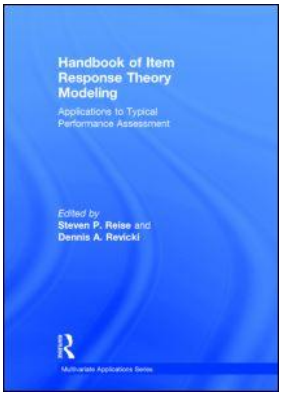

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Introduction

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch1
Steven P. Reise, Dennis A. Revicki
Published online on: 16 Dec 2014

Part I

Fundamental Issues in Item Response Theory

This page intentionally left blank

1 Introduction

Age-Old Problems and Modern Solutions

Steven P. Reise and Dennis A. Revicki

The statistical foundation of item response theory (IRT) is often traced back to the seminal work of Lord, Novick, and Birnbaum (1968). The subsequent development, research, and application of IRT models and related methods link directly to the need of large-scale testing companies, such as the Educational Testing Service, to solve statistical as well as practical problems in educational assessment (i.e., the measurement of aptitude, achievement, and ability constructs). Daunting problems in this include the challenge of administering different test items to demographically diverse individuals across multiple years, while maintaining scores that are comparable on the same scale. This test score comparability problem traditionally has been addressed with "test-score equating" methods, but now more routinely, IRT-based "linking" strategies are used (see Chapter 19).

The application of IRT models and methods in educational assessment is now commonplace (e.g., see most any recent issue of the *Journal of Educational Measurement*), especially for large-scale testing firms that employ on their research staff dozens of world-class psychometricians, content experts, and item writers. The application of IRT models, and related statistical methods in the fields of personality, psychopathology, patient-reported outcomes (PRO), and health-related quality-of-life (HRQOL) measurement, in contrast, has only recently begun to proliferate in research journals. In these noneducational or "typical performance" domains, the application of IRT has gained popularity for much the same reasons as in large-scale educational assessment; that is, to solve practical and technical problems in measurement.

The National Institutes of Health (NIH) Patient Reported Outcome Measurement Information System (PROMIS®), for example, has developed multiple item banks for measuring various physical, mental, and social health domains (Cella et al., 2007; Cella et al., 2010). Similarly, the Quality of Life in Neurological Disorders (www.neuroqol.org) and NIH Toolbox (www.nihtoolbox.org) have also employed IRT methods of scale development and item analysis. One of the chief motivations underlying the application of IRT methods in these projects was to solve a long-standing and well-recognized problem in health outcomes research; namely, that for any important construct, there are typically half a dozen or so competing measures of unknown quality and questionable validity. This chaotic measurement situation, with dozens of researchers studying the same phenomena using different measurement tools, fails to promote good research and inhibits the cumulative aggregation of research results.

Large-scale IRT application projects, such as PROMIS®, have raised awareness not only of the technical and practical challenges of applying IRT models to psychological or PRO data, in general, but also has uncovered the many and varied special problems and concerns that arise in applying IRT outside of educational assessment (see also Reise & Waller, 2009). We will highlight several of these critical challenges later in this chapter to set a context for the present volume. Before doing so, however, we note that thus far,

standard IRT models and methods have been imported into noneducational measurement contexts, and essentially without modification. In other words, there has been little in the way of "new models" or "new statistical methods" uniquely appropriate for PRO or any other type of noneducational data (but see Chapter 13).

This equalitarian—the same IRT models and methods should be used for all constructs, educational or PRO—was perhaps critical in early stages of IRT exploration and application in new domains. Inevitably, we believe, further progress will require new IRT-based psychometric approaches particularly tailored to meet measurement challenges in noneducational assessment. We will expand on this in the final chapter. For now, prior to previewing the chapters in this edited volume, in the following section, we briefly discuss some critical differences between educational and noneducational constructs, data, and assessment contexts, as these relate to the application of IRT models. We argue that although there are fundamental technical issues in applying IRT to any domain (e.g., dimensionality issues, assessing model to data fit), unique challenges arise when applying IRT to noneducational data due to the nature of the constructs (e.g., limited conceptual breadth, questionable applicability across the entire population), and item response data (e.g., non-normal latent trait distribution issues).

