AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 4, 211–220

The Emergence of Item Response Theory Models and the Patient Reported Outcomes Measurement Information Systems

Steven P. Reise
University of California, Los Angeles

Abstract: Item response theory (IRT) models emerged to solve practical testing problems in large-scale cognitive achievement and aptitude assessment. Within the last decade, an explosion of IRT applications have occurred in the non-cognitive domain. In this report, I highlight the development, implementation, and results of a single project: Patient Reported Outcomes Measurement Information Systems (PROMIS). The PROMIS project reflects the state-of-the-art application of IRT in the non-cognitive domain, and has produced important advancements in patient reported outcomes measurement. However, the project also illustrates challenges that confront researchers wishing to apply IRT to non-cognitive constructs. These challenges are: a) selecting a population to set the metric for interpretation of item parameters, b) working with non-normal quasi-continuous latent traits, and c) working with narrow-bandwidth constructs that potentially have a limited pool of potential indicators. Differences between cognitive and non-cognitive measurement contexts are discussed and directions for future research suggested.

Keywords: PROMIS, Item Response Theory, Patient Reported Outcomes.

1 Introduction

Item response theory (Embretson and Reise, 2000) measurement models and associated applications (e.g., linking items to form a common item pool, the administration of computerized adaptive tests (CAT), the assessment of differential item functioning (DIF), scaling individual differences) were developed to solve practical problems in the context of large-scale multiple-choice based cognitive achievement and aptitude testing. At least in the United States, it is reasonable to argue that at present, cognitive aptitude and professional competence (nursing skill) measurement is dominated by IRT methods and procedures, as opposed to traditional classical test theory methodologies.

In recent years, IRT applications have also exploded in the non-cognitive domains of personality and psychopathology assessment, and especially in patient reported outcome (PRO) measurement (see Reise and Waller, 2009). Although the application of rigorous psychometric models in PRO measurement is most welcome, I argue that the emigration of IRT methodologies from cognitive to non-cognitive assessment raises new and interesting challenges. The central goals of this report are thus to: a) raise awareness of a large-scale project that is using IRT methods to revolutionize PRO assessment, b) provide examples of current research findings arising from this project, and based on these

findings, c) argue that non-cognitive assessment contexts present new challenges for IRT modeling.

This paper is divided into four parts. In the first Section, I review commonly stated virtues of IRT modeling in order to set a context for their application in the PRO domain. Second, rather than reviewing all recent applications of IRT in non-cognitive assessment, I focus on describing the background and research results arising from a single large-scale project called Patient Report Outcomes Measurement Information System (Cella et al., 2007, 2010). In a third Section, I review several key differences between cognitive and non-cognitive constructs and measurement contexts, and how the results of PROMIS may partially reflect these differences. Finally, I conclude by suggesting some future research directions.

2 Why IRT in Health Related Outcomes?

Regardless of whether a researcher is interested in measuring cognitive or non-cognitive constructs, the relative benefits of IRT models and associated methods have been well described in the research literature (e.g., Hays, Morales, and Reise, 2000) and in popular texts (Embretson and Reise, 2000; Hambleton and Swaminathan, 1985). Reise, Ainsworth, and Haviland (2005), for example, argue that assuming model-to-data fit and appropriate assumptions have been met, the chief virtues of IRT, relative to traditional methods are the following. IRT methods offer the ability to:

A) rigorously study how items function differently across examinee populations, that is, they facilitate the assessment of differential item functioning (DIF);

B) place individuals who have responded to different items onto a common scale;

C) derive individual scores that have good psychometric properties;

D) more thoroughly understand the psychometric properties of items and scales through inspection of item parameters and information functions;

E) create order in various research fields by having a common item pool and latent scale for key constructs, rather than many competing fixed-length instruments; and,

F) develop computerized adaptive testing (CAT) systems or static short-forms for precise and efficient assessment of individual differences.

Several of the advantages listed above are of critical importance in PRO assessment. For example, constructs such as pain, fatigue, depression, and anxiety are arguably universal phenomena that may display qualitative differences in symptom expression across countries, cultures, ages, and gender. Thus, advantages A and B – the ability to identify qualitative variation in trait manifestation but still retain the possibility of scaling individual differences on a common metric – are paramount.

