Comments on

Developing Tailored Instruments: Item Banking and Computerized Adaptive Assessment

Prepared for the conference "Advances in Health Outcomes Measurement," Bethesda, Maryland, June 23-25

David Thissen[1]

L.L. Thurstone Psychometric Laboratory

University of North Carolina at Chapel Hill

Item response theory (IRT) is a collection of statistical models and methods used for two broad classes of purposes: item analysis and test scoring. IRT models are mathematical descriptions of the (mini-)theory that the probability of a particular response to an item depends only on the level(s) of the respondent on the latent (unobserved) variable(s) being measured by a test or questionnaire and the characteristics of each item. Usually, item responses are answers to questions, but combinations of responses or other observations may be suitable data for IRT. Examples of latent (unobserved) variables measured by tests or questionnaires include verbal proficiency, depression, and headache impact.

Item analysis using IRT is based on a collection of item parameters, which have values that depend only on the item and the reference scale (which is usually defined with respect to some population; more will be said about that later). One or more parameters may be used for each item; taken together they are sufficient to characterize the probability of any particular item response as a function of the variable(s) the item measures, or indicates. A derivative function of

the item parameters that is particularly useful in any discussion of item banking or adaptive testing is the (Fisher) information function, that describes the quantity of additional precision that is obtained by presenting that item, or asking that question, at any particular level of the latent variable being measured.

Test scoring using IRT produces a numerical description of the distribution on the variable(s) being measured for persons a particular sequence of item responses. Usually that description includes a measure of central tendency (the average, either loosely speaking or not, as the IRT scale score), and a description of variability (which is reported as a standard error of measurement). The IRT scale score may be computed using only the item parameters and the responses of a single individual to any arbitrarily selected set of items.

The advantages of IRT over the traditional true score theory arise from this last point—the relative ease with which items may be mixed and matched to produce alternate forms or short forms of a test or questionnaire. The same technology we use to produce alternate 80-item forms of statewide grade school mathematics tests can be used to produce short forms of health status questionnaires that yield scores on the same scale as longer forms—it has been, in the case of the SF-36 and the SF-12 and the SF-8 (http://www.sf-36.org/). Orlando, Sherbourne, and Thissen (2000) describe the use of IRT to link scores on alternate forms of a depression instrument (with overlapping items). Carrying these ideas to an extreme, Paul Holland has remarked that computerized adaptive tests are simply (very large) collections of alternate forms (sometimes of various lengths).

Item Banks

The idea of an item bank was developed in Britain in the 1960s to provide educators with large pools of IRT-calibrated items that could be mixed-and-matched to provide classroom

assessments of achievement that could be scored on a common reference scale (Choppin 1968, 1976; Wood, 1976; Wood & Skurnik, 1969). Item banks are used widely in educational measurement, more or less visibly, as the sources of alternate paper-and-pencil test forms and/or adaptive tests that are scored on some common scale. The essential idea of this conference, and the recent "roadmap" initiative, is that these ideas may also be useful in the measurement of health outcomes. But that activity brings all of the details of IRT to the forefront of health outcomes measurement, leading to a few comments on various relevant topics:

The Reference Population

The father (or perhaps more accurately, one of the parents) of item banking, Robert Wood (1976), wrote that "The availability of trustworthy item parameter estimates is essential for any measurement application of item banking…" (p. 252). "Strictly speaking, it is not correct to say that the latent trait models provide invariant item parameter estimates. Only if a common scale… is used from group to group will this be true" (p. 254). In most applications of IRT this common scale is defined by the mean and standard deviation of some reference population.

Bjorner, Kosinski, and Ware (in press) and Bjorner (2004) suggested sensible alternatives for the reference population that define the metric for health outcomes IRT scale scores: They wrote "for generic health status measures it may be convenient to standardize… to a general population… for disease-specific concepts, the metric could be based on a well-defined population of people with the given disease." It is, in any event, important to remember that an IRT-scaled test, or item bank, *has* a reference population; and it is important to choose that population carefully.

