HEALTH EDUCATION RESEARCH
Theory & Practice

Vol.21 (Supplement 1) 2006
Pages i19–i32
Advance Access publication 31 July 2006

Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach

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

Abstract

This paper compares the approach and resultant outcomes of item response models (IRMs) and classical test theory (CTT). First, it reviews basic ideas of CTT, and compares them to the ideas about using IRMs introduced in an earlier paper. It then applies a comparison scheme based on the AERA/APA/NCME 'Standards for Educational and Psychological Tests' to compare the two approaches under three general headings: (i) choosing a model; (ii) evidence for reliability—incorporating reliability coefficients and measurement error—and (iii) evidence for validity—including evidence based on instrument content, response processes, internal structure, other variables and consequences. An example analysis of a self-efficacy (SE) scale for exercise is used to illustrate these comparisons. The investigation found that there were (i) aspects of the techniques and outcomes that were similar between the two approaches, (ii) aspects where the item response modeling approach contributes to instrument construction and evaluation beyond the classical approach and (iii) aspects of the analysis where the measurement models had little to do with the analysis or outcomes. There were no aspects where the classical approach contributed to instrument construction or evaluation beyond what could be done with the IRM approach. Finally, properties of the SE scale are summarized and recommendations made.

Introduction

Item response models (IRMs) underlie many of the advances in contemporary measurement of the behavioral sciences, including assessment of the information provided by a particular item, criterion referenced assessment, computerized adaptive testing and item banking. Making full use of such advances requires knowledge of IRM that few yet possess. Comparison with the more well-known procedures associated with classical test theory (CTT) should help to situate new concepts of IRM and understand how each approach might contribute to measurement. In this paper, an analysis is conducted of a self-efficacy (SE) scale for exercise comparing IRM and CTT to illustrate the differences and similarities between these approaches. First, some basic ideas of the CTT approach are reviewed, to compare with the IRM ideas introduced in another paper in this volume [1]. A comparison scheme is described based on the AERA/APA/NCME 'Standards for Educational and Psychological Tests' [2] and used to compare the two approaches under three general headings: (i) choosing a model, (ii) evidence for reliability and (iii) evidence for validity. The paper comments on the characteristics of the validity evidence that seem to be presented for instruments like the SE scale. Finally, the paper

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

doi:10.1093/her/cyl053

ends with a summary of the recommendations one might make for the SE scale following the item response modeling approach.

Basic differences between IRM and CTT

Both IRM and CTT are used to (i) help develop instruments and (ii) check on their reliability and validity. The CTT approach is by far the most widely known measurement approach, and, in many areas, is the most widely used for instrument development and quality control. In the classical approach, a particular instrument establishes its respondents' 'amount' of a particular characteristic (i.e. ability or attitude in educational tests or psychological instruments) based on their raw score across all the items on the instrument. Instrument developers using the classical approach assume that the observed score (X) obtained is composed of the true score (T), representing the true assessment of this characteristic on this particular instrument, plus an overall error (E) [3]:

$$X = T + E.$$

In theory, the error or noise represents the variability if each respondent took the instrument many times without remembering previous trials or changing in the characteristic measured [4, 5]. The X and T are both indicators of the interaction between the instrument and the respondents' characteristic; neither presumes to indicate an amount of characteristic directly. The difficulty of the instrument depends on the amount of the characteristic of the respondents who take it, while the amount of the characteristic of the respondents is assessed by their performance on the instrument [6]. The inherent confounding between instrument and sample makes comparisons between different instruments related to the same characteristic and between groups of respondents having different amounts of that characteristic a challenge, although, if we stick to just one instrument form, then the situation is not so complex. In addition, treating the raw scores as if they were linear measures with the same standard error throughout their range biases statistical methods based on that assumption [7].

In contrast, IRM uses item responses to create a linear (logit) scale that represents 'less' to 'more' of a characteristic or latent variable like a particular ability, trait or attitude, to name a few possibilities. Because of this linearity, the relationship between respondent location and item location on the scale of that latent variable can be compared directly [7]. [We assume that the reader has already read an introduction to item response modeling, such as in Wilson *et al.* [1, 8] (this volume), or is otherwise familiar with it.] The location of an item is modeled to be independent of the locations of the respondents in the sense that any respondent in the group, at any location, has an estimated probability of endorsing that item. Item responses on instruments are used to estimate item locations and standard errors and respondent locations and standard errors. The advantage of locating items on the same scale as respondents is not cost free: IRM requires that items perform in ways that conform to certain assumptions (i.e. they must have reasonably good 'fit'). The resulting scale can be interpreted as indicating the probability that a particular respondent with a particular estimated location will endorse a given item. With these conceptual differences in mind, a more direct comparison of the standard methods used in the classical approach and item response modeling becomes possible.