A second set of critical advantages of particular relevance to PROs are E (forming an item bank) and F (administering CAT or creating short-forms). The field of PRO research is characterized by numerous competing measures of ostensibly the same or highly similar constructs. Thus, a critical problem in research is that different labs use different measures (and short-form versions), thus making results incompatible across research sites.

Furthermore, many PRO measures are not subjected to intensive psychometric scrutiny or standardization across important clinical groups. Finally, many PRO measures are exceedingly long and waste valuable research participant time.

Thus, the formation of a common item pool to measure key constructs, and to scale individuals onto the same metric using IRT-based CAT and/or short-forms produces tremendous efficiency as well as assists in making sure that research results are comparable across different research teams. In addition, by virtue of establishing a common latent trait scale, future proposed measures, as well as "gold-standard" legacy measures, can be evaluated in terms of item and scale information, relative to an existing standard. In the following section, I highlight a large research project that was specifically designed to take advantage of the benefits of IRT, namely, the Patient Reported Outcomes Measurement Information Systems (PROMIS) project (Cella et al., 2007, 2010).

3 The PROMIS Project

To my knowledge, PROMIS (`http://www.nihpromis.org/default.aspx`) is the largest application of IRT methods in the non-cognitive domain. The following two quotes succinctly summarize the project's goals:

"The NIH PROMIS Roadmap initiative is a 5-year cooperative group program of research designed to develop, validate, and standardize item banks to measure patient-reported outcomes (PROs) relevant across common medical conditions" (Cella et al., 2007, p. 3). "PROMIS will build and validate common, accessible item banks to measure key symptoms and health concepts applicable to a range of chronic conditions, enabling efficient and interpretable clinical trial and clinical practice applications" (Cella et al., 2007, p. 4).

This is not the appropriate context for providing a detailed history of PROMIS (see Cella et al., 2007, 2010), or to provide a step-by-step outline of the PROMIS psychometric methods (see Reeve et al., 2007; Pilkonis et al., 2010). Rather, in this section, I briefly review the specification of initial constructs, item pool development strategy, sampling plan, and subsequent IRT item calibrations and the formation of item banks. An up-to-date set of publications based on the PROMIS project can be found at the following link. `http://www.nihpromis.org/Web\%20Pages/Publications\%20and\%20Reports.aspx`.

A critical first step in the PROMIS project was the adoption of the World Health Organization distinction among physical, mental, and social health. This overarching framework directed the research project toward the initial specification of "core domains" such as physical functioning, emotional distress, and social functioning. In turn, these core domains served as key targets for the specification of subdomains and the development of IRT calibrated item pools.

After exhaustive literature review, an initial library of over 7000 potential self-report items was identified (Pilkonis et al., 2010). These items differed in stem, response format, reference time frame, and instrument of origin. Thus, one of the early tasks was to rewrite these items to have a standard format (five response options) and time frame (last 7 days). Expert panels were also formed who binned items together into content domains, and winnowed down items that were judged poorly written or overly redundant. The re-

duced pool of candidate trait markers contained 1100 items (DeWalt, Rothrock, Yount, and Stone, 2007). Finally, cognitive interviews were conducted on patient groups in order to identify content areas that were not covered by the current pools. These pools then served as the basis for collecting item response data for an initial IRT calibration. See `http://www.assessmentcenter.net/ac1` for a listing of final items.

The design of the PROMIS project presented two interesting challenges. First, given the large amount of items, how should the sampling be conducted? This issue was resolved by using two different types of sampling strategies. First, items were divided up into random subsets and block-sampling was used. Second, some individuals received all the items from one or more domains (e.g., all depression and anxiety items). The latter strategy was used for calibrating item parameters, while the former design was used for investigating structural and other psychometric questions.

A second interesting challenge was deciding what population(s) should be sampled and how should the metric of the item parameters be identified? In total, over 20000 individuals responded to PROMIS items. A majority of those responses were collected from a non-clinical general population sample via the internet. In addition, item response data were collected from a number of "clinical" disease populations (e.g., coronary artery disease, liver disease, arthritis or rheumatism, asthma, diabetes, depression, Parkinson's disease). Most importantly, a scale setting sample sub-sample was created reflecting demographics proportional to the 2000 U.S. census. Thus, PROMIS item parameters and scale scores are interpreted relative to a U.S. population.