As Bjorner *et al*. point out, it is not necessary for a sample from the reference population to respond to every item; there are many so-called *linking* data-collection designs that permit joint

calibration of smaller sets of items to create or expand one large bank. However, it must be understood that throughout the process of construction and maintenance of an item bank, linking must be used: Either sets of common items or common persons must provide a bridge to the common scale for every items' parameters.

In addition, it is probably wise to give some thought to the reporting metric in the construction either of tests or of the item bank itself. Test theorists often work in standardized (*z*) scores in which the reference population has a mean of zero and a variance of one. However, the decimals in such scores, and the negative scores for half of the population, make those an unfortunate choice for many uses. Bjorner (2004) suggests the time-honored and often-used *T* score scale (mean 50; standard deviation 10). However, the possible confusion of *T* scores with percentiles has led some to use alternatives (such as the *T* score scale with a decorative zero at the end of each score, otherwise known as the "College Board" scale).

Calibration: On Dimensionality

As mentioned earlier, Wood (1976, p. 252) observed that "The availability of trustworthy item parameter estimates is essential for any measurement application of item banking." For many unidimensional item response models, effective technology exists to provide item parameter estimates based on data from a calibration sample from the reference population, or a sample of data that includes both items that provide a link to the reference population and additional items. A literature of simulation studies indicates that sample sizes of the order of 500 to 1000 appropriately chosen respondents provide adequate data for item calibration; in some cases as few as 250 may be sufficient.

What does "appropriately chosen" mean? First, the calibration sample that defines the scale with respect to some defined population should be representative of that population. Second, it is

essential that the variation in the sample be sufficient to ensure that at least some respondents select each of the response options for each item. For example, if only individuals with extreme levels of dysfunction select "strongly agree" for some extremely worded items, the sample must include such individuals. That means either that a single sample must be sufficiently large to include such persons, or additional (linked) targeted samples must be drawn. Bjorner, Kosinski, and Ware (in press) make these points.

What if the construct to be measured is multidimensional? That is, what if there are distinct aspects of the construct such that some individuals may score high on one and low on another, while other individuals may show the opposite pattern? In this case, theory and practice diverge: There is multidimensional IRT (MIRT). However, except in fairly special cases current implementations in software do not provide the same kind of "trustworthy item parameter estimates" that can be obtained from unidimensional procedures. At this time, the best approach is to divide the items into unidimensional subsets, and (effectively) construct several item banks, and tests. This may take the form of several distinct item banks and tests, or a single bank in which items are coded according to their domain and a test that may appear unitary to the respondent, but in reality comprises several distinctly scored scales. In a computerized adaptive test (CAT), if there are several scales administered sequentially and the scores on those scales are correlated, it may be useful to choose initial items on some of the scales based on the score already obtained on other scales.

How does one know if the construct is multidimensional? Bjorner, Kosinski, and Ware (in press) observe that "There are no definitive rules for deciding when multidimensionality or local dependence is of sufficient magnitude to pose problems." Evaluation of dimensionality is like (indeed, is part of) evaluation of construct validity: There is no way to obtain a simple answer

that the construct underlying an item set is unidimensional or multidimensional. Instead, the data analyst accumulates evidence from various sources. Bjorner *et al*. recommend item factor analysis of tetrachoric or polychoric correlations, as well as comparisons between IRT slope parameter estimates computed from larger and smaller sets of the items. Both of those approaches provide useful evidence, and a greater quantity of useful evidence if they are combined with good substantive knowledge about the construct(s) being measured. There is, at this time, no substitute for skilled data analysis and judgment in support of a decision to treat a collection of items as unidimensional (locally independent, yielding one score) or a multidimensional set that might be best divided into subsets yielding several separately-scored tests.