In a classical analysis, the investigator uses raw scores to compute statistics such as means, variances, reliability coefficients, item discrimination measures, point-biserial statistics for item categories, total scores and errors of measurement for the instrument as a whole. In an item response analysis, the investigator uses raw scores to estimate item and respondent locations plus all standard errors. These parameters can be used to calculate the equivalent of the statistics mentioned from classical analysis, as well as additional ones such as the variation in standard errors and thus, the information available across the range of the latent variable. The focus on items in the item response analysis also allows a more extensive assessment of the functioning of response choices within items—the item categories—and the coverage of content the instrument is supposed to measure.

The data and instrument

A specific instrument and data set will be used to illustrate the similarities and differences between the two approaches. The data were provided by the Behavior Change Consortium (BCC) [9], which collected data from 15 different projects explicitly studying major theoretical approaches to behavior change and the interventions related to them. The different projects investigated mediators (or mechanisms) of behavioral change interventions directed at tobacco use, sedentary lifestyle and poor diet. A common hypothesized mediator for change in many of the studies was self-efficacy. Our analyses included only baseline data regarding self-efficacy; no post-intervention data from the BCC were included for this paper.

Self-efficacy is defined as 'a specific belief in one's ability to perform a particular behavior' [10, p. 396]. One instrument developed to measure self-efficacy for exercise, 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' [10, 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'. The instrument authors used summary scores that averaged the responses across items if at least 13 items were completed [10].

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

Two of the projects used the same SE scale for exercise [10] to assess the amount of self-efficacy subjects said they had, and related that to changes in the dependent variables surrounding exercise program persistence. Investigators for both projects postulated self-efficacy as a hypothesized mediating variable, speculating that people who felt confident that they could maintain an exercise program through various adverse conditions would be more likely to reap the health benefits of increased physical activity [11].

The original article describing the SE scale [10] reported a typical summary of the information from a CTT analysis: the total score at baseline averaged 74.3% confident with a standard deviation of 16.72. The internal consistency of the scale was 0.90, and the test–retest correlation was 0.67 (n = 62, $p <$ 0.001), although the time period between tests was not explicitly stated. Baseline SE scale scores were positively correlated with adherence to an exercise program in both the first and last 6-month periods during the 1-year assessment (r = 0.42 and 0.44, respectively), and added significantly to the model of adherence in a multiple regression analysis in which self-motivation was insignificant [10].

Note that the wording of certain items (Items 6 and 13) for data gathering in one project used 'exercise' in the text (this is what is shown in Table I), as opposed to the term 'physically active' which was used in the gathering of data in the second project. For the purposes of this paper, we assumed that this word change did not affect the results.

The classical and item response analyses for this paper were performed using the software package ConQuest [12]. The majority of the item response analyses was described in another paper in this volume [1]. Any extensions are described when they are discussed below. The classical analyses are quite standard, and are no different from those outlined in a standard text such as that of Cronbach [4]. Details necessary for interpreting the results are discussed below.

Results: comparing analyses

The scheme for comparing the two approaches utilizes psychometric concepts of model fit and common aspects of reliability and validity to perform a parallel assessment of the SE scale. The particular aspects of reliability and validity incorporated are based on the terminology and scheme proposed in the latest edition of Standards for Educational and Psychological Tests [2]. The scheme has been described in detail by Wilson [13]:

(i) choosing a model [13, Chapter 6]
(ii) evidence for reliability and measurement error [13, Chapter 7]
(iii) evidence for validity [13, Chapter 8], including
  (a) evidence based on instrument content,
  (b) evidence based on response processes,
  (c) evidence based on internal structure,
  (d) evidence based on other variables and
  (e) evidence based on the consequences of using a particular instrument.

Commonly used alternative terms for (a) through (e) include content validity, evidence based on cognitive interviews or think alouds, construct validity, external or criterion validity and consequential validity. The results of the comparisons are summarized in Table II.

Choosing a model

The classical true score model, $X = T + E$, is not one that can be formally rejected since it is one equation with two unknowns. Hence, there is no step of 'model choice' in the classical approach. To

Table II. *Standards framework for comparing the two approaches*

Comparison framework	CTT approach	Item response modeling approach
Choosing a model	The same model is always 'chosen' $X = T + E$	'Fit' of persons and items to specific model can be calculated and evaluated—may be informative Alternative models can be compared, to explore measurement implications
Evidence for reliability		
Reliability coefficients	Cronbach's α = 0.91	MML reliability = 0.92
Standard error of measurement	Constant value = 7.66	Varies by raw score, see Fig. 2 Will be important when the shape of the curve has measurement consequences (see text for examples)
Evidence for validity		
Based on instrument content	No contribution	Modeling item endorsability may influence choice of items and response categories to span the range of levels needed to cover content
Based on response processes	Not relevant	Not relevant (yet)
Based on internal structure	No equivalent to Wright map Item discrimination index and point-biserial correlation can be used for item analysis DIF not available in CTT (although it could be addressed using other methods)	Wright maps provide a basic tool Mean for respondents in each category can be used for item analysis DIF can be addressed directly as an item parameter
Based on other variables	Same	Same
Based on the consequences of using an instrument	Same	Same

illustrate the steps one would take to choose a model in the case of the IRM approach; one can first examine the section of Wilson *et al.* [1] labeled 'fit statistics', where the fit of both items and persons to the partial credit model was examined for the SE scale. The details of that will not be repeated here, but, in summary, note that (i) *all* the items appeared to be fitting reasonably well and (ii) the information about person fit can be used to flag individuals for whom an estimated location was perhaps not sufficient to convey an adequate summary of their full set of responses. Note that in the classical context, no results address how well the items are represented, partly, at least, because formally there are no items in the CTT model itself. Of course, under the rubric of 'item analysis', several characteristics of items are indeed examined, and one could argue that these constitute a sort of 'model fit' (more on this later).