Educational Versus Noneducational Measurement

It is well recognized that psychological constructs, both cognitive and noncognitive, can be conceptualized as being hierarchically arranged, from very general to middle level, conceptually narrow to specific behaviors (Clark & Watson, 1995).[1] Since Loevinger (1957), it has also been well recognized (although not necessarily realized in practice by scale developers) that the position of a construct in this hierarchy has profound implications for all aspects of scale development, psychometric analyses, and ultimately validation of test score inferences.

Almost by definition, measures of broad bandwidth constructs (intelligence, verbal ability, negative affectivity, general distress, overall life satisfaction, or QOL) must have heterogeneous item content to capture the diversity of trait manifestations.[2] In turn, item intercorrelations, item-test correlations, and factor-loadings/IRT slopes are expected to be modest in magnitude, with low communality. Moreover, resulting factor structures may (must?) be multidimensional to some degree, perhaps with a strong general factor and several so-called group or specific factors corresponding to more content-homogeneous domains (see Chapter 2).

On the other hand, just the opposite psychometric properties would be expected for measures of conceptually narrow constructs (mathematics self-efficacy, primary narcissism, fatigue, pain interference, germ phobia). That is, in this latter context, the content diversity of trait manifestation is very limited (by definition of the construct), and as a consequence, item content is homogeneous with the conceptual distance between the item content and the

1 Interestingly, these authors attribute the apparently inexhaustible proliferation of individual difference constructs and measures to this hierarchical structure, which can be cleaved in an infinite number of ways.

2 We are assuming here that for the construct of interest, there is a latent variable underlying, or causing, variation in item response. Such a measurement model has been termed an "effects" indicator model by Bollen and Lennox (1991). If the construct of interest were better represented by a "cause" indicator measurement model, then IRT models, which assume an underlying latent trait, are questionable. Moreover, in a cause indicator model, item content diversity would be required to form a census of indicators (see Bollen & Lennox, 1991 for further discussion).

latent trait being slim. In turn, this can result in very high item intercorrelations, item-test correlations, and factor-loadings/IRT slopes. In factor analyses, essential unidimensionality would be the expectation, as would high item communalities. Finally, in contrast to broadband measures, where local independence violations are typically caused by clusters of content-similar items, in narrowband measures, local independence violations are typically caused by having the same item content repeated over and over with slight variation (e.g., "I *have problems concentrating,*" "I *find it hard to concentrate,*" "I *lose my concentration while driving,*" "*It is sometimes hard for me to concentrate at work*").

In our judgment, applications of IRT in educational measurement have tended toward the more broadband constructs, such as verbal and quantitative aptitude, or comprehensive licensure testing contexts (which also involve competencies across a heterogeneous skill domain). In contrast, we argue that with few exceptions, applications of IRT in noneducational measurement have primarily been with constructs that are relatively conceptually narrow. As a consequence, IRT applications in noneducational measurement contexts present some unique challenges, and the results of such applications can be markedly different from a typical IRT application in education.

For illustration, Embretson and Reise (in preparation) report on an analysis of the PROMIS® anger item set (see Pilkonis et al., 2010), a set of 29 items rated on a 1 to 5 response scale. Anger is arguably conceptually narrow because there simply are not that many ways of being angry (especially when rated within the past seven days); that is, the potential pool of item content is very limited, unlike a construct, say, such as spelling or reading comprehension where the pool of items is virtually inexhaustible. Accordingly, alpha was 0.96, and an eigenvalue ratio of around 15 to 1, suggesting unidimensionality, or at least a strong common factor. Fitting a unidimensional confirmatory factor analysis resulted in an "acceptable" fit by conventional standards. However, univariate and multivariate Lagrange tests indicated 407 and 157 correlated residuals needed to be estimated (set free), respectively. This unambiguous evidence against the data meeting the unidimensionality/local independence assumption was not due to the anger data being in any real sense of the term "multidimensional," with substantively interpretable distinct factors, but rather as having many sizeable correlated residuals (violations of local independence), likely caused by highly similar item content.