(Admittedly) in an educational context, McLeod, Swygert, and Thissen (2001) and Swygert, McLeod, and Thissen (2001) illustrate the gradations of gray among tests of various degrees of unidimensionality through multidimensionality. Even though responses to a set of items may be *detectably* multidimensional, they may be only slightly so. In such cases, a scale with practical usefulness may be constructed by acting as though the item set is unidimensional. Stout's (1990) concept of "essential unidimensionality" provides a statistical definition of "slightly multidimensional" that may be useful in many contexts.

In addition, the assembly of a multi-scale battery does not preclude the subsequent computation of a single score if that proves useful as a summary or outcome index. The classical example is the intelligence, or IQ, score: Almost all intelligence tests are actually batteries comprising several scales, each with distinct scores. However, for many purposes a single linear composite of those scores (with such names as "full scale IQ") is a summary widely used, for

example, to predict academic performance. It is likely that in many health-related contexts, a similar "total score" may be useful.

That last word is what is most important in testing: What is useful? If one is measuring the outcome of an intervention or treatment, is it necessary to measure the distinct effects on several subscale scores, because those effects may differ, or is it the case that the treatment has similar effects on the collection of subscale scores, in which case the most powerful analysis would summarize that aggregate effect? The question "what is useful?" is not a question that test theory answers; it is a substantive question that gives *direction* to test theory in the construction of item banks and scales.

Many research groups in test theory (including my own) are actively engaged in the development of procedures that may, in the future, make feasible multidimensional item calibration and test scoring. However, even if that problem is solved, there remain other related challenges that may arise in the context of the measurement of health outcomes. For example, some disorders are best indicated by the presence of symptoms that are not strictly cumulative (that is, it is not the case that more severe levels of disorder are indicated by *more* symptoms; the symptoms may be exchangeable or alternative to each other). Historically the development of test theory has been driven by measurement problems presented by substantive research questions, and the measurement of health outcomes provides new problems to solve as well as new impetus to solve old problems. Increased support for purely methodological research in psychological measurement could produce results that would be broadly useful in research in the health sciences.

On Dichotomous and Polytomous Response Scales

Adaptive testing is most motivated in the context of items with dichotomous responses. The information function of IRT makes the reason clear: Highly discriminating dichotomous items provide information over a relatively narrow range of the variable being measured. Above and below that range, positive or negative responses are nearly certain and so no information is obtained by asking the question. Adaptive administration attempts to use items only for persons for whom they will be informative, and it can be successful in shortening the test by eliminating the administration of useless items.

The usefulness of adaptive testing is less clear for items that use Likert-type response scales. Indeed, Likert-type response scale items can be described as self-adapting: Persons low on the variable being measured may be choosing between "strongly disagree" and "disagree" while persons high on that variable may be choosing between "agree" and "strongly agree." There is a sense in which a single item with a Likert-type response scale is a conglomeration of several dichotomous items. The development of polytomous item response models makes that clear. The information functions for highly discriminating polytomous items may be relatively high over a relatively wide range of the scale. Explicit adaptation by item selection may not be so useful with such items.

However, there are circumstances under which scales comprising Likert-type items may usefully adapt, and these circumstances are likely to arise in the measurement of health outcomes. One such circumstance is when the range of the measurement scale is very wide, relative to the information "coverage" for any particular item. Bjorner, Kosinski, and Ware (2003) illustrate this in their Figure 1, showing the information curves for two items measuring headache impact: "I feel irritable because of my headaches" and "I feel desperate because of my

headaches." The first of those items yields some probability of a response more positive than "definitely false" for many in the general population. However, it massively understates headache impact for the subpopulation prone to migraines, for whom variation in responses to the second question yields information.

Another (albeit related) context in which adaptation may be useful even with polytomous responses arises for scales that are intended to cover a wide developmental range. (This is really just another situation in which a very broad range of variation is being measured.) For a scale that is intended to cover a wide developmental range, the dichotomous vs. polytomous response question may also reverse: It may be easier for very young or older persons to deal with dichotomous responses (yes/no, true/false) than with representations of a Likert-type scale, so one might be motivated to use adaptation as an alternative to polytomous response scales to obtain broad information coverage.