Finding evidence for misfitting items can help the instrument designer understand the essential nature of the set of items (e.g. lack of unidimensionality) that are defining a latent variable. Misfitting respondents can alert the measurer to alternative points of view or response styles that were not considered in the design of the original instrument. A different type of fit issue can be addressed by asking if an alternative model would have done a better job. For example, the partial credit model (used above) allows each item to have a different pattern of relative differences in endorsability when making the transition from one category to the next (the 'item step parameters' as defined in Wilson *et al.* [1]). A more parsimonious model, called the 'rating scale model' [14], allows each item to differ in its overall endorsability, but constrains the relative endorsability of the steps to be the same across all the items. Some researchers find that this is a reasonable model to consider as an alternative as the labels of the response alternatives for the SE scale items are identical across items [14]. The different item stems may not interact with the way the respondents interpret the percentages, hence, one might expect that the parameters that govern the relative endorsability of, say, 80 versus 90% might be approximately equal across items. Two ways to examine the difference in fit of the two models are (i) to compare them using a likelihood ratio test (possible because the rating scale model is a constrained version of the partial credit model) and/or (ii) to compare them using the same fit indices as were used in the previous paragraph. The likelihood ratio test can be calculated by finding the difference between the deviances (i.e. twice the loglikelihood) for the two models (provided by the ConQuest computer program). In this case, it turns out to be 336.23 (df = 117), and when this is compared with the critical value for a χ^2 distribution at $\alpha = 0.0001$ (which is 182.61), we can see that the difference is indeed highly statistically significant. The same conclusion is drawn upon noting that the fit statistics for the step parameters, which were found to be innocuous for the partial credit model, are all above the usual maximum for the rating scale model. The effect size for this can be gauged by looking at the pattern of thresholds in a graph showing the locations of respondents and item thresholds, the Wright map shown in Fig. 1 (this was introduced in Wilson *et al.* [1]). Note that there are considerable differences in the relative location of the thresholds across items—e.g. between Thresholds 9 and 10 in the columns for Items 1 and 13. These differences could imply quite considerable interpretational differences between scales resulting from the two models. Hence, we can conclude that the use of a uniform set of options ('10%', '20%', etc.) for the SE scale has not resulted in a uniform pattern of responses from the respondents.

Evidence for reliability

The consistency with which respondents are measured is commonly reported in two ways: (i) using a reliability coefficient, which attempts to give an overall indication of how consistent the respondents' responses are and (ii) the standard error of measurement, which attempts to indicate the amount of variation one might expect, given a certain pattern of responses across the items. Under the classical approach, the internal consistency reliability coefficient most commonly used for polytomous data is Cronbach's α [4]. In this case, it

```
Logit  Respondents                                   Items
                 1    2    3    4    5    5    7    8    9    10    11    12    13    14
-----------------------------------------------------------------------------------------
                |                    10                  10                            10
                |          10                                 10
                |                         10
                |     10                                                  10
                |10                            10   10
                |
                |
                |               10                                  10
                |
               X|                                                                      9
                |                                                               10
               X|9
                |                     9                   9
               X|
   1           X|          9
               X|
               X|     9                    9         9                    9
               X|                                              9
              XX|               9             9         8                             8
              XX|
              XX|8                    8                            9
              XX|
            XXXX|          8                            7                      9      7
           XXXXX|
            XXXX|7    8                       8
         XXXXXXX|               8    7    8         8                    8
         XXXXXXX|6         7                            6     8                       6
         XXXXXXX|                     6       7                    8
      XXXXXXXXXX|     7     6    7                   7                    7     8
   0    XXXXXXXX|                          7
         XXXXXXX|               6             6     6   5     7    7     6     7
         XXXXXXX|5    6               5                                               5
       XXXXXXXXX|          5               6                   6    6           6
        XXXXXXXX|               5             5     5   4                5
       XXXXXXXXX|4    5     4                                                  5      4
          XXXXXX|               4    4                             5           4
           XXXXX|     4                    5   4     4   3     5    4     4            3
            XXXX|               3    3                                         3
             XXX|3    3     3              4         3         4    3           2
             XXX|     2          2                                 2     3
              XX|          2               3   3         2     3
               X|                     2              2         2                       2
               X|2                         2   2                          2
               X|
                |
  -1            |
                |
                |
                |
                |
                |1
                |                     1                  1
                |          1
                |
                |               1                                                      1
                |
               X|
                |
                |                                    1
                |     1                     1   1               1     1      1     1
```

Fig. 1. Wright map of item thresholds for SE scale analyzed polytomously (each ‘X’ represents 3.7 cases).

was calculated to be 0.91. For the item response approach, an equivalent coefficient can be calculated as a by-product of the marginal maximum likelihood (MML) estimation algorithm, and turns out in this case to be a very similar 0.92. This reliability value can be used to predict the effect on reliability of reducing or increasing the number of items using the Spearman–Brown formula [4]. For the SE scale, starting from $r = 0.92$, the reliability would be predicted to be reduced to 0.91 by deleting one item, to 0.89 by using only 10 items and down to 0.85 using just half the items.