In sum, item responses to conceptually narrow measures such as anger are clearly highly influenced by a single common dimension (what else could items like "I *stayed angry for hours*" be measuring?), but are not, statistically speaking, truly unidimensional/locally independent, as commonly applied unidimensional IRT models assume. Importantly, item slope parameters (and, thus, test information) may be artificially high because of these unmodeled local independence violations. On the other hand, responses to measures such as anger cannot be readily fit to multidimensional models such as a correlated-factors, second-order, testlet, two-tier, or bifactor measure, because items do not cluster neatly into content domains. One either has to decide that the measure is sufficiently unidimensional such that the item parameters are estimated accurately, or start deleting items displaying local dependencies, with the realization that attempting to eliminate all local independence violations may result in a three-to-five-item bank.

As illustrated earlier, differences between educational and noneducational constructs, in particular their level of conceptual breadth, can be consequential for IRT analyses, in particular dimensionality assessment. However, it is by no means the only consequential difference. We argue that for many educational constructs where IRT models are applied, it is reasonable to assume a continuous, normally distributed latent variable in the population of relevant examinees. Often, this population is readily defined (e.g., all public school 8th graders in California), and data are collected on almost the entire population (sans absentees).

Moreover, for the test developer, it is relatively straightforward to generate multiple items that extend across the trait continuum from "easy" items requiring low-level skills to "hard" items requiring greater knowledge or skills (i.e., have item location parameters that span the latent trait range).[3] As a consequence, test information will be spread out across the latent trait range, and meaningful interpretations of latent trait scores can be made at either end of the latent trait continuum from low to high ability levels.

In the measurement of health outcomes or psychopathology, we argue that the measurement situation and the item response data often differ greatly from that discussed earlier. For example, often for constructs, such as pain interference, pain behavior, fatigue, or depression, scores are not normally distributed in the nonclinical, general population. For example, the distribution of PROMIS® pain behavior and pain interference scores in the general population are highly skewed. Based on completed IRT analyses on pain behavior items, including a sample of individuals with varying levels of chronic pain, the results supported two distributions, one of no pain and pain, and then if pain, a more normal distribution of pain behavior scores (Revicki et al., 2009). Analysts often assume a normal distribution for the latent trait during the estimation of item parameters and model fit (see Chapter 4). Violating the assumption of normal distributions for the latent trait may bias the estimates of IRT slope and threshold parameters, although the extent of this bias attributable to various levels of non-normality needs further research. Extreme cases of highly skewed and non-normal distributions may require alternative IRT modeling approaches. In the case of the PROMIS® pain behavior item bank, a hybrid nominal-partial credit IRT analysis provided very good model fit to these data (Revicki et al., 2009).

Related to this non-normal latent trait is a similar problem with a slightly different origin. Consider again the anger example described previously. The low end is not mild cynicism, negativity, irritability, being "upset" or "frustrated," but more likely the complete absence of anger reactions (within the past seven days). The construct of anger might be what Lucke (see Chapter 13) refers to as a unipolar trait—definable only on one end of the scale—and Reise and Waller have been referring to as a quasi-trait since at least 1990 (but see also Reise & Waller, 2009). The concept of a unipolar or quasi-trait is even better illustrated by constructs such as depression, sex addiction, belching/flatulence symptoms, and tobacco use. For these types of constructs, low scores are not necessarily below average on the trait, or low on the trait, but rather the trait is simply not applicable to them.