In comments on an earlier draft of this essay, Bryce Reeve (personal communication, May 24, 2004) asked whether it might be possible to have the same item calibrated with a dichotomous response format, and with a polytomous response scale, in the same item bank. I have not previously seen that idea proposed, and upon reflection I believe it would be rare that it would be *exactly* the same item stem, because some adaptation of wording would likely be required to make "yes/no" vs. "yes/sometimes/no" (or a five-point agree-disagree scale) semantically acceptable as response alternatives. Nevertheless, the two would be the "same item" (stem); but one would treat them as a special kind of pair of items in calibration. They would be a "special kind of" pair of items because they could never be administered to the same person without violating the assumption of local independence (indeed, a casual definition of perfect local dependence is "asking the same question twice"). But if one wanted to construct a scale

with dichotomous responses for one group of persons (say, children), but with the potentially greater response load and precision of the Likert-type response format for another group of persons (say, adults), then one certainly could calibrate a single stem with two different response scales as two distinct items—as long as one remembered to avoid asking the same person both questions in a pair.

Indeed, to expand upon this topic a little, Steinberg & Thissen (1996) have described research that compares the performance of (exactly) the same item stems with different numbers of Likert-type response alternatives. Comparing four-, to seven-alternative scales, we found that the levels of the latent variable associated with the response with the same label (like "agree") differed depending on how many other response alternatives were present! This leads to the comment that "the same" item with any change in the response format is best treated as a different item.

In addition to that thought-provoking question, Bryce Reeve (personal communication, May 24, 2004) also pointed out that a distinct advantage of the polytomous response format arises in the measurement of some health outcomes domains in which it is difficult to create large numbers of discrete items. For example, in the measurement of the consequences of pain, say from migraine headaches, how many ways are there to ask whether the pain distracts one from work? A graded response scale can easily be used to obtain data on the degree to which the pain is distracting, without repetitive, redundant questions. This use of Likert-type response scales is likely to be separate from adaptation, because it is a use that provides more precision of measurement on a scale that may be inherently short because only a small number of suitable items exist.

In any event, there are no significant differences between the styles of IRT item analysis and test scoring for scales that use polytomous-response items as opposed to those that are based on dichotomous responses. Bjorner, Kosinski, and Ware (in press) illustrate the use of Muraki's (1992, 1997) generalized partial credit (GPC) model with Likert-type responses; Samejima's (1969, 1997) graded model is a frequently used alternative.

Bjorner, Kosinski, and Ware (in press) suggest that a so-called nonparametric technique such as that provided by Ramsay (1997, 2000) can be used to confirm that the item response categories are approximately ordered as expected, before item calibration with parametric models is begun in earnest. That is excellent advice. A parametric alternative is to use Bock's (1972, 1997) "nominal" model which does not require that the response categories are ordered, but will show order if it is present. Note, however, that either of these alternatives actually requires more data than the basic graded item calibration. If each response category is fitted with a trace line that is effectively independent of all of the other response categories, then there must be sufficient data *in each response category* to permit estimation of the relationship (slope) of the response category as a function of the latent variable being measured. This requires (at least) tens of persons responding in each category (including the extreme categories), whereas the ordered models can be estimated with as few as one or two persons in an extreme category (assuming there are many more in other categories, of course!).

Differential Item Functioning (DIF) and (Possibly) Conditional Scoring

Bjorner, Kosinski, and Ware (in press) and others suggest that tests for differential item functioning (DIF) (Thissen, Steinberg, and Wainer, 1993) are useful parts of the item calibration process. That is certainly true, primarily because the presence of DIF is highly informative about the (potentially differential) validity of test items.