The classical standard error of measurement is calculated as a function of the reliability coefficient and the standard deviation of the raw scores. It is, by assumption, a constant, not varying for different scores. In this case, it turns out to be 7.66 (in raw score units). In the item response modeling approach, the relationship between the estimated location and the standard error of measurement is not a constant, but varies with the location of the respondent (thus it is also called the 'conditional' standard error of measurement). This relationship results from the proximity of the respondent's location and the item parameters (usually, there are more item parameters estimated toward the middle, hence, the relationship is usually 'U'-shaped). The specific relationship for the SE scale data is shown in Fig. 2. Because of the non-linear relationship between the raw score metric and the logit metric, it is not straightforward to compare the values of these two standard errors of measurement. Of greater import is the shape of the relationship between the standard error of measurement and location, which can be informative in different measurement contexts. For example, if the tails lift too high, then that might indicate that measurements at the extremes are to be treated with caution. Or, if one is using a cut score, then the relationship could be used to examine whether the particular set of items used was an optimal set (i.e. by examining how close the minimum point is to the cut). Of course, if the non-linearity of the relationship is important, as it is in these two instances, then the linearity assumption of the classical approach is a drawback.

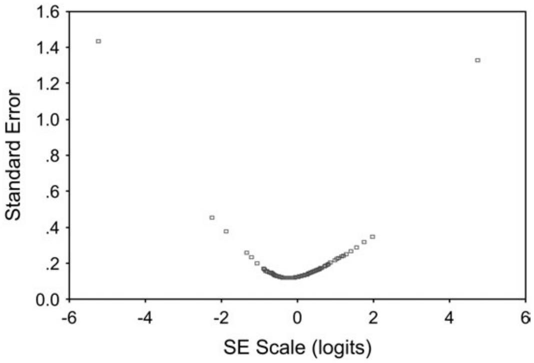

Fig. 2. The standard error of measurement for the SE scale (each dot represents a different score).

Evidence for validity

Note that the structure of the discussion about validity below is based on the 1999 Standards for Educational and Psychological Testing [2]. These differ quite markedly from the structures presented in earlier editions of the 'Standards'. Those more familiar with the older categories may wish to update their knowledge before reading further.

Evidence based on instrument content

It is quite possible to develop an instrument's content in the same way regardless of the potential application of either a classical or an item response modeling approach. The intent is to formulate items that 'cover' all areas of the content of interest, and, in the CTT approach, is frequently performed by topic area or other subdivision. However, to do so is to ignore one of the major advantages of item response modeling. The focus of item response modeling on 'the item' gives a direct connection between the meaning of the item content and the location of the item on the latent variable. The Wright map illustrates the connection between the latent variable and item and respondent locations, enabling a very rich repertoire of interpretations of the relative locations of respondents and items (see Wilson *et al.* [1] for more on this).

The SE scale is a negative example of this criterion—the scale is dominated by the response categories rather than the items (as is clearly shown in Fig. 1). The categories, '0%', '10%', etc., like many merely numbered categories, are not conducive to meaningful interpretation. Effectively, the respondent is left to intuit what relative differences in '10% of self-efficacy' might be and respond accordingly. That thinness of interpretational possibilities limits the possibilities for the user to interpret the results, and also limits the possibilities for someone gathering internal validity evidence (made evident in a later section).

In contrast, there is a strong tradition of instruments being designed with the intent of building in strong content, and matching that content with features of a measurement model. This dates back to the work of Guttman [15], and has reached prominence in the work of Wright *et al.* (e.g. [16, 17]). A recent account that builds on this work [13] shows multiple examples of such content structures (called 'construct maps') in which the relative behaviors of a generic respondent having various amounts of the construct are arranged in order on one side, and potential items are arranged in order of endorsability on the other.

Evidence for validity based on response processes

Evidence based on the response process consists of studies of how respondents react to the items and the instrument as a whole. These might consist of 'think alouds', 'exit interviews' or 'cognitive interviews' with samples of respondents. To date, there have been no studies that explicitly relate this type of evidence with the features of measurement models, and such evidence was not available with our secondary analysis of the baseline SE scale data, so this aspect of the framework is not relevant to our comparison at this time. It is possible that this connection may be made in the future, however.

