DEVELOPING TAILORED INSTRUMENTS: ITEM BANKING AND COMPUTERIZED ADAPTIVE ASSESSMENT

Chih-Hung Chang, Ph.D.

Northwestern University Feinberg School of Medicine

Introduction

Better health outcomes measurement and management demands high quality assessment tools in order that health care providers can evaluate directly from patients' perspectives the value of a given treatment as well as the efficacy of specific pharmaceuticals and medical devices. The unmet need is a practical, time-sensitive, and user-friendly data acquisition and information delivery system that can capture and present clinically-meaningful, responsive data about patient health status in real time. Also needed is a psychometrically robust and clinically meaningful indicator of patient-reported outcomes that is interpretable by physicians and patients. Several challenges remain.

First, most health outcomes data acquisitions still rely on paper-and-pencil (P&P) format, which is time consuming, labor intensive, and limits their optimal usefulness in real-time clinical decision making. Secondly, physicians want clinically relevant and responsive questions that can be aggregated to derive a single index and scores available for use during the medical encounter. This requires a comprehensive bank of generic and disease-specific questions shown to be relevant and sensitive to the concerns of patients and it also requires appropriate scoring methods to summarize multidimensional concerns into a global summary index to guide alternative treatments. Lastly, using brief and yet comprehensive forms to obtain adequate information and desired measurement precision is often preferred. This is especially important for patients with limited functionality. Lengthy fixed-format questionnaires, which might result in content that is

redundant and unnecessary questions, often frustrate patients because of the time needed for completion. For routine health outcomes assessment to be implemented in the clinic setting, then, the assessment tools must be brief and easy for the patient to complete, impose little or no burden on clinic staff to collect and analyze, and provide clinically relevant information at the time of the clinical encounter.

The intelligent selection of questions based on prior knowledge about the patient, such as computerized adaptive testing (CAT), is the key to make health outcomes assessment efficient in clinical settings. A good CAT system utilizing item response theory (IRT) would in a sense simulate a clinical interview by focusing in on what's relevant and ignoring what is not. A successful CAT requires: 1) a pre-calibrated item bank using proper IRT models; 2) a procedure for initial item selection; 3) a scoring method; 4) an item selection method during test administration; 5) terminating rules to stop; and 6) a reliable computerized delivery system. Figure 2 depicts a CAT algorithm.

What is an item bank and how will it be useful?

More than just a collection of items or questions, an *item bank* is comprised of IRT-calibrated items that develop, define and quantify a common theme and thus provide an operational definition of a trait.[1, 2] The items in the bank are concrete manifestations of positions along the continuum that represent differing amounts of that trait. A bank is as good as its coverage of the entire continuum of the latent trait being measured and could contain all items that could be written to measure a trait. A good polytomous-scored item bank, often the case in health outcomes assessment, should contain items with high item information and test information along the underlying continuum being measured.

A well hierarchically constructed health outcomes item bank, i.e., a central item bank capturing multidimensional constructs and branches aimed for specific domains (see Figure 1), can provide a basis for designing the best possible set of questions (a "test" or "short form") for any particular application. An item bank with IRT-calibrated items makes it possible to compare health outcomes of patients who completed different sets of questions in the bank. Not only does this allow for tailored, "adaptive" testing, it also allows one to compare health outcomes of patients across studies which have used various tailored questionnaires and their original full versions. Patients can be administered the selection of bank items most appropriate to their level of health outcomes (e.g., quality of life). And because all items, exhibiting no differential item functioning, drawn from the bank are calibrated onto one common scale (metric), we can compare health status between diverse groups of patients, even when there are no common items administered for the different groups. Finally, a good item bank with wide ranging item location along the continuum defined by the IRT models also enables researchers or clinicians to select items to construct a wide variety of tailored instruments/surveys, depending on the target populations to be measured and the purpose of the assessment.

What is CAT?

Adaptive testing is a process of test administration in which items are selected on the basis of the examinee's responses to previously administered items,[3, 4] This process utilizes an algorithm to estimate person ability or latent trait to choose the next best item and to administer the test under test specifications such as content coverage and test length. In fact, the capacity to locate all examinees on the same continuum, even if they have not been administered any items in common, gives rise to the possibility of a test that is individually tailored to each examinee. IRT-based adaptive tests can be greatly facilitated by a computer because of the computational

requirements of the algorithm and the logistics of item and data management. CAT is a special type of computerized testing and is based on the psychometric framework of IRT that targets the difficulty or location of test questions to the ability or latent trait of the person. With CAT, each person needs to answer only a sample of items in order to obtain an accurate estimates what would have been obtained had the entire set of items been administered.

