score	Logit estimate	Standard error
0	−3.53	1.51
1	−2.34	0.91
2	−1.73	0.73
3	−1.29	0.65
4	−0.93	0.60
5	−0.60	0.58
6	−0.30	0.56
7	−0.01	0.56
8	0.29	0.56
9	0.59	0.58
10	0.92	0.61
11	1.29	0.66
12	1.74	0.74
13	2.36	0.91
14	3.56	1.51

More than two response categories

The discussion and graphs above provided interesting ways to interpret output from the measurement model when the data were dichotomous (i.e. just two response categories). In this section, these principles are generalized to items with more than two ordinal response categories (called 'polytomous' data). First, we need to develop a somewhat simpler way to express Equation (2). Some algebra will show that, following Equation (2), the ratio of the probability of 1 and 0 is a relatively simple expression $e^{\theta-\delta_i}$. Then, taking the log of that, we get

$$\log\left(\frac{P(X_i = 1)}{P(X_i = 0)}\right) = \theta - \delta_i. \quad (3)$$

Now, the 'odds' of an event is the proportion of times that an event occurred compared with the times it did not occur. Thus, Equation (3) gives an expression for the log of the odds of a '1' (as opposed to a '0'). The log of the odds is often called the 'logit'. Thus, Equation (3) can be rewritten as

$$\text{logit}(1:0) = \theta - \delta_i. \quad (4)$$

This expression highlights the simple relationship between the person location and the item location in the Rasch model.

This gives a way to generalize the dichotomous expression in Equation (4) to a polytomous relationship. Consider the case of items with five ordered response categories or scores: 0, 1, 2, 3 and 4. Suppose that we assumed that the logit relationship in Equation (4) held between item scores 0 and 1 as

$$\text{logit}(1:0) = \theta - \delta_{i1}, \tag{5}$$

where the item location δ_i. has been relabeled δ_{i1} to denote that this is just for the 0/1 comparison. Then, just repeat this for the pair of scores 1 and 2:

$$\text{logit}(2:1) = \theta - \delta_{i2}. \tag{6}$$

Repeat this again for the subsequent pairs 2 and 3 and 3 and 4:

$$\text{logit}(3:2) = \theta - \delta_{i3}, \tag{7}$$

$$\text{logit}(4:3) = \theta - \delta_{i4}. \tag{8}$$

These four equations are sufficient to generalize the Rasch model to polytomous data where there are five ordered response categories. The parameters δ_{ik} are known as 'step parameters'—they govern the probability of making the 'step' from Score $k-1$ to Score k [15]. For example, look at Equation (8): the relationship says that, if a respondent is in response category either 3 or 4, then the relative probability of being in response category 4 is a function of $\theta - \delta_{i4}$. That is, it is a function of the difference between the person location and the step parameter. In general, similar equations can be developed for any finite number of ordered categories, and there will be one equation less than the number of categories. This is because they are related to the comparison between score categories $k-1$ and k, and there is one less step comparison than there are categories, and so one less step parameter. See Wright and Masters [15] for a lengthier discussion of these step parameters and their interpretation. Thus, for the SE scale, there will be 10 steps for the 11 categories of response (0–100%).

To develop a graphical expression for the polytomous case, we start with the IRF for $X_i = 0$, where 0 = 0% level of SE. In the dichotomous case, the probability of $X_i = 0$ is simply 1.0 minus the probability of $X_i = 1$. In other words, the probability of $X_i = 0$ plus the probability of $X_i = 1$ equals 1. The IRF for $X_i = 0$ is shown in Fig. 4. The curve is the one in Fig. 2 turned upside down! The equivalent for a polytomous item will generalize this by adding more curves to the figure and dividing the probability into more than two segments. Just as in the dichotomous case, for any logit value, the sum of the probabilities of all possible responses must equal 1.