There are three obvious consequences of applying IRT models to unipolar health outcome or psychopathology traits. First, if the measure is given to a "healthy" sample, there will be many zero item scores, and total scores will be highly skewed. It is not at all clear whether even item parameter estimation strategies that can account for non-normal latent trait distributions (see Chapter 4) can salvage viable item parameters in this circumstance. Second, if the full range of the latent variable is not meaningful, then it should be difficult if not impossible to write items with location parameters that span the range of the latent variable. Instead, one would expect that they would be highly skewed. Indeed, this is exactly what is found for anger (see Pilkonis et al., 2010), where none of the first threshold parameter estimates, based on the graded response model, is below negative one and almost half of the parameters are positive.

Reise and Waller (2009) argued that application of a polytomous IRT model to noneducational measures has seldom, if ever, resulted in item location parameters spread across

3 We are by no means implying that creating a test or item bank that spans the complete ability range is an easy task. We merely are pointing out that a range of ability from low to high is more readily definable in educational assessment.

the latent trait range (see also Embretson & Reise, 2000 for a similar argument regarding a popular self-esteem scale). In fact, studies routinely report the opposite—that locations are clustered tightly at one end of the scale and that test information is highly peaked. At the very least, such findings have clear implications for the development of banks of items that provide items with differential precision across the trait range (see Chapter 16) and for the viability of computerized adaptive testing (see Reise & Henson, 2000 for further commentary). However, with those cautions raised, research has demonstrated that for the PROMIS® depression and fatigue item banks, multiple items are needed for precise assessment across the trait continuum and that CAT scores outperform static short forms (Choi et al., 2010; Lai et al., 2011).

A third consequence of applying IRT to unipolar traits, especially for clinically related patient outcomes (e.g., pain behaviors, depression), is that it is often unclear who the norming population should be. Decisions about the population for identifying the latent trait scale can affect the scale and item parameters. For example, if a clinical population is selected for item calibration and setting the scale for a depression item bank, the metric of the latent trait can be identified based on a sample of patients with depressive disorders. Selecting a clinical sample would result in an extended range of scores for the clinical depressed sample, and compression of the depression latent trait scores in the general, nonclinical population. If a general population sample is selected for item calibration and setting the metric, the opposite would occur, that is, there would be a greater spread of scores in the general population with compression of scores for the depressed clinical population. There is no right or wrong way to set the metric for the latent trait; however, decisions about the calibration population have implications for the scale metric.

Finally, and related to this issue of unipolar traits, there are a number of cases where health outcome measures are configured as presence-severity items (Liu & Verkuilen, 2013). These measures first ask the respondent to indicate whether an event or symptom is present, and if affirmative ask for a rating of severity, frequency, bother, distress or effect. For these kinds of latent constructs and item configuration, alternative models may need to be considered for the IRT analyses, such as the nominal response model (Liu & Verkuilen, 2013; Chapter 18), or new IRT models such as Lucke's unipolar models (see Chapter 13) may be required.

Brief Preview of Chapters

This summary of the unique challenges of applying IRT to noneducational data provides a context for the present volume, given that many of the examples in the various chapters illustrate IRT methods using health outcome data. To further contextualize the following chapters, we provide a very brief description of the motivation and some of the central themes of each.

Part I: *Fundamental Issues in Item Response Theory*

This part includes a set of seven chapters that tackle foundational issues relevant to the application of IRT models in any substantive domain, educational and noneducational. Because unidimensional IRT models are, by far, the most commonly applied, in Chapter 2 Steven Reise, Karon Cook, and Tyler Moore review the definition of (uni) dimensionality—as something belonging to data and not a construct—and describe how the concept of unidimensionality has been traditionally assessed using a variety of statistical indices. One unique feature of their chapter is the argument that psychological data are never strictly unidimensional, and, thus, the critical question in applied research is

determining whether the multidimensionality inherent in the data is sufficient to bias item parameter estimates. In turn, they suggest a "comparison modeling" approach where the parameters from a bifactor model are evaluated relative to those estimated under a unidimensional model. This chapter sets the foundation and provides complementary material for Chapters 9 and 11, in which alternative approaches to scaling individuals on a single dimension in the presence of multidimensional data are detailed.