Why CAT?

CAT has been used successfully in educational, licensing and achievement testing, personality assessment, and military personnel selection. The use of CAT in health outcomes measurement is still in its infancy. Theoretical development of IRT was crucial to achieving these CAT advances.[5] The power to do this does not exist comfortably within the confines of traditional true score theory and yet is a natural outgrowth of IRT. CAT has been shown to reduce test length without loss of precision[6] and to provide better measurement with the same test length as a conventional full-length test (assuming an adequate item bank).

A CAT administration platform for health outcomes assessment has a number of advantages over conventional testing: 1) compared to traditional paper-and-pencil tests, CAT method is efficient, requiring fewer questions to arrive at an accurate estimate; 2) it allows for immediate report; 3) with its IRT underpinnings, CAT allows users who chooses to use different instruments to communicate with one another on a common metric, as defined and measured by the IRT models; 4) in CAT, the items being selected are tailored dynamically to the level of the individual; in principle, then, CAT reduces patient boredom or frustration by eliminating the problem of items with excessive floor or ceiling effects; and 5) since CAT automates test administration, scoring and recording, human clerical error can be eliminated.

A global (central) health outcomes item bank?

The success of a CAT administration platform depends on the comprehensiveness of the available questions or items and the breadth and depth of the item bank for test administration (see Figure 1 for example). A robust CAT administration program cannot function as it should be with a limited pool of items, or items of poor quality. To realize many of the measurement advantages of adaptive testing, the item bank from which items are selected must contain high-quality IRT-calibrated items for many different levels of the health outcomes continuum. In addition, IRT-derived item banks must satisfy the assumptions of the IRT models (e.g., unidimensionality) that underlies the item calibration, administration, and scoring, so that items can be better cataloged in the hierarchically structured banks (see Figure 1) while also taking multidimensional nature of the health into account. The development of such item bank is a continuous process and more time and resources should be committed to expanding the size and quality of item banks. For example, a new item can then be added to the existing bank if its position on the underlying continuum can be understood clinically or practically. This means that the contents of the bank can be expanded as needed.

Health related questions will be asked of patients presenting for many levels of care during the course of treatment. These questions must vary to some degree, but should also contain some commonality across delivery applications so that cross-diagnostic, cross-provider, and cross-departmental comparisons can be made. Successful implementation of item banking and CAT in health outcomes measurement will allow us to provide an efficient measurement platform using relatively brief measurement to facilitate medical decision-making.

How to develop an item bank?

Lots of patient-reported outcomes questionnaires regarding health status, symptoms, well-being, decision-making, and satisfaction have been developed, and their contents and response scales appear to share much in common. How to integrate these available instruments into a central item bank that is clinically meaningful and psychometrically sound require enormous efforts and collaborations.

Two approaches have been used to build initial item banks. Top-down approach starts with a large pool of items followed by a series of factor analyses to create definable clusters of items. These clusters are then subject to IRT analyses to ensure unidimensionality resulting in several sub-banks. Bottom-up approach usually starts with a small and better defined set of items (e.g., depression) and then build up to establish higher-order dimension (e.g., mental health).

Other issues

Unidimensionality vs. multi-dimensionality: An inherent problem in health outcomes assessment using CAT is the apparent multidimensional nature of the assessment tools where the items within domains are more highly correlated than items between domains. This condition leads to violation of the conditional independence assumption and results. Gibbons and Hedeker[7] derived an item-response model for binary response data exhibiting the bi-factor structure (each item has a non-zero loading on the primary dimension and a second loading on its specific sub-dimension) and developed a practical parameter estimation method. The bi-factor model has been extended to handle polytomous response data and is a promising model to deal with the complexity and multidimensionality nature of health.

How to link items? With the robustness of several IRT models to handle items with different response scales and missing data, concurrent calibration or co-calibration makes it possible to link items from different existing onto the same matrix. Balanced incomplete design

is also practically useful to administer different sets of items (each linking items) to different people and link all the items onto the same matrix, when administering all the items to everyone is strictly impossible in healthcare setting.

Item content: Most instruments use somewhat similar contents, if not exactly the same, and result in content redundancy. Content-overlapping items can be reconciled and/or removed. In order for the item banks to be useful for patients with limited literacy, the requirements to comprehend the questions need to be carefully examined. Rewording some existing items, with help from linguistic experts, might be necessary. One other critical issue has to do with copyright. Most instruments are developed for specific studies and are proprietary and their inclusions to a national item bank remain legally challenging.