Thus, an equivalent of Fig. 4 for a polytomous (in this case, five category) item is shown in Fig. 5. Note that in this graph, the curves are cumulative versions of the category response functions. That is, what is shown is the cumulative probability of being in successive score categories: first, being in score category 0; then being in score categories 0 or 1; then being in score categories 0, 1 or 2; etc. Note that the probability of a 0 response instead of any other response is very high at the lower end of the scale; that the probability of a 0 or 1 response rather than a 2, 3 or 4 response decreases as you get higher on the logit scale and that the probability of a response other than a 4 steadily decreases, becoming close to 0 at the highest end of the scale.

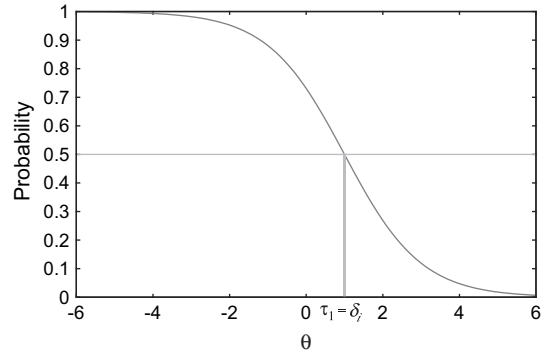

Fig. 4. Item response function for 0 for an item with endorsability of 1.0.

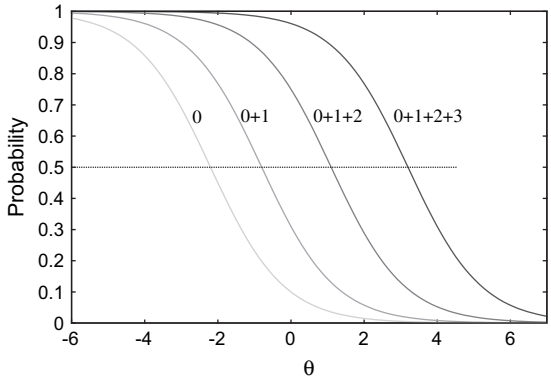

Fig. 5. The cumulative category response functions for a polytomous item.

A Wright map for the polytomous item identifies the critical points where these cumulative curves intersect with the horizontal line where probability equals 0.5, shown on Fig. 5. These points are known as 'Thurstone thresholds': the *k*th Thurstone threshold is the point at which the probability of the scores below k is equal to the probability of the scores k and above (and that probability is, of course, 0.50). For example, the intersection of the first curve with the straight line (τ_1) is the point at which responses 1, 2, 3 and 4 together become more likely than a response of 0; the intersection of the second curve with the straight line (τ_2) is the point at which responses of 2, 3 and 4 together become more likely than responses 0 and 1 combined; the intersection of the third curve with the straight line (τ_3) is the point at which responses 3 and 4 together become more likely than responses 0, 1 and 2 combined; etc. Note that, except in the dichotomous case, the Thurstone thresholds, in general, are NOT the item parameters δ_{i1}, ..., δ_{i4} in Equations (5)–(8). Some people find this confusing, but we have chosen this way to represent the category response functions because it avoids some complexities that arise in interpreting the δ_{ik} parameters due to the fact that they are defined relative to pairs of categories (as above). Nevertheless, the relative locations of the Thurstone thresholds are very useful for interpretive purposes (as discussed below).

The SE scale example, continued

The same data analyzed in a dichotomous format were also analyzed in their original format with 11 categories, and the resulting Wright map is shown in Fig. 6. This map has the same general layout as Fig. 3. In particular, note that the same types of information are given on the left-hand side of the map. But the right-hand side looks a bit more complicated—and that is because the number of categories has been restored to its original count. Each item now has 10 Thurstone thresholds, one between each pair of the ordered response categories. The first threshold for each item, governing the transition from marking 0% to marking 10% confident that the respondent could be effective under that item's adverse condition is depicted at the bottom of the Wright map, being the easiest to surpass. The 10th threshold for each item, governing the transition from being 90 to 100% confident are shown at the top of the Wright map, being the most difficult for respondents to choose.