In the educational context in which concern about DIF originated, items that display DIF with respect to any of several demographic grouping variables are routinely eliminated from tests that measure achievement of aptitude. This is because items that exhibit DIF reduce the validity of tests if they are included, and scored identically. (Conventional notions of fairness associated with the use of educational tests as contests demand that items be scored identically for all examinees.)

However, in the context of the measurement of health or health outcomes, in which accurate measurement is most highly valued and issues of "fairness" are not likely to arise, it might be possible to use otherwise good items that exhibit DIF by explicitly modeling the DIF—that is, by using different parameters for the same item for different groups. For example, a class of items that is well known to exhibit DIF between men and women comprises items that ask about "crying" on depression scales. Most depression scales include such an item, and most exhibit DIF: different trace lines for men and women. Because crying is an indicator of depression, it may be desirable to use such items. But to compute depression scores that are comparable, it may be necessary to use different parameters for the crying item for men and women. This explicitly models the fact that (at least in this culture) asking men about crying is different (so it has different item parameters) than the same words used to ask women about crying.

In the educational context, DIF has been investigated primarily to examine the relation of item responses with demographic grouping variables such as sex or ethnic categorization. However, in the context of health outcomes measurement, it is important to consider the possibility that DIF may exist between diagnosed and healthy groups, or between individuals in different disease groups. Indeed, in the area of mental health measurement, both Hancock (1999) and Reeve (2000) detected DIF between diagnosed and "normal" groups on some items on the

MMPI-2. It is easy to imagine that similar DIF could appear between relatively healthy and diagnosed/treated groups in health outcomes measurement, by a very simple mechanism: Participants in medical care or treatment, be they practitioners or patients, may develop different interpretations of the meanings of health-related vocabulary. Then, when those words are used to frame questions or to define response alternatives, persons in the treated group may interpret the questions differently, and respond differently, than healthier persons. It may be very useful to investigate the possibility of DIF between diagnosed and "normal" groups in the item calibration process. In addition, with respect to the idea of the "reference population" discussed in earlier section of this essay, it is important to remember that when items are "moved" from calibration with one reference population to be used to measure members of another population, the possibility of DIF should also be considered and evaluated. (To be concrete, consider items calibrated with a cancer population subsequently used to measure persons in another chronic disease population; the item parameters may not be the same, and if they are not the same, that is DIF.)

Mode Effects

Bjorner, Kosinski, and Ware (in press) observe that "the data collection method for item calibration should be the same as the data collection method in the final CAT (most often a computer interface…)" This advice is offered to avoid problems with so-called "mode effects" (Mead and Drasgow, 1993) on the item parameters. Items may become more or less commonly endorsed, or more or less related to the latent variable they are intended to indicate, depending on the administration medium. Generally such effects are small, but to avoid problems it is best to follow the advice of Bjorner *et al*. The possibility exists that mode effects may also arise with alternative computerized presentation of items (e.g., for a questionnaire delivered directly to the

computer screen by local software vs. web-administration relying on a browser for display). That possibility has been studied little, if at all.

When results from computer-administered and paper-and-pencil questionnaires are to be compared, Steinberg, Thissen, and Wainer (1990, pp. 192-196) recommend that classical "multitrait-multimethod" (Campbell and Fiske, 1959) studies may be used to examine the degree to which the scores are comparable across presentation media.

Local Dependence and Context Effects

Mixing and matching items to assemble alternate forms, short forms, and computerized adaptive tests assumes that items are fungible, and their measurement properties remain as described by their item parameters regardless of context. This may not always be the case. Steinberg (1994, 2001) has documented small, but real, changes in item parameters as a function of context. Several studies of local dependence [including one of our own, by Chen and Thissen (1997)] have indicated that the responses to some items may interact with the responses to others, especially for very similar items or those that share some common stimulus.