In total, the above sampling strategies were used to develop and standardize 11 item pools: anger, anxiety, depression, fatigue, pain behavior, pain impact, physical function, satisfaction with participation in discretionary social activities, satisfaction with participation in social roles, sleep disturbance, and wake disturbance. All item responses were scrutinized for monotonicity, unidimensionality and local dependence violations, fit to the graded response model (Samejima, 1969), and differential item functioning between important demographic groups (age, gender, education).

After these preliminary psychometric analyses, final item banks were formed for each construct. According to Cella et al. (2010), these pools have high internal consistency (coefficient alpha > 0.95), and scale information is spread widely across the trait range sufficiently to allow precise measurement across the trait range. In addition, correlations among bank scores and linked "gold-standard" legacy measures were moderate to strong. Finally, short-forms of 10 items or less correlated with full bank scores greater than 0.90 (see also Choi, Reise, Pilkonis, Hays, and Cella, 2010). A few more specific results will be noted in the next section.

4 Cognitive Versus Non-Cognitive Measurement Contexts

It is tempting to believe that all psychological constructs are equal and that measurement is measurement (i.e., IRT methods can be used for every nameable construct). However, the modeling of individual differences must always be sensitive to the assessment context and the phenomenon the researcher is attempting to assess. As noted earlier, IRT models and methods were developed to solve practical problems in large-scale multiple-choice

achievement and aptitude testing. It is thus appropriate to ask, does the application of IRT methods beyond these domains lead to any new challenges or can standard IRT methods be applied as usual?

Reise and Waller (2009) have pointed out a number of key differences between cognitive and non-cognitive constructs and measurement contexts. Although reasonable minds may disagree with one or more points, these authors argue that cognitive and non-cognitive measurement often differs in the following ways. Specifically, relative to graded response rating scale PROs, in large-scale multiple-choice aptitude testing:

A) it is often easy to define the relevant examinee population;

B) it is often reasonable to assume a continuous normally distributed latent trait (quantitative ability) and, in turn, that test scores will be normally distributed;

C) because of the time and effort extended toward specifying age and grade appropriate learning objectives, content domains are better understood;

D) for some constructs, there exists an almost limitless item pool (e.g., reading comprehension, spelling, algebra); in turn,

E) it is relatively easy to write dichotomously scored items that span the difficulty range (i.e., easy items requiring low level skills versus hard items that require more knowledge or complex cognitive skills); and finally,

F) important ability constructs are often broad-bandwidth (e.g., verbal ability) – and thus no one item can perfectly reflect the construct – and both ends of the latent trait continuum are interpretable (i.e., it is meaningful to talk about low and high ability individuals).

The above features ostensibly represent a typical context in achievement or aptitude testing. Now I will consider how the context of PROMIS differs from the above and how that may impact observed results. In particular, I consider three challenges to IRT modeling that arise in PRO assessment. These are: a) selecting an appropriate population to set the metric for interpretation of the item parameters, b) working with non-normal quasi-continuous latent trait distributions, and c) working with narrow-bandwidth constructs that potentially have a limited pool of potential indicators.

First, as is well known, the scale of the latent variable in IRT is arbitrary and must be identified in some way. Outside the context of Rasch modeling, it is common in IRT to specify that in some calibration population, the mean of the latent trait is zero and the standard deviation is 1.0. Thus importantly, the trait level and variability of the calibration sample importantly affects the metric used to assess individual differences, and the interpretation of item parameters. For example, in a cognitive assessment context, a spelling test written for 6th graders would look "hard" if calibrated on a sample of 4th graders, and "easy" if calibrated on a sample of 8th graders. Analogously, in PRO assessment, whether a clinical or non-clinical population is selected for identifying the latent trait scale, can importantly affect the scale for the item parameters.

The PROMIS constructs are health outcomes (pain impact, sleep disturbance, fatigue, anger) that are commonly observed in some clinical populations and less so in others, in particular in non-clinical "healthy" populations. This creates an interesting challenge, namely, how should researchers identify the scale for the item parameters? This is an

important choice because it affects not only the item parameters, but also determines how scale scores and change scores will be interpreted. As noted above, PROMIS investigators used a sample that represented the 2000 US population to set the metric for all item banks. The effect of this choice is that the population latent trait distribution is positively skewed (poor health is the high end), and latent trait scale scores for some clinical groups (who score higher on a trait compared to normals) may be compressed.