There are a number of differences between Figs 3 and 6. Looking from left to right, there are now more possible total raw scores and hence more bars in the histogram of respondents' locations (this is because the categories are no longer collapsed for recoding as 0 or 1). The shape of the respondent distribution is different also. Figure 6 approximates a normal curve when the same data are analyzed polytomously. Note that the skewness, mentioned above as being expected, but not observed in the dichotomized data, is evident here.

Just as for Fig. 3, one can gauge the approximate probability relationships between the items and persons using the map itself. For example, at the bottom of Fig. 6 the very lowest scoring respondents have about a 0.50 probability of responding that they are at or above '10% confident' that they can be effective under the adverse conditions in about half of these items, but have a very tiny probability of responding at '100% confidence' for these items. Going to the top end of the figure, the highest scoring respondents have a 0.50 probability of responding that they are at or above '90% confidence' on Item

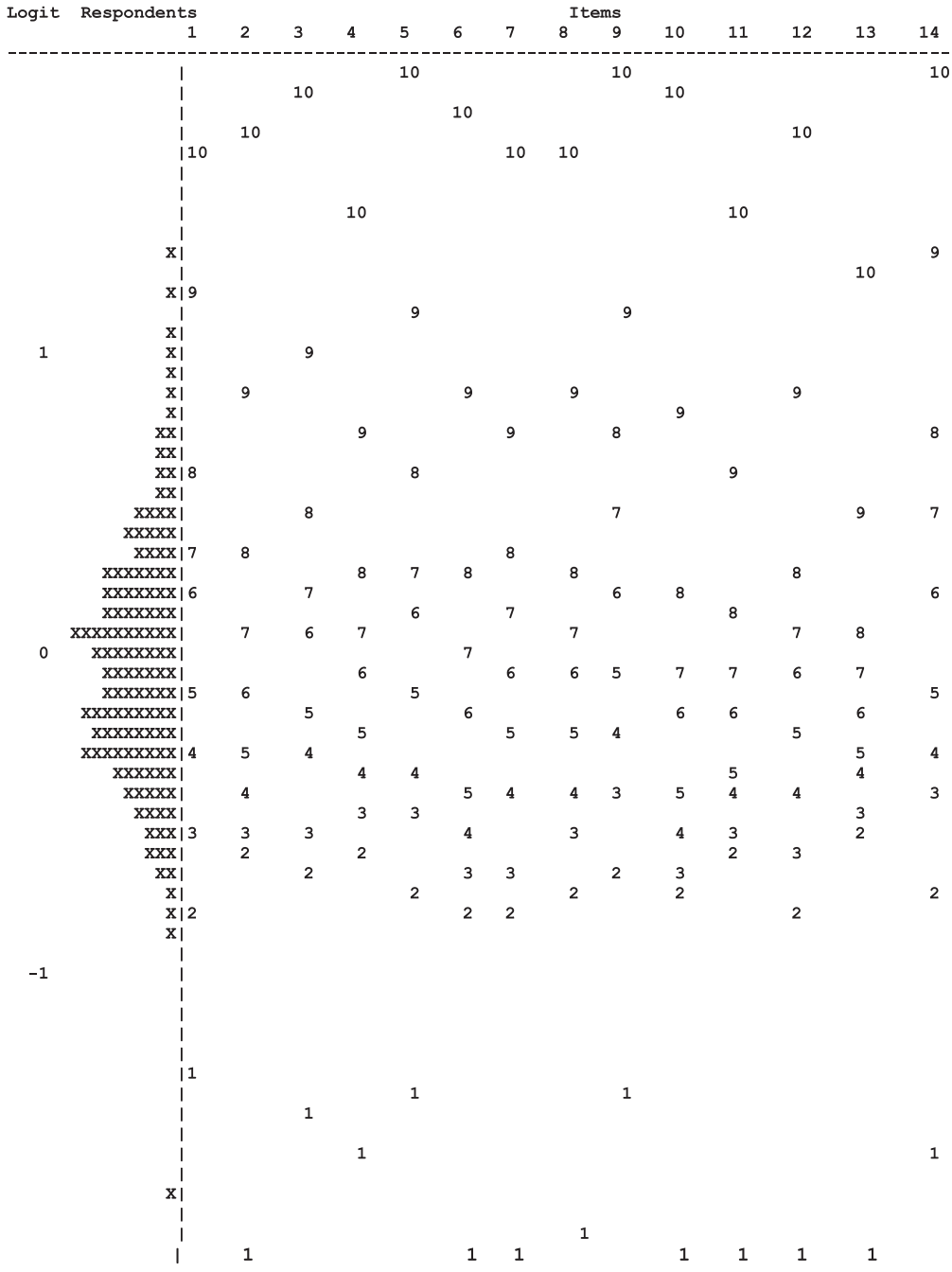

Fig. 6. Wright map of item thresholds for SE scale analyzed polytomously (each 'X' represents 3.7 cases).

14 and an even higher probability of responding at or above 90% confidence to the other items.

What one would do at this point in a standard interpretation of the Wright map (e.g. see Wilson [8]) would be to seek to interpret the meanings that can be attached to portions of the construct, using interpretations of the wording of items, and the wording of the categories. In this case, it is clear from the Wright map that the most important characteristic differentiating higher from lower is the set of response categories rather than the specific items themselves. Unfortunately, the categories are nothing other than percentages: '0', '10', etc., which gives almost no basis for meaningful interpretation. Thus, by the nature of its response options, the SE scale offers no worthwhile opportunity for interpretation of its internal structure.

The use of Wright maps such as those depicted in Figs 3 and 6 aids one in interpreting the usefulness of individual items, as well as the relative location of each respondent. Knowing that an item is easy or difficult to endorse for the population of respondents in question can help in the evaluation of that item and its usefulness in the instrument. Knowing that a respondent is located toward the bottom or the top of the construct can help in the evaluation of that respondent's perception of SE in the case of the SE scale.

Interpretations and errors