Evidence based on internal structure

One major criterion for internal or construct validity is an *a priori* theoretically based hypothesis about the order of item endorsability, the ease with which respondents rate items strongly. A very useful tool for investigating this is the Wright map (discussed in Wilson *et al.* [1]; also see Wilson [13], Chapter 6). As mentioned above, in the case of the SE scale, no such expectations were developed. Hence, this source of validity evidence is not available. Even if there were, the Wright map shown in Fig. 1 demonstrates that there is no discernible empirical order to the items of the SE scale. Instead, the feature of the SE scale items that maps out the SE scale variable is the transitions between the categories. Unfortunately, the labels chosen for these categories, 10%, 20%, etc., are not interpretable. The one thing you might expect would be that the categories line up across the page, but that is clearly not the case here, other than for the extreme categories. For examples of cases where the Wright map has been used successfully as a support for internal validity, see Wilson [13].

The Wright map is useful for a number of other purposes, as well. For example, one generic threat to the usefulness of an instrument is a mismatch between the locations of the items and the respondents on this map. Examination of the Wright map for the SE scale shows that the instrument is free from certain types of problems that sometimes occur in instrument development: (i) there are no significant gaps in the locations of the item thresholds and (ii) the range of the item parameter locations matches quite well the spread of the respondent locations. For illustrations of what these problems look like on a Wright map, and what it means for the instrument, see Wilson [13].

Other evidence that the items are operating as intended is available in somewhat different forms from both the classical and item response modeling approaches. An important piece of evidence that an item is functioning as expected is that the increasing response levels of the item are operating consistently with the instrument as a whole. The CTT indicator of this at the item level is the item discrimination index (Table III). The CTT indicator at the category level is the point-biserial correlation between the respondent's choice of that category and the raw score. As the categories increase from lowest to highest, if they are working well, then one would expect these correlations to increase from negative to positive. Where they do not, that is

Table III. *Item point-biserial correlations and item discrimination (disc.) for the SE scale*

Item	Item disc.	Point-biserial correlation for each category										
		0	10	20	30	40	50	60	70	80	90	100
1	0.59	−0.31	−0.26	−0.17	−0.17	−0.06	0.04	0.14	0.21	0.29	**0.21**	**0.06**
2	0.75	−0.30	−**0.36**	−0.17	−0.17	−**0.22**	−0.08	0.03	0.17	0.25	0.30	0.36
3	0.75	−0.32	−**0.38**	−0.16	−**0.17**	−0.15	0.00	0.07	0.24	0.27	**0.26**	0.33
4	0.67	−0.30	−**0.32**	−0.19	−0.10	−0.07	−**0.10**	0.00	0.13	0.23	0.28	0.30
5	0.67	−0.29	−0.28	−0.19	−0.16	−0.13	0.01	0.09	0.22	0.26	**0.24**	0.26
6	0.71	−0.27	−**0.29**	−0.22	−**0.26**	−0.15	−0.12	−0.01	0.12	0.24	0.32	**0.30**
7	0.55	−0.16	−**0.27**	−0.16	−**0.18**	−0.17	−0.01	0.13	**0.05**	0.20	0.21	0.24
8	0.73	−0.34	−0.29	−0.21	−0.18	−**0.19**	−0.05	0.01	0.13	0.26	0.32	0.33
9	0.65	−0.35	−0.27	−0.17	−0.12	−0.08	0.09	0.15	0.17	0.23	0.24	0.24
10	0.71	−0.30	−**0.33**	−0.13	−**0.24**	−**0.17**	−0.14	−0.05	0.16	0.22	0.29	0.34
11	0.71	−0.31	−**0.33**	−0.17	−**0.20**	−0.16	−0.15	−0.03	0.08	0.27	**0.26**	0.35
12	0.68	−0.29	−**0.33**	−0.18	−**0.21**	−0.12	−0.08	0.06	0.15	0.29	**0.22**	**0.26**
13	0.67	−0.29	−0.28	−0.22	−**0.25**	−0.16	−0.11	−0.07	0.09	0.23	0.26	0.35
14	0.70	−0.33	−0.31	−0.19	−0.19	−0.07	0.07	0.16	0.19	0.32	**0.23**	**0.20**

Values that appear to be out of increasing order from left to right for any item are shown in bold.

evidence that the successive categories are not working as they should. Table III shows the point-biserial correlations for the SE scale data. Although the discrimination indexes seem to indicate that the items are all acting quite well, there are numerous cases where the expected order is not observed (cases where the right-hand values are less than the left-hand values are shown in bold in the table). In particular, there appears to be some fairly consistent problem between the first and second categories. However, this seems inconsistent with the information provided by the item discrimination index, and may be due to problems in interpreting correlation coefficients in small or restricted samples.

The analogous information for the item response modeling approach is shown in Table IV: these are the means of the locations of the respondents in each category. As can be seen in Table IV, there are far fewer instances where the order is other than expected, and given the small number of cases in some of the categories (especially at the extremes), these instances can be largely ignored. The mean is an inherently simpler index than the correlation coefficient, and may be giving a clearer picture in this case [13].