Various "testlet" or "multi-stage" adaptive designs have been put forward to assemble tests in the context of local dependence or potential context effects (Armstrong, Jones, Koppel, and Pashley, 2004; Luecht, DeChamplain, and Nungester, 1997; Luecht and Nungester, 2000). These designs administer blocks of items, adapting only between blocks. The blocks are designed so that content balance is maintained (if needed), so context effects are minimized because an item is seen only in a few carefully considered contexts, and local dependence can be managed by keeping any such instances within blocks and scoring between blocks. While the problems that invite testlet or multi-stage adaptive designs may be more rare in the measurement of health outcomes than they are in education, it is good to know that alternatives are available if the

situation requires. Such testlet-based or multi-stage adaptive designs are usually longer than item-level adaptive tests, so these designs would only be used in contexts requiring the precision of relatively long scales.

Uses of Item Banks

As implied in the preceding sections, item banks may be used to support a wide variety of test assembly strategies: They may be used as the source of the item pools for conventional item-selection CATs—that is the most visible, headlining motivator for this conference. However, they can equally well be the source of items for various kinds of "testlet-based CATs" or multi-stage tests. Those may be computerized, and indistinguishable to the respondent from item-selection CATs; or, in sufficiently simple designs they may be administered as paper-and-pencil instruments that appear to have skip patterns.

Item banks may also be the source of a variety of fixed questionnaires that may vary in length, all scored on the same scale as a consequence of the underlying IRT models (Chang, 2004). Of course, measurement precision will vary with both the length of the scale and the type and degree to which it is adaptive. For fixed questionnaires of varying length, conventional methods can be used to compute reliability and measurement precision (which will be a function of the score level when IRT is used). For more tests with more complex structures, Bjorner, Kosinski, and Ware (in press) recommend computer simulations to determine measurement precision; that is likely the best solution to what otherwise could be an intractable problem.

In any event, item pools are the test-theoretic analog to the "pools" that were the source of life on earth a long time ago. The pools are the source of the items, with known properties—the lifeblood of psychometrically sound measurement.

Some Good, (Likely) New Ideas Arising in CATs for Health Outcomes, and Some New (Old) Problems

The development of CATs for health outcomes measurement has already contributed additional options for CAT design that have not received consideration in the educational context. It is always risky to point to an idea and pronounce it "new," but here are candidates:

Doubly-adaptive CATs (CAATs?). Surveys, including those documenting health behaviors and health outcomes, have often used so-called "skip-patterns" to ask more sensible series of questions (and to maintain the conversational nature of the interviewer-interviewee exchange in personal-interview surveys). It has been only natural to include skip patterns, as well as IRT-based item selection, in CATs for health measurement such as the one on the web at www.AmIHealthy.com. We note that skip patterns are a kind of adaptation. Indeed, skip patterns represent a particular multi-stage testing structure, with (blocks of) one-item routing tests followed by blocks of additional items. As long as the item set is globally unidimensional, unidmensional IRT-based scoring provides a means to compute comparable scores even in the presence of skip patterns—that is the augmentation of standard survey practice by IRT. Mixing and matching skip patterns with other CAT item selection strategies may well be novel in the developing adaptive measurement of health-related outcomes, but it may also be a widely useful idea.

Variable-criterion variable-length stopping rules. The CAT presented at the website www.AmIHealthy.com also uses a very interesting set of stopping rules for the variable-length CAT: The test is short, with relatively low precision measurement, for those who will obtain scores on the healthy end of the continua measured, and the test is longer, with more precise measurement, for those whose scores may indicate some degree of pathology. Bjorner, Kosinski,

and Ware (in press) describe this idea; to my knowledge it is new. It is true that many CAT systems in the educational achievement domain, especially for licensure tests, have stopping rules that are optimized to give most precision for persons near pass-fail decision points on the scale. But the idea of a planned smooth increase in precision from respondents who are healthy (and about which there is little concern) to those who are less healthy is another new contribution to the CAT toolbox.