Another aid to the interpretation of item and respondent location is examination of the reliability, using the standard error. Recall that each location is an estimate. That means that it is subject to a degree of uncertainty. This uncertainty is usually characterized using the standard error of the location—the so-called 'standard error of measurement'. This quantity, which is calculated by the software along with the estimate, indicates how accurate each estimate is. For example, if a respondent scored 10 on the dichotomous version of the SE scale (see Table III), then the respondent's location is 0.92 logits and the standard error of the respondent's location is 0.61. This is usually interpreted by saying that the measurer is uncertain about the exact location of the respondent, but that it is centered approximately on 0.92 logits with a 95% confidence interval ranging from −0.28 to 2.12 or a raw score of 6–12 out of 14. This is a fairly wide confidence interval, spanning a quite wide part of the range of the instrument from its lowest score location to the highest. This observation corresponds to the fairly low reliability for the dichotomized scale 0.78. Note that the reliability coefficient being used here is one based on the logit metric rather than the score metric as are the classical reliabilities (KR-20, Cronbach–Guttman alpha, etc.), but it is based on an analogous approach, and can be interpreted in an analogous way [16]. Note that researchers in behavioral sciences have typically used a 0.70 reliability (i.e. Nunally and Bernstein [17]) as a lower bound of acceptability.

The precision of the original polytomous scoring of the SE scale was also computed. A respondent who scored 106 on the SE scale (see Fig. 7) has a logit score of 0.535 with a standard error of 0.165. The 95% confidence interval is (0.214, 0.856), a range of 0.64 logits, from just above a score of 91 to just below a score of 116. Although a 25-point raw score range may seem wide for a 95% confidence interval around a respondent's raw score, it is an improvement on the case for the dichotomized data. For the polytomous data, the reliability is a more respectable 0.92.

Figure 8 shows, for the polytomous data, how the standard errors differ across the range of the construct, reflecting the greater number of item thresholds nearer the middle of the ability distribution, and fewer item thresholds at the extremes. Because these standard errors vary depending on the person's location, these are also called conditional standard errors of measurement. Sometimes, this is displayed as the 'Information' of an instrument. The Information is the reciprocal of the square of the standard error and is helpful in investigating where the instrument measures most precisely on the construct.