One of the fundamental assumptions of IRMs is that the item response function (IRF) (discussed in Wilson *et al.* [1]) is invariant throughout the population being measured. If the IRF differs according to which subgroup a respondent is in, then that is referred to as ‘differential item functioning’ (DIF). The most common way to think about this is that an item would be ‘harder’ for similar people from one group than from another. (i.e. harder to endorse at a higher level, etc.). DIF does not require that the subgroups differ in their mean scale locations. When subgroups have different mean scale locations on the latent variable, this is commonly referred to as ‘differential impact’. There are several ways to investigate DIF under an item response modeling approach. One approach that is particularly straightforward is to add an item by group interaction parameter, γ_{ig} into the underlying relationship. Thus, the IRF without DIF is given as in Equation 1,

$$\text{Probability}\,(X_i = 1\,|\,\theta,\, \delta_i) = \frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}}, \qquad (1)$$

where θ is the person estimate and δ_i is the item endorsability parameter. Then the relationship

Table IV. *Means of SE scale locations for respondents selecting each category*

Item	Mean respondent location for each category										
	0	10	20	30	40	50	60	70	80	90	100
1	−1.03	−0.40	−0.27	−0.20	−0.09	0.03	0.20	0.28	0.38	0.62	0.74
2	−1.75	−0.44	−0.40	−0.29	−0.29	−0.10	0.00	0.16	0.24	0.48	1.37
3	−1.14	−0.49	−0.31	−0.22	−0.19	−0.02	0.07	0.25	0.30	0.49	1.59
4	−1.18	−0.39	−0.39	−0.17	−0.10	**−0.12**	−0.02	0.17	0.23	0.43	0.90
5	−1.07	−0.40	−0.29	−0.22	−0.14	0.00	0.09	0.25	0.37	0.50	1.72
6	−2.53	−0.42	**−0.57**	−0.45	−0.22	−0.14	−0.04	0.11	0.20	0.50	1.22
7	−1.52	−0.39	−0.38	−0.23	**−0.24**	−0.01	0.16	**0.04**	0.27	0.29	0.92
8	−1.77	−0.37	**−0.41**	−0.27	−0.25	−0.06	−0.02	0.12	0.25	0.49	1.17
9	−1.11	−0.35	−0.22	−0.17	−0.11	0.07	0.17	0.24	0.43	0.52	2.24
10	−3.53	−0.44	−0.36	**−0.40**	−0.25	−0.15	−0.08	0.15	0.18	0.38	1.33
11	−2.34	−0.38	**−0.49**	−0.36	−0.23	−0.17	−0.07	0.06	0.25	0.31	0.92
12	−2.09	−0.45	−0.35	−0.30	−0.16	−0.10	0.07	0.17	0.29	0.32	1.04
13	−3.01	−0.30	**−0.56**	−0.37	−0.28	−0.16	−0.12	0.09	0.21	0.31	0.73
14	−1.32	−0.41	−0.24	**−0.25**	−0.11	0.03	0.17	0.28	0.52	0.63	2.46

Values that appear to be out of increasing order from left to right are shown in bold.

incorporating the DIF effect is expressed in Equation 2:

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

Gender was chosen to illustrate IRM analysis of DIF because it is binary and thus simpler than age or race, for example. Estimates of the item by gender interactions provided by the ConQuest software [12] for the SE scale, along with their standard errors, are shown in Table V. Calculation of the approximate 95% confidence intervals for these interaction effects using the usual formula (estimate plus or minus 1.96 times the standard error) shows that all the confidence intervals contain zero, and an omnibus test of parameter equality gives a χ^2 statistic of 13.021, on 14 df ($p = 0.525$). Thus, the SE scale did not display any statistically significant DIF with respect to gender in this sample. Examples where DIF has been found to be important, and the implications of this for the instrument, are discussed in Wilson [13].

Table V. *Estimates and standard errors for DIF parameters indicating gender–item interaction from IRM analysis of the SE scale (*n *= 504, female = 394)*

Item	Estimate	Standard error
1	0.003	0.022
2	−0.007	0.022
3	0.002	0.022
4	−0.001	0.022
5	0.018	0.022
6	0.026	0.023
7	−0.039	0.022
8	−0.024	0.022
9	0.043	0.022
10	−0.037	0.023
11	0.001	0.023
12	0.002	0.022
13	0.006	0.023
14	0.006	0.022

No estimate is statistically significant.

The discussion of how an issue like DIF can be incorporated directly into the IRM is intended as an illustration of one of the general strengths of this approach: IRMs can be expanded to investigate theoretical and measurement complexities. In contrast, investigation of DIF is not available through any of the standard statistics of the classical approach. While one could use an additional technique, such as logistic regression using the raw scores, to look for evidence of DIF, it would not link directly to the CTT approach (as it does not utilize the standard error of measurement in the analysis). Other DIF possibilities arising out of the origins of the CTT approach include structural equation modeling (SEM) and factor analysis to see if responses differ by subgroup.

Evidence based on other variables

Commonly called 'external validity', this form of evidence is examined in a similar way in both approaches by comparing respondents' measures on the instrument of interest with their behavior or responses to other instruments. There is quite a long list of validity studies available for different versions of the SE scale, many of these have provided evidence for the relationship of the SE scale with physical activity behavior [18] or assessed its ability to predict change in physical activity behavior [19–27]. In the item response modeling approach, this would be carried out in a very similar manner, differing only in that it would use respondent location instead of total score for comparison.