Another possible option would have been to set the scale for the item parameters on a clinical population (e.g., depression items where the metric of the latent trait is identified in a sample of in-patients or out-patients with any psychiatric diagnosis). The effect of selecting this "clinical" sample identification strategy would be to spread scores in the clinical group, and compress latent trait scores in the general population. The point here is not to argue for a particular approach, in fact the PROMIS approach is consistent with the original project mandate. However, I raise this issue to demonstrate a key challenge for IRT modeling that exists in PRO assessment, which seldom is problematic in cognitive abilities testing.

Another set of challenges arises from the nature of PRO constructs. Specifically, it is almost never safe to assume that constructs such as pain behavior, physical function, fatigue, or depression are normally distributed in a general non-clinical population. The degree to which non-normality of the latent distribution is consequential for IRT modeling is debatable and needs further research. However, note that it is common to assume a normal distribution for the latent trait during the estimation of the item parameters using marginal maximum likelihood estimation. To the degree that violating the normality assumption biases or otherwise distorts the item slope and threshold parameter estimates, applications of IRT (e.g., DIF assessment, linking items from different scales) based on those parameters are problematic.

Moreover, many PRO constructs are unipolar (quasi-continuous and defined on only one end). For example, the low end of anger is not love, it is the absence of anger, and the low end of pain impact, it is the absence of pain impact. This may appear trivial at first glance, but it is arguable that the uni-polar (defined only on one end) quasi-continuous nature of some PRO constructs makes it challenging to write items that span the range of the latent trait. It is challenging because the low end of the trait has little meaning (Reise and Waller, 2009). There is some evidence of this in PROMIS.

Consider the depression, anger, and anxiety scales item parameter reported by Pilkonis et al. (2010). As mentioned above, these banks were created through a comprehensive identification of items relevant to depression and make use of five-point response formats and a 7-day time frame. Ostensibly, the use of multi-point response options allows the items to spread threshold parameters (and thus item information) across the trait range. But this is not what was found; In the depression, anger, and anxiety pools item threshold parameters (four per item because there are five response categories) are highly clustered on the high (depression) end of the scale.

For example, in the anxiety, anger or depression item banks there is not a single first threshold parameter estimate (graded response model) below -1.0 and around half of those thresholds are positive. This implies that people have to be above the mean on the latent trait in order to respond in category two and above ("rarely" and above), rather than in category one ("never"). In other words, despite efforts to allow discriminations across

the trait range by using a five-point rating scale, most response options are "hard" and require above average trait standing to endorse. As a consequence, scale information is peaked at the high end.

Perhaps such findings are expected for serious clinical constructs like depression, anger, or anxiety, due to the fact that trait indicators consists of relatively rare symptoms. Nevertheless, it is arguable that, given the time and effort in developing item content, and given the five-point response format designed to spread information out across the range, if the PROMIS project could not identify items that differentiate among low trait individuals, perhaps such items simply cannot be written for some PRO constructs. Note this is not to say that low trait individuals are not measured well with the PROMIS item banks. In fact, because items tend to be highly discriminating, item information is high across the trait range despite the fact that threshold parameters tend to be clustered in the depressed trait range (see Cella et al., 2010, Table 2).

Finally, relative to constructs such as verbal and quantitative ability, some PRO constructs are conceptually narrow and have a limited pool of potential indicators. For example, there are only a very limited number of non-redundant indicators of constructs like fatigue and sleep disturbance. In some respects, these features facilitate IRT modeling. For example, such measures tend to be highly unidimensional and items highly discriminating; the ratio of the 1st to 2nd eigenvalue in the pain impact item bank is an astonishing 35 to 1; the Mokken scalability coefficient for the fatigue item bank is an amazing 0.71 (Lai et al., 2010); and the depression item bank has four items with graded response model slope parameters greater than 4.0 (logistic metric), implying that there is little conceptual distance between the symptom and latent construct – these items are redundant with the latent variable.