Similarly, the item locations also have a standard error. In typical measurement situations, where there are many more respondents than items, the

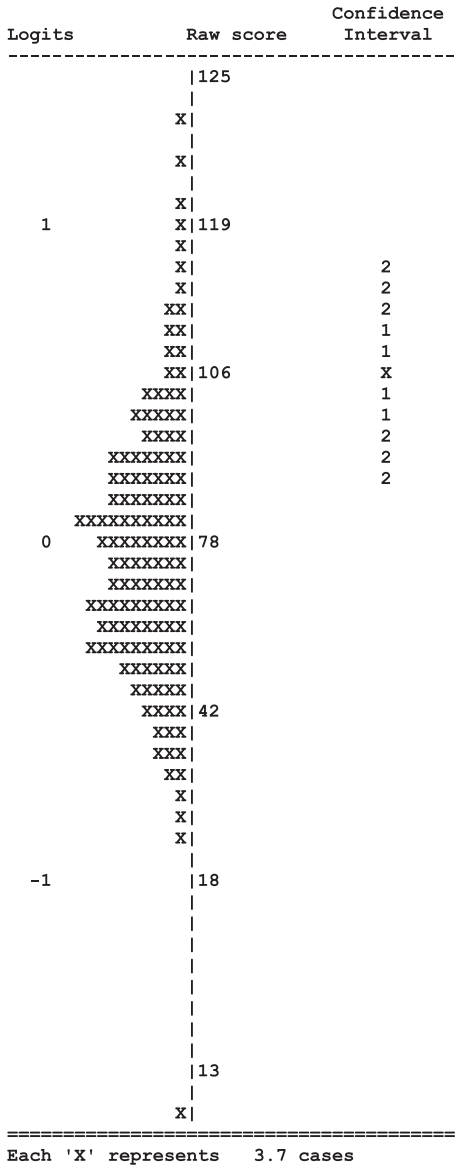

Fig. 7. A Wright map for the polytomous SE scale, showing 67% (1) and 95% (2) confidence intervals for a respondent with a score of 106.

item standard errors are quite a lot smaller than the respondent standard errors. For example, the standard errors of the SE scale item estimates range from 0.021 to 0.024. In many applications, the item standard errors are small enough to ignore

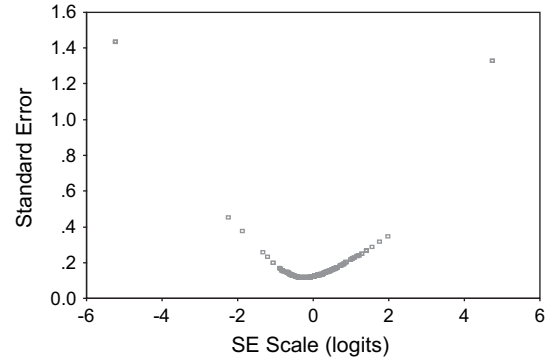

Fig. 8. The standard error of measurement for the SE scale (each circle represents a different score).

when interpreting the item locations. However, it is important to keep in mind that they are estimates subject to error, just as are the respondent location estimates. One situation that requires use of the item standard error is the calculation of item fit statistics, used in the assessment of the fit of a model.

Model fit

The gathering of evidence that the mathematical models that are being used are appropriate is generally termed the investigation of 'fit'—here it is discussed with respect to items. There is more than one approach to investigating fit—each approach tends to emphasize one aspect of the model over the other. In this section, the emphasis will be on consideration of how well the shapes of the empirical item characteristic curves are captured by the curves generated by the estimated item parameters. Most fit investigations begin by examining the residuals—the difference between the observed score and the expected score for a particular person and item:

$$Y_{in} = X_{in} - E_{in}, \tag{9}$$

where Y_{in}, X_{in} and E_{in} are the residual, the observed score and the expected score for person n responding to Item i, respectively. The expected score is given by

$$E_{in} = \sum_{k=1}^{K_i} kP(X_{in} = k|\theta, \boldsymbol{\delta}_i), \tag{10}$$

where K_i is the number of response categories for the item and $\boldsymbol{\delta}_i$ is a vector of the parameters for Item i. While we do not expect every response of a respondent to an item to have a small residual, we do expect that the distribution of these residuals across the instrument will meet certain standards, falling within a particular range specified by the measurer. Thus, fit indices usually consist of various ways of looking at the distribution of the residuals, their means and variances, etc.