This idea has led Dave Weiss (personal communication, May 24, 2004) to speculate that for some health-related outcomes the very dimensionality of the construct may be different on the healthy and less-healthy "ends" of the continuum. That is, individual differences among relatively healthy individuals may be relatively unidimensional, but as health problems become more severe they may do so in several distinct "directions" and thus become multidimensional. Weiss made the analogy with vocabulary, which is relatively unidimensional near the lower end but becomes more multidimensional as individuals develop specialized vocabularies. This is an extremely interesting suggestion, that might be studied using multi-group item factor analysis to compare low- and high-scoring populations. Indeed, this is a specific hypothesis about DIF based on diagnostic groups.

Item bank maintenance, and item parameter drift. The facts that item banks require maintenance (adding and deleting items), and item parameters may drift over time, have long been recognized in the application of IRT to educational measurement. Indeed, the original name of earlier versions of the computer software now known as BILOG-MG (du Toit, 2003) was "BIMAIN" (Muraki, Mislevy, and Bock, 1991), in which the two components of the name referred to items with binary responses, and item pool maintenance in the context of item parameter drift (Bock, Muraki, and Pfeiffenberger, 1988). However, in practice in educational

measurement, items are largely replaced, after they age or are publicly released. In the measurement of health outcomes, it is likely that the same items may be used for decades. If that happens, it will be necessary to periodically re-calibrate item banks to investigate and correct for the possibility that any number of social, cultural, or linguistic changes might have induced changes in the relations of the items with the constructs being measured.

In his review of an earlier version of this essay, Bryce Reeve (personal communication, May 24, 2004) wrote "In the future, we may be faced with the existence of multiple item banks that are created by different developers. If the multiple item banks are calibrated using different IRT models (e.g., one using the rating scale model, one using the generalized partial credit model, and one using the graded response model), how difficult would it be to link those item banks?" He answers his own question by saying "The optimal solution would be for the developers to release their item response data…" That is clearly the case. While IRT certainly provides straightforward mechanisms to compute scores based on responses to items calibrated using different models, the linking task remains. If, for example, a new item bank is developed to augment an older item bank, and a goal is to use items from both the older and newer item banks together in new instruments, all items in both banks must use the same reference scale. While there are a number of ways to accomplish that, certainly the best method from a statistical point of view is to jointly calibrate the "new" items with the "old" items. This is the way that scales are maintained over the long term in educational measurement, for example for the National Assessment of Educational Progress (Yamamoto and Mazzeo, 1992). Thus, to be truly useful over the long term, item banks should be accompanied by some mechanism to maintain, as a reference, the original calibration data set to serve as a "bridge" with potential new items and data.

Conclusion

It may be as fascinating to watch item banking and CAT develop in the health arena in the early 21st century as it was to be a part of the early development of CAT for educational measurement in the 1970s and 1980s. Those specializing in the measurement of health outcomes can learn, and have learned, from the experience of educational measurement. However, the measurement of health outcomes has its own unique features and will no doubt invite its own unique CAT systems, as it already has.

References

Armstrong, R.D., Jones, D.H., Koppel, N.B., & Pashley, P. (2004). Computerized Adaptive Testing With Multiple-Form Structures. *Applied Psychological Measurement*, *28*, 147-164.

Bjorner, J.B. (2004). Developing Tailored Instruments: Item Banking and Computerized Adaptive Assessment. Presentation at the conference entitled "Advances in Health Outcomes Measurement," Bethesda, Maryland, June 23-25

Bjorner, J.B., Kosinski, M., & Ware, J.E. (in press). Computerized adaptive testing and item banking. In P. Frayers & R.D. Hays (Eds.), Quality of life assessment in clinical trials (2nd Edition). New York: Oxford University Press.

Bjorner, J.B., Kosinski, M., & Ware, J.E. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT™). *Quality of Life Research*, *12*, 913-933.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories .*Psychometrika, 37*, 29-51.