Although such findings suggest near perfect Guttman like relations between symptoms (items) and latent trait, such "super homogeneity" in measures is not always desirable. For example, the well known attenuation paradox implies that too narrow a measure may not predict heterogeneous criterion well. More relevant to IRT, such homogeneity may defeat the purpose of applications like item banking and CAT. Recall that in aptitude testing, the reason that large pools of IRT calibrated items are built is because no one item perfectly represents the latent trait and thus researchers need to ask several items to hone in on an individual's trait standing. When item sets become too homogeneous, too discriminating (remember item information is related to the square of item slope in the graded response model) or too "scalable" in a Mokken sense, the need for item banks, CAT, or even short-forms with more than one item becomes questionable. On the other hand, research by Choi et al. (2010) and Lai et al. (2010) has shown that for several PROMIS item banks (depression and fatigue, respectively) administering multiple items is indeed necessary for precise measurement across the trait range, and moreover, CAT does indeed outperform fixed-length short-forms.

5 Conclusion

By using IRT modeling to create a set of high quality psychometrically sound item banks to assess important constructs, PROMIS has revolutionized PRO assessment. Beyond

demonstrating that IRT models can be applied productively outside the domain of cognitive testing, the advantages of the PROMIS initiative include: a) increasing the comparability of research results across research teams, b) improving the psychometric quality and efficiency of PRO measurement, and c) creating a set of standardized metrics for PRO domains by which the effectiveness of interventions can be judged.

Nevertheless, despite these successes, I argue that the application of IRT models and associated methods in the PRO domain presents interesting challenges. Beyond the tricky problem of selecting an appropriate calibration sample to identify the item parameters, these challenges include: a) the estimation of item parameters under non-normal latent trait conditions, b) identifying items that provide information across the trait range in the presence of quasi-continuous traits, and c) writing non-redundant items in the presence of narrow-band constructs that have a limited pool of potential trait indicators.

I note that Woods (2006) has been working on alternative item parameter techniques that simultaneously estimate a non-normal latent trait distribution and the item parameters. As for dealing with quasi-continuous highly skewed latent traits where the low end (non-disease) of the trait scale is difficult to interpret, future "hybrid" models that combine latent class (depressed versus not depressed) with latent trait (if depressed, how much?) models may prove valuable. Finally, one solution to the narrow bandwidth problem is to design an assessment system around a bifactor framework, where each item serves as both a measure of a general global disposition (generalized distress), as well as one or more specific narrow bandwidth constructs (anxiety, depression, anger). An example of such a project can be found in Gibbons, Grochocinski, Weiss, Bhaumik, and Kupfer (2008).

There is no doubt that the future will bring better technical methods of assessing the effects of IRT model violations, more accurate and diagnostic model fit indices, improved parameter estimation techniques, new models that handle multidimensionality, and new DIF detection strategies that allow the consideration of multiple-populations. Also of equal importance are advancements in how latent variable modeling should be adapted to fit in this new PRO domain. Given the differences in assessment contexts described above, a researcher cannot blindly take IRT modeling technologies and transport them to a new domain. At the least, serious consideration of the meaning of the latent trait (Borsboom, 2005) in PRO measures and the meaningfulness of an identified latent trait scale is required to: a) advance our understanding of the response processes underlying item responses, b) assess the meaningfulness of change scores, and c) ultimately "validating" the application of IRT models.

Acknowledgements

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Northwestern University, PI: David Cella, PhD, U02AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR-52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford

University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, PhD, Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD. See the web site at `http://www.nihpromis.org` for additional information on the PROMIS cooperative group.

References

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., and Yount, S. (2010). *Initial item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS) network: 2005-2008.* (under review)

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., and Reeve, B. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*, S3-S11.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., and Cella, D. (2010). *Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms.* (in press)

DeWalt, D. A., Rothrock, N., Yount, S., and Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*, S12-S21.

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

Gibbons, R. D., Grochocinski, V. J., Weiss, D. J., Bhaumik, D. K., and Kupfer, D. J. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychological Services*, *59*, 361-368.

Hambleton, R. K., and Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Hays, R. D., Morales, L. S., and Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*, 517-527.

Lai, J., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2010). *Development of the PROMIS fatigue item bank, computerized adaptive testing, and short-forms.* (under review)

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, T., and Cella, D. (2010). *Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger.* (under review)

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., and Teresi, J. A. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*, S22-S31.

Reise, S. P., Ainsworth, A. T., and Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*, 95-101.

Reise, S. P., and Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.

Woods, C. M. (2006). Ramsay-curve item response theory to detect and correct for non-normal latent variables. *Psychological Methods*, *11*, 253-270.

Author's address:

Steven P. Reise
Department of Psychology
Franz Hall, UCLA
Los Angeles, CA 90095
email: reise@psych.ucla.edu