For example, one way to detect differences from what we expect is to compare how much the actual residuals vary with how much they would vary randomly if the data fit the model. This is just what we do to calculate the so-called 'mean square fit statistic' [8, 15]. When the observed residuals are varying about as much as we expect the mean square should fall within a particular range ~1.0. When mean square values are >1.0, the observed variance is greater than the expected—and that can be interpreted as implying that the IRF slope indicated by the data is flatter than expected. When mean square values are <1.0, then the observed variance is less than the expected—and that can be interpreted as saying that the IRF slope indicated by the data is steeper than expected. In considering the interpretation of these results, note that items with a mean square >1 will be those that contribute less toward the overall estimation of the latent variable, and hence those that lie outside the specified range are the ones that are most problematical for measuring and should be attended to first. Items with a mean square <1 and outside the specified range are also problematical for measuring; lower variance means responses were less random than predicted. However, after the items above the upper limit have been deleted, and the item set recalibrated, it is often the case that some, if not all of the items that were below the lower limit are no longer in that critical region. There are several ways to create fit indices like this—the one shown above is often termed the 'weighted' mean square or sometimes the 'infit' mean square because the calculation corrects for occasional outliers that may affect the 'outfit' mean square.

As an effect size, there is no absolute criterion for what is a desirable range to specify for a weighted mean square value, but 0.75 (=3/4) is a reasonable lower bound and 1.33 (=4/3) is a reasonable upper bound [18]. A second fit index, the weighted t, uses a transformation that attempts to make the weighted mean square into a standard normal distribution [15], and is sometimes used to test the statistical significance of the mean square [15]. But, with large sample sizes, one can expect that this t statistic will show significant values for many items; hence, a safer strategy is to consider as problematical only those items that show as misfitting on both the mean square and the t statistics.

With this background, now look at Fig. 9 [18], which shows the weighted mean square for the average item locations for the SE scale data. The weighted mean square indicates that all of the items are fitting within reasonable bounds, with respect to

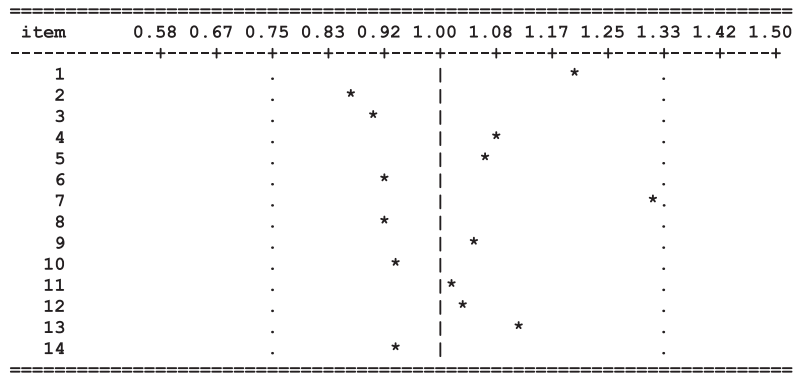

Fig. 9. Fit results for the SE scale data (weighted mean square for average item location).

the estimated average locations, although Item 7 is near the border with an infit mean square of 1.32. If we had looked at just the weighted *t* statistics, three of the 14 items would have looked out of bounds, i.e. the *t* statistics for three of these average location parameters are significant at the $\alpha = 0.05$ level, but recall, from above, that is not sufficient. We would be interested in items where both the mean square and the *t* indicate problems, and none of the items qualify. For the relative step parameters, only two *t* statistics are significant, and none of the mean squares is out of bounds. Thus, the overall finding is that the SE scale data fit the polytomous model reasonably well. If both the mean square and the *t* indicated problems, users of the scale should examine the specific items or steps to determine whether they should be revised, replaced, deleted or considered along a separate construct [8].