After establishing that the data are appropriate for application of an IRT model, perhaps the most fundamental topic in all of IRT is item parameter estimation. In Chapter 3, Li Cai and David Thissen provide a comprehensive review and discussion of modern full-information (based on the complete item response matrix) approaches to unidimensional item parameter estimation, including explanation of the Metropolis-Hastings sampler and Robbins-Monroe method. Although parameter estimation approaches in unidimensional IRT models have been around for a long time, these traditional methods are not entirely adequate to handle the computational challenges presented by new types of multidimensional IRT models, such as the bifactor or two-tier models (see Chapters 8 to 12). The methods presented in Chapter 3 have straightforward extensions to polytomous item responses and multidimensional models.

Traditional marginal maximum likelihood approaches to IRT item parameter estimation typically assume a normal prior distribution (implemented through quadrature points and weights) for the latent trait, and then item parameters are estimated assuming this distribution is reasonable. It has long been a concern that this normality assumption may not be appropriate in many noneducational measurement contexts, especially PRO measurement. Accordingly, in Chapter 4, Carol Woods describes the problems in item parameter estimation when the latent trait is non-normal and then reviews statistical methods for estimating the latent trait distribution simultaneously with the item parameters in the context of unidimensional item response theory models. It is also shown that item parameters and estimated latent trait scores are more accurate when the shape of the latent trait distribution is estimated, rather than assumed normal.

Chapter 5, authored by Rob Meijer, Jorge Tendeiro, and Rob Wanders, could have been placed as Chapter 3 because it deals with the use of nonparametric IRT methods (NIRT) to explore whether item response data are consistent with the assumptions underlying the fitting of a parametric IRT model. Nevertheless, we decided to group Chapters 5, 6, and 7 because they all are relevant to the topic of model fit, albeit approached from different vantage points. Specifically, Chapter 5 provides insight into a number of commonly used NIRT methods and demonstrates how these methods can be used to describe and explore the psychometric quality of PRO measures. The authors also emphasize consideration of the degree to which specific IRT models are "robust" to violations of assumptions and provide practical advice for applied researchers.

Chapter 6, by Alberto Maydeu-Olivares, introduces cutting-edge methods for evaluating overall IRT model fit based on analyses of the contingency table. As is well known, for any sizeable length test, the complete item response contingency table will be sparse because of many item response patterns not being observed. In turn, this has made it nearly impossible to use the discrepancy between the observed response patterns and those predicted from the estimated model to judge fit. Maydeu-Olivares reviews traditional methods for assessing the overall model fit and describes new limited information on overall goodness of fit statistics and methods for assessing approximate fit and piecewise assessment of fit. Complementing Chapter 6, Pere Ferrando, who has published extensively on the topic of person fit in noneducational settings, presents a comprehensive review of statistical methods for evaluating how consistent an individual's item response pattern is with an estimated IRT model in Chapter 7. He summarizes the importance of evaluating

person fit, details the main methods for assessing person fit, and describes methods for diagnosing the causes and implications of poor person fit.

Part II: *Classic and Emerging IRT Modeling Approaches*

In this part, a number of different, cutting-edge methods for IRT modeling are summarized. Chapters 8 through 11 all deal with the emerging umbrella topic of multidimensional IRT, but in very different ways. Michael Edwards, R. J. Wirth, Carrie Houts, and Andrew Bodine, in Chapter 8, explore the concepts underlying dimensionality and present some of the challenges researchers face when trying to choose between different models. These conceptual issues are illustrated with both simulated and real data examples before turning to a broader discussion of how the issue of dimensionality may affect PROs. In Chapter 9, Brian Stucky and Maria Edelen summarize the structure of traditional multidimensional models with an emphasis on the bifactor and more recent generalizations such as the two-tier models (Chapter 10). They then describe complications that arise in interpreting multidimensional IRT item parameters and propose a method for creating unidimensional scales from multidimensional item response data using the results from a bifactor model.