Bock, R.D. (1997). The nominal categories model. In W.J. van der Linden & Ronald K. Hambleton (Eds), *Handbook of item response theory*. New York: Springer-Verlag.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Chang, C-H. (2004). Developing Tailored Instruments: Item Banking and Computerized Adaptive Assessment. Presentation at the conference entitled "Advances in Health Outcomes Measurement," Bethesda, Maryland, June 23-25

Chen, W.H. & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Choppin, B.H. (1968). An item bank using sample-free calibration. *Nature*, *219*, 870-872.

Choppin, B.H. (1976). Recent developments in item banking: A review. In D.N.M. De Gruijter & L.J.Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 233-245). New York: John Wiley & Sons.

du Toit, M. (Ed.) (2003). *IRT from SSI.* Lincolnwood, IL: Scientific Software International.

Hancock, T.D. (1999). Differential trait and symptom functioning of MMPI-2 items in substance abusers and the restandardization sample: An item response theory approach. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Luecht, R. M., DeChamplain, A., Nungester, R. J. (1997). Maintaining content validity in computerized adaptive testing. *Advanced in Health Sciences Education*, 3, 29-41.

Luecht, R. M. & Nungester, R. J. (2000). Computer-adaptive sequential testing. In C. Glas & W. J. van der Linden (Eds). *Computer-Adaptive Testing*, pp. 117-128. Dordrecht, The Netherlands: Kluwer Academic Publishers.

McLeod, L.D., Swygert, K., & Thissen, D (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (Pp. 189-216). Mahwah, NJ: Lawrence Erlbaum Associates.

Mead, A.D. and Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, *114*, 449–58.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E. (1997). A generalized partial credit model. In W.J. van der Linden & Ronald K. Hambleton (Eds), *Handbook of item response theory*. New York: Springer-Verlag.

Muraki, E., Mislevy, R.J., & Bock, R.D. (1991). *PC-BIMAIN: Analysis of item parameter drift, differential item functioning, and variant item performance* [Computer software]. Chicago, IL: Scientific Software, Inc.

Orlando, M., Sherbourne, C.D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*, 354-359.

Ramsay, J.O. (2000). *TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data* [Computer software]. Montreal, P.Q.: McGill University, Department of Psychology.

Ramsay, J.O. (1997). A functional approach to modeling test data. In W.J. van der Linden & Ronald K. Hambleton (Eds), *Handbook of item response theory*. New York: Springer-Verlag.

Reeve, B.B. (2000). Item- and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17, *34*, Part 2.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & Ronald K. Hambleton (Eds), *Handbook of item response theory*. New York: Springer-Verlag.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66*, 341-349.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology* ,*81* , 332-342.

Steinberg, L., & Thissen, D. (1996). *The empirical consequences of response-category recombination, as viewed with item response theory*. Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada, June 27-30.

Steinberg, L., Thissen, D. & Wainer, H. (1990). Validity. In H. Wainer, N. Dorans, R. Flaugher, B. Green, R. Mislevy, L. Steinberg & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187-231). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stout, William F. (1990), "A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation", *Psychometrika*, *55*, 293-325

Swygert, K., McLeod, L.D., & Thissen, D (2001). Factor analysis for items scored in more than two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (Pp. 217-250). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. & Wainer, H. (1993) Detection of *differential item functioning* using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, 67-113.

Wood, R. (1976). Trait measurement and item banks. In D.N.M. De Gruijter & L.J.Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 247-263). New York: John Wiley & Sons.

Wood, R. & Skurnik, L.S. (1969). Item banking: A method for producing school-based examinations and nationally comparable grades. Slough: National Foundation for Educational Research.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics*, *17*, 155-173.

---

[1] Thanks to Jakob Bjorner and Chih-Hung Chang for their support in the development of this session, and to Bryce Reeve and Dave Weiss for extremely useful comments on an earlier draft. Any errors that remain are, of course, my own.