Evidence based on consequences of using an instrument

The final form of validity evidence relates to consequences. In some senses, this is the most important type of evidence. For example, if the use of an instrument was to differentiate accurately between groups of people who would respond well to intervention, but its inaccuracy in differentiating led to false exclusion or inclusion of large numbers of people, then the instrument can hardly be said to be successful, no matter what the other forms of evidence say. However, as with the previous form of validity evidence, there are no major differences in how it would be investigated under the two approaches.

Discussion

This paper has addressed several important points in the comparison of the CTT and IRM approaches to measurement. There were a number of ways in which the CTT and IRM approaches were consistent in the sorts of issues that they addressed and the results they obtained. For example, the CTT concepts of reliability and standard error of measurement have equivalents in the IRM approach. The reliability for the SE scale was found to be similar under both approaches. The standard error of measurement was expressed in different metrics under the two approaches, but the reliability indicates that they were effectively not very different. Another example of method concordance is evidence concerning external variables. Both approaches use correlations between criterion variables and either the raw score or the respondent locations. There are ways in which the item response modeling approach can be extended in order to estimate better the underlying relationship using multilevel extensions of the IRMs [28], but, at a conceptual level, the operations are basically the same.

There were some interesting and important differences between the two approaches, with IRM generally extending the CTT approach. For example, the concept of model fit is not available with the classical $X = T + E$. Although this simplifies the application of the classical approach, it is also an important limitation, depriving the analyst of the creative power of model building. A second important feature in the IRM approach is the common scale for respondents and items, embodied in the Wright map. This gives an immediacy of interpretation that allows non-technical intuitions constructively to inform instrument development and interpretation. For example, if the IRM approach had been used in the development of the SE scale, an interpretational perspective might have been developed that would make the process of interpreting results from the SE scale a more intuitive and useful exercise.

By designing instruments according to the ideas of criterion referencing as pioneered by Wright [7], the unique properties of the Rasch model can be used to help build rich interpretations into measurement. For example, response categories that lend themselves to greater interpretation could help researchers to understand what aspects of self-efficacy are easy to endorse, and which can be endorsed only by those who have a great deal of confidence in their ability to overcome obstacles to exercising. Looking in more detail into the standard error of measurement, the fact that the error varies across the construct, as revealed in the IRM approach, indicates that there can be important differences in how the two approaches deal with reliability. For example, those who have very high or very low self-efficacy have higher standard errors of measurement, so any interpretation of association between these scores and behavioral change should be somewhat more suspect.

There are important aspects of validity evidence that are not different between the approaches. For example, evidence based on external variables, evidence based on response processes and evidence based on consequences are all areas where the specific measurement approach adopted is not markedly important to the use of such evidence.

Finally, the evidence that is usually gathered in studies of instrument validity in the area of health outcomes research does not address all aspects of validity proposed in the accepted Standards [2]. This is particularly noticeable for evidence based on response processes and evidence based on consequences. Neither the original reports on the instrument nor the BCC data include these types of evidence (although, regarding the latter, such evidence would not contribute much to a comparison of the two approaches).

Perhaps the most important difference between the CTT and the IRM approaches is the DIF. DIF detection is essentially absent from the classical approach, partly at least, because items themselves are not a formal part of CTT. A measurer may use SEM or factor analysis for ways to determine DIF, but direct assessment of a DIF parameter is readily available within the IRM approach. Thus, the item response modeling approach can do *all* that you can do in the classical approach when it comes to assessing items and instruments, and it can do a great deal 'more'.

There could be ways to extend the classical approach to achieve many, perhaps even most, of the possibilities of item response modeling. In fact, many of the techniques that are the bread and butter of large-scale analysis using IRMs (such as, say, different equating techniques) are available in one form or another in extensions of the standard classical approach. One example is generalizability theory, which can offer many useful insights—unfortunately, this and many other topics such as computerized adaptive testing are beyond the scope of an introductory paper. But the problem is that each of these extensions involves a unique set of extra assumptions that pertain only to that extension. The advantage of the item response modeling approach is its ability to be extended to incorporate new features of the measurement situation into the model, features such as extra 'facets' of measurement (like raters), extra dimensions (e.g. behavior as well as attitude) or higher-level units of observations (e.g. families or medical practices). Some possibilities for these extensions are given in successor papers to this one. A more comprehensive account of such extensions is given in De Boeck and Wilson [29].