Although briefly described in the preceding chapters, in Chapter 10, a two-tiered item factor analysis (or IRT) model is outlined in more detail by Wes Bonifay. A two-tier model is an IRT model with more than one general factor (which may be correlated) and multiple primary factors nested within each general factor. As such, the two-tier model is a parent model that subsumes the correlated factors, bifactor, and testlet response IRT models. Data analyses are used to demonstrate the psychometric advantages of the two-tier item factor analysis model.

The final chapter on multidimensional IRT, Chapter 11, is authored by Edward Ip and Shyh-Huei Chen. They do not propose a new multidimensional IRT model, per se, but rather a method of scaling individuals on a single dimension in the presence of multidimensionality. Specifically, this chapter details projective IRT models, which are a class of statistical methods for collapsing a multidimensional latent space down into a unidimensional latent space that reflects the common dimension assessed by all the items. Ip and Chen provide both Monte Carlo simulation results and several real-data applications to illustrate the method. They also provide a comparison of the projection methodology with the results from a bifactor model.

Chapters 12 and 13 each present "new" approaches to IRT modeling that do not fit easily into old rubrics such as "multidimensional IRT." First, explanatory IRT (EIRT) modeling is a relatively new but emerging approach to IRT modeling that heretofore has captured the interests of many educational researchers. In Chapter 12, Paul De Boeck and Mark Wilson describe how EIRT is based on finding explanatory covariates for items (i.e., variation in locations) and persons (i.e., variation in trait standing), and how in contrast to traditional IRT models, the latent variable is not viewed as causal. The authors provide a demonstration of the model in the domain of self-reported aggression and describe how the approach may be useful in PRO measurement more generally.

As noted in the first part of this chapter, it is often the case in noneducational measurement that constructs are not fully bipolar (i.e., scores are only interpretable on one end of the scale). In Chapter 13, using a gambling addiction scale as an example, Joseph Lucke introduces a new class of unipolar item response models. A distinguishing feature of these types of models is that, unlike traditional IRT models where the mean of the latent trait is defined as zero, in unipolar models, zero is the lowest possible latent trait score (corresponding to individuals with "no symptoms" or "no meaningful trait level"). Lucke also

presents relevant derivatives and information functions, and discusses the fact that models can yield similar item response curves, but very different information functions.

Finally, in noneducational measurement, especially PRO measurement, polytomous item response formats are the norm, and dichotomous response formats the exception. Consequently, polytomous IRT models are more commonly used. However, there are many proposed polytomous IRT models, which begs the question, which are best for PRO data, or does it really matter? In Chapter 14, Remo Ostini, Matthew Finkelman, and Michael Nering discuss issues associated with selecting polytomous IRT models for various applications. They summarize the more commonly applied polytomous IRT models, including some of their more salient differences. The chapter also considers strategies for selecting among different polytomous IRT models and reports on some research that describes how the strategy may play out in practice.

Part III: *Using IRT Models in Applied Problems*

As we noted at the beginning of this chapter, IRT psychometric methods allow researchers to solve both statistical and practical testing problems that are otherwise either not possible or extremely challenging using traditional classical test theory–based approaches. This last part is devoted to chapters that describe how IRT models can be successfully employed in research and practice. One of the primary uses of IRT models is to estimate an individual's position on a common latent dimension or dimensions. Although methods for accomplishing this task with standard unidimensional models are fairly well known, few researchers understand estimating latent trait scores in multidimensional models. To address this, Anna Brown and Tim Croudace summarize problems and solutions for scoring individuals based on multidimensional IRT models in Chapter 15. Models described include the correlated factors, second-order, and bifactor models.