Conclusion

As the first in a series of papers utilizing BCC data to illustrate the use of IRM analysis, this paper introduces the basic elements of IRM. Data from the SE scale for exercise were analyzed using both dichotomous and polytomous models. The item response model analysis shows that the polytomous items seem to cover the content well, and provide reliable information about both the items and the respondents along the construct of SE for exercise. This information about the items and the respondents in this data set provides a foundation for determining the usefulness of the SE scale in measuring SE and thus offers a basis for interpreting validity evidence for this instrument as an assessment of behavioral measures in the BCC studies. A later paper will compare these findings with those from an analysis using the CTT, and appraise evidence for reliability and validity.

Acknowledgements

We would like to thank Louise Mâsse formerly of the National Cancer Institute (NCI), and now from the University of British Columbia, for organizing the project on which this work is based and for providing crucial guidance throughout the writing of the paper. The views presented in this paper represent those of the authors and not those of the NCI. Thanks also to Tom Baranowski, from Baylor College of Medicine, for helpful comments. Support for this project was provided by NCI (Contract No. 263-MQ-31958). We thank the BCC for providing the data that were used in the analyses. Any errors or omissions are, of course, solely the responsibility of the authors.

Conflict of interest statement

None declared.

References

1. Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904; **15**: 72–101.
2. Cronbach LJ. *Essentials of Psychological Testing*, 5th edn. New York, NY: Harper & Row, 1990.
3. Guttman L. A basis for scaling qualitative data. *Am Soc Rev* 1944; **9**: 139–50.
4. Ory MG, Jordan PJ, Bazzarre T. The Behavior Change Consortium: setting the stage for a new century of health behavior-change research. *Health Educ Res* 2002; **17**: 500–11.
5. Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health Educ Res* 2006; **21**(Suppl 1): i19–i32.
6. Allen DD, Wilson M. Introducing multidimensional item response modeling in health behavior and health education research. *Health Educ Res* 2006; **21**(Suppl 1): i73–i84.
7. Mâsse LC, Allen DD, Wilson M *et al*. Introducing equating methodologies to compare test scores from two different self-regulation scales. *Health Educ Res* 2006; **21**(Suppl 1): i110–i120.
8. Wilson M. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum, 2005.
9. Garcia AW, King AC. Predicting long-term adherence to aerobic exercise: a comparison of two models. *J Sport Exerc Psychol* 1991; **13**: 394–410.
10. King AC, Friedman R, Marcus BH *et al*. Harnessing motivational forces in the promotion of physical activity: the Community Health Advice by Telephone (CHAT) project. *Health Educ Res* 2002; **17**: 627–36.
11. Coday M, Klesges LM, Garrison RJ *et al*. Health Opportunities with Physical Exercise (HOPE): social contextual interventions to reduce sedentary behavior in urban settings. *Health Educ Res* 2002; **17**: 637–47.

12. Wu ML, Adams RJ, Wilson MR. *ACER ConQuest: Generalised Item Response Modelling Software* [computer program]. Hawthorn, Australia: ACER (Australian Council for Educational Research) Press, 1998.
13. Adams RJ, Wilson M. Formulating the Rasch model as a mixed coefficients multinomial logit. In: Engelhard G, Wilson M (eds). *Objective Measurement III: Theory Into Practice*. Norwood, NJ: Ablex, 1996, 143–66.
14. Fischer GH, Molenaar IW (eds). *Rasch Models: Foundations, Recent Developments, and Applications*. New York, NY: Springer-Verlag, 1995.
15. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: MESA Press, 1982.
16. Mislevy RJ, Beaton AE, Kaplan B *et al*. Estimating population characteristics from sparse matrix samples of item responses. *J Educ Meas* 1992; **29**: 133–61.
17. Nunally JC, Bernstein IH. *Psychometric Theory*, 3rd edn. New York, NY: McGraw-Hill, 1994.
18. Adams RF, Khoo ST. *Quest* [computer program]. Melbourne, Australia: ACER Press, 1996.

Received on September 6, 2005; accepted on August 24, 2006