Final note

So, what has been learned by analyzing the SE scale data using an IRM approach that one would not have gained by using a CTT approach? If a measurer wishes to order item presentation from those that are easiest to those that are hardest to endorse, the results of an IRM analysis of empirical data can establish such an order for future studies; future researchers must determine if this is valuable. Possibly, the most important insight has been gained not in deep technical analysis, but from a diagram. The Wright map (Fig. 1) showed that there is ample scope for incorporating meaningful category levels into the SE scale, but the current method of labeling the categories as percentages allowed for little in the way of interpretation. Perhaps, if the IRM approach had been available to the instrument's authors, a richer basis for interpretation would have been built in from the beginning, possibly with fewer categories per item but each associated with a meaningful level of self-efficacy. The Wright map also showed that the categories were well-located with respect to the respondent locations, and also that there were no important gaps in the coverage of the latent variable, both positive features of the SE scale. The more detailed information about standard error of measurement indicated that the measurement at the two extremes was much less accurate than in the middle. (From Fig. 2, one can see that the measurement error at the extremes ranged from two to three times larger than the minimum, excluding the two most extreme data points.) The analyses indicated that with the current number of categories, several items could be dropped to shorten the measure with very little change in overall reliability, but the paucity of information already noted at the extremes of the range indicates the danger in assuming accuracy throughout the range would likewise be unaffected. Examination of the Wright map may give clues as to which items provide the most redundant information across the span of the content if a discerning researcher proposed to delete any items. The DIF analysis gave some comfort, as it showed that the SE scale was acting in a fairly consistent way, at least with respect to males and females. The CTT analysis, in terms of the point-biserial correlations, would likely lead to some concern about the respondent interpretations of the categories, but this was shown to be somewhat less problematical in the IRM analysis. Nevertheless, the IRM analysis indicates possible problems with respondent interpretations particularly of 20 and 30% response categories, so future investigations of the SE scale might start by considering alternative ways to label categories of responses.

Acknowledgements

We would like to thank Louise Mâsse formerly of the National Cancer Institute (NCI), 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. Support for this project was provided by NCI (Contract No. 263-MQ-31958). We thank the Behavioral Change Consortium 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. Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Educ Res* 2006; **21**(Suppl 1): i4–i18.
2. American Educational Research Association, American Psychological Association, National Council for Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 1999.
3. Traub RE. Classical test theory in historical perspective. *Educ Meas Issues Pract* 1997; **16**: 8–14.
4. Cronbach LJ. *Essentials of Psychological Testing*, 5th edn. New York: Harper & Row, 1990.
5. Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904; **15**: 72–101.
6. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, 1991.
7. Wright B. A history of social science measurement. *Educ Meas Issues Pract* 1997; **16**: 33–45, 52.
8. 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.
9. 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.
10. Garcia AW, King AC. Predicting long-term adherence to aerobic exercise: a comparison of two models. *J Sport Exerc Psychol* 1991; **13**: 394–410.
11. 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.
12. Wu ML, Adams RJ, Wilson MR. *ACER ConQuest: Generalised Item Response Modelling Software* [computer program, Version]. Hawthorn, Australia: ACER (Australian Council for Educational Research) Press, 1998.
13. Wilson M. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum, 2005.
14. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**: 561–73.
15. Guttman L. A basis for scaling qualitative data. *Am Sociol Rev* 1944; **9**: 139–50.
16. Wright BD, Stone M. *Best Test Design*. Chicago, IL: MESA Press, 1979.
17. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: MESA Press, 1982.
18. Trost SG, Owen N, Bauman AE *et al.* Correlates of adults' participation in physical activity: review and update. *Med Sci Sports Exerc* 2002; **34**: 1996–2001.

19. Poag-DuCharme KA, Brawley LR. Self-efficacy theory: use in the prediction of exercise behavior in the community setting. *J Appl Sport Psychol* 1993; **5**: 178–94.
20. Sallis JF, Hovell MF, Hofstetter CR. Predictors of adoption and maintenance of vigorous physical activity in men and women. *Prev Med* 1992; **21**: 237–51.
21. Sallis JF, Hovell MF, Hofstetter CR *et al.* Explanation of vigorous physical activity during two years using social learning variables. *Soc Sci Med* 1992; **34**: 25–32.
22. McAuley E. Self-efficacy and the maintenance of exercise participation in older adults. *J Behav Med* 1993; **16**: 103–13.
23. Marcus BH, Eaton CA, Rossi JS *et al.* Self-efficacy, decision-making, and stages of change: an integrative model of physical exercise. *J Appl Soc Psychol* 1994; **24**: 489–508.
24. Muto T, Saito T, Sakurai H. Factors associated with male workers' participation in regular physical activity. *Ind Health* 1996; **34**: 307–21.
25. McAuley E, Jacobson L. Self-efficacy and exercise participation in sedentary adult females. *Am J Health Promot* 1991; **5**: 185–91.
26. Miller YD, Trost SG, Brown WJ. Mediators of physical activity behavior change among women with young children. *Am J Prev Med* 2002; **23**: 98–103.
27. Wilbur J, Miller AM, Chandler P *et al.* Determinants of physical activity and adherence to a 24-week home-based walking program in African American and Caucasian women. *Res Nurs Health* 2003; **26**: 213–24.
28. Adams RJ, Wilson M, Wu M. Multilevel item response models: an approach to errors in variables regression. *J Educ Behav Stat* 1997; **22**: 47–76.
29. De Boeck P, Wilson M (eds). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer-Verlag, 2004.

Received on September 7, 2005; accepted on May 12, 2006