Time Frame: Different time frames have been used for acute or chronic conditions, therefore, it is of important to select proper time frame in the CAT administration when using combined questionnaires.

Response categories: Although IRT models can handle mixed response categories, it is cognitively challenging for patients, particularly the elderly, to switch their frame of reference to answer different questions. Similar to the paper-and-pencil format, where items with the same response choices are clustered together, it seems reasonable to create some uniform formats for ease of administration. Whether to utilize dichotomous responses or polytomous responses merits further study.

Differential item functioning: Detection of DIF across different groups is an important exercise in order to establish valid item parameters for general and specific applications. This process will enable us to properly allocate items to distinct bank and instruct the CAT algorithm to- appropriately select items for different disease populations and ethnic groups.

Automatic creation of short forms: One added bonus after successful creation of item banks, researchers can provide specific study-related parameters and then a short form targeted to those specifications can be automatically created and formatted for use as a paper-and-pencil administration instrument.

Infrastructure of a national health outcomes item bank: Developing a national item bank requires interdisciplinary collaborations. Clinicians, information scientists, psychometricians, and researchers from other discipline need to work together to establish a central item bank with branches for CAT applications in health outcomes measurement.

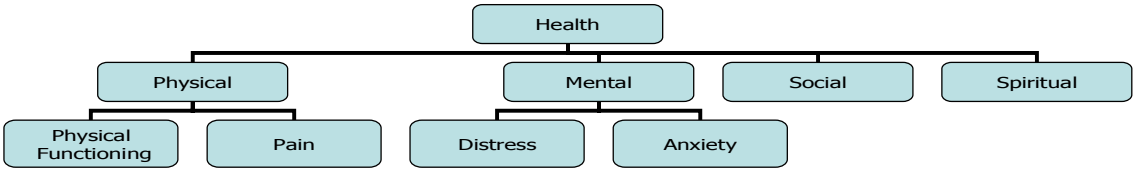

Figure 1: Hieratically-structured Central Item Bank

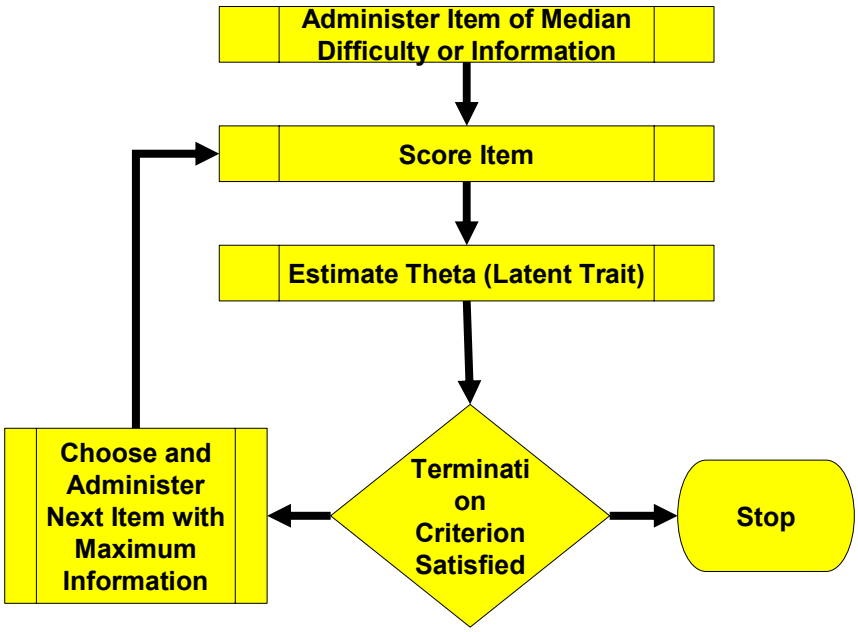

Figure 2. CAT Algorithm

References

1. Haksar L. Design and Usage of an Item Bank. *Programmed Learning & Educational Technology.* 1983;20(4):253-262.

2. Choppin B. Item Bank Using Sample-Free Calibration. *Nature.* 1968;219(5156):870-&.

3. Weiss DJ, Kingsbury GG. Application of Computerized Adaptive Testing to Educational-Problems. *Journal of Educational Measurement.* 1984;21(4):361-375.

4. Reckase MD. Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues & Practice.* 1989;8(3):11-15.

5. Sands WA, Waters BK, McBride JR, eds. *Computerized adaptive testing: From inquiry to operation*; 1997.

6. Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement.* 1984;21(4):361-375.

7. Gibbons RD, Hedeker DR. Full-Information Item Bifactor Analysis. *Psychometrika.* SEP 1992;57(3):423-436.