A further distinguishing feature of IRT models is that the theory promotes the development of item banks—sets of items all measuring a single construct with known IRT item parameters. The PROMIS® project, cited previously, has developed many such item banks for PRO constructs. The creation of item banks stands in marked contrast to the historical practice of researchers creating their own preferred measures. In Chapter 16, Dennis Revicki, Wen-Hung Chen, and Carole Tucker provide an overview and summary of methods for developing and evaluating item banks for patient-reported health outcomes. They cover concept identification, qualitative research, item bank development, the basics of the psychometric evaluation of an item bank and resultant measures, and review issues for future consideration in item bank development. Concepts and methods are illustrated with examples from the NIH-sponsored PROMIS® project.

Another often touted advantage of IRT models is that they provide an elegant framework for defining and assessing differential item functioning (DIF)—when the relation between the latent trait and the item responses is not equivalent across examinee populations. Accordingly, in Chapter 17, Roger Millsap, Heather Gunn, Howard Everson, and Alex Zautra summarize methods for evaluating DIF (sometimes referred to as measurement invariance research, or as item or test bias research although DIF does not necessarily imply bias). These authors review definitions of measurement invariance and how violations of invariance are distinguished from simple group mean and variance differences in scores. They then demonstrate how contemporary IRT methods are applied to empirically evaluating measurement invariance.

Yet another claimed advantage of IRT modeling is that it provides a superior method for studying the psychometric properties of items and item category functioning. To illustrate the latter, Kathleen Preston and Steven Reise (Chapter 18) discuss and summarize

methods for evaluating and diagnosing problems with items using the under-used nominal response model (NRM). The NRM can be viewed as a parent model for the generalized partial credit, partial credit, and rating scale models. Preston and Reise illustrate several useful applications of the nominal response model including exploring whether category boundary discrimination parameters vary within an item, whether an item has too many response options, and whether response options are well ordered. The chapter combines Monte Carlo simulations and real data examples.

As noted previously, if individual examinees respond to different sets of test items that measure the same construct, the metrics for the two item sets must be "linked" such that scores are comparable. There is an extensive literature on linking methods for unidimensional IRT models in the educational measurement literature. In Chapter 19, Jonathan Weeks provides a foundation for understanding issues that should be considered when performing either unidimensional or multidimensional test linking. With the emergence of applications of multidimensional IRT models, this latter topic is of critical importance looking forward.

One final potential advantage of IRT lies in the domain of studying change, growth, or development. In Chapter 20, John McArdle, Kevin Petway, and Earl Hishinuma summarize the issues involved in the application of IRT methods for handling growth and changes in scale scores. A real data example of theory testing is provided based on longitudinal data collected from high school students measured in 9th, 10th, 11th, and 12th grades on the *Center for Epidemiological Studies—Depression Scale* drawn from the *Hawaiian High School Health Survey* project. To conclude this volume, in Chapter 21, Steven Reise and Dennis Revicki provide a summary of new IRT problems and future directions for IRT applications in health outcomes assessment.

References

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*, 1179–1194.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (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., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. *Quality of Life Research*, *19*, 125–136.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (in preparation). *Item response theory* (a volume in the Multivariate Applications Series). New York: Routledge/Taylor & Francis Group.

Lai, J., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*(10 Suppl), S20–S27.

Liu, Y., & Verkuilen, J. (2013). Item response modeling of present-severity items: Application to measurement of patient-reported outcomes. *Applied Psychological Measurement*, *37*, 58–75.

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph Supplement 9. *Psychological Reports*, *3*, 635–694.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, T., & Cella, D. (2010). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger. *Assessment*, *18*, 263–283.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, *7*, 347–364.

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

Revicki, D. A., Chen, W. H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., Jensen, M. P., & Keefe, F. J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain*, *146*, 158–169.