ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

The Science and Practice of Item Response Theory in Organizations

Jonas W.B. Lang[1,2] and Louis Tay[3]

[1]Department of Human Resource Management and Organizational Psychology, Ghent University, B-9000 Gent, Belgium; email: jonas.lang@ugent.be

[2]Business School, University of Exeter, EX4 4PU Exeter, United Kingdom

[3]Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907, USA

Annu. Rev. Organ. Psychol. Organ. Behav. 2021. 8:311–38

First published as a Review in Advance on November 11, 2020

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-012420-061705

Keywords

measurement, psychometrics, models, research methods, theory testing, validity

Abstract

Item response theory (IRT) is a modeling approach that links responses to test items with underlying latent constructs through formalized statistical models. This article focuses on how IRT can be used to advance science and practice in organizations. We describe established applications of IRT as a scale development tool and new applications of IRT as a research and theory testing tool that enables organizational researchers to improve their understanding of workers and organizations. We focus on IRT models and their application in four key research and practice areas: testing, questionnaire responding, construct validation, and measurement equivalence of scores. In so doing, we highlight how novel developments in IRT such as explanatory IRT, multidimensional IRT, random item models, and more complex models of response processes such as ideal point models and tree models can potentially advance existing science and practice in these areas. As a starting point for readers interested in learning IRT and applying recent developments in IRT in their research, we provide concrete examples with data and R code.

INTRODUCTION

Measurement is foundational to science, but it is frequently a challenging endeavor that leads to controversy. One idea that has gained traction in many scientific disciplines is to advance the scientific discourse on measurement through the use of formal mathematical or statistical models (Mitchell et al. 2017, Tal 2017). Models make assumptions around measurement explicit and testable. Frequently, models also have value for practice because they can help in retrieving vital information from complex response patterns. In the social sciences, model-based views of measurement are frequently closely associated with the term item response theory [IRT (Borsboom 2006, De Boeck & Wilson 2004, Drasgow & Hulin 1990, Kubinger 2009, McClimans et al. 2017, Wilson et al. 2008)].

The focus of the current article is on describing what IRT is and how it can be used to advance science and practice in organizations. We start by discussing the central role of measurement in both science and practice and then move on and review some of the core ideas and motivations behind using IRT models. We discuss the more traditional view of IRT as a scale development tool, and the more recent perspective that IRT models have broader applications and can serve as useful tools for theory testing. Specifically, IRT can be used to improve researchers' understanding of human behavior in different areas of organizational research by modeling how respondents react to variations of items or behavioral scenarios and by testing theories of the response process. The remainder of the article then describes specific IRT models and their application in four key research and practice areas: testing, questionnaire responding, construct validation, and measurement equivalence of scores. Our review of IRT in these areas is not intended to be exhaustive. Instead, it focuses on highlighting IRT models, research, and perspectives that have the potential to advance research and practice in these areas.

MEASUREMENT, SCIENCE, AND PRACTICE

Measurement is an integral part of most scientific disciplines and typically describes some regular way of linking theoretical and abstract constructs or variables to empirical observations (Michell 2015, Tal 2017). In daily life, many measurement tasks such as measuring the volume of gas in a car's tank, measuring time using a modern quartz watch, or measuring temperature using a modern digital thermometer are perceived as routine activities. Few readers would question the accuracy of these measurements, and some readers may similarly view measurement as a pedestrian activity in science. However, measurement issues are frequently at the center of scientific advancement. Many scientific discourses focus on questions surrounding the appropriateness of measurement instruments as representations of theoretical quantities, discrepancies between different measurement instruments for the same theoretical construct, or the accuracy and replicability of measurements. In line with these observations, it is not surprising that major scientific steps forward are frequently closely tied to innovative new ways of performing measurement. As a motivation for readers to consider the value of studying measurement (and IRT methods later in this article), we provide two examples of how measurement innovations can advance science from other research fields.

As a first example, consider navigation at sea in the eighteenth century (Johnston et al. 2015). Accurate navigation at sea was only partly possible until the mid-1750s because navigators were unable to determine longitude on the equator accurately. This situation changed through major innovations in watchmaking in the 1750s and the resulting availability of highly accurate watches that could be used to measure time on ships—so-called marine chronometers. To calculate longitude, navigators could use marine chronometers to compare the local time (on the basis of known positions of stars on the horizon) with a referent time (on the chronometer) at a known point

(typically the observatory in Greenwich). These developments accelerated discoveries and global trade over the next decades.

As another example, consider the measurements that ultimately led to the wide acceptance of Albert Einstein's theory of general relativity (Coles 2019). Einstein published his general theory of relativity in 1915, but the theory was only widely accepted when astronomer Arthur Stanley Eddington realized that Einstein's theory predicted a different effect of gravity on light than Isaac Newton's theory of gravity. Eddington reasoned that the theories could be tested against each other by recording the deflection of light by the gravitational field of a star (the sun) during a total solar eclipse. He started two expeditions to measure this deflection by recording the position of stars on photographic plates during a total solar eclipse in 1919. These measurements ultimately supported Einstein's theory. Today, Einstein's theory is, for instance, crucial for the functioning of the global positioning system [GPS (Ashby 2003)].

The two examples we described demonstrate why measurement is frequently critically important in science and also in using science in practice. Admittedly, the two cases may seem somewhat removed from the way measurement is frequently done in organizational research. Organizational research typically focuses on measuring behaviors, attitudes, feelings, or observations of individuals, teams, or organizations. The measurement routine most commonly involves collecting a series of responses to questions (or items) and simply aggregating these responses. The rationale for aggregating responses in this manner is typically not explicitly discussed but is commonly based on classical test theory. Classical test theory (Gulliksen 1950) makes the assumption that observations from a set of similar items or observations consist of a linear effect of a typically not more closely defined latent variable or process and random error. As the error in each item is random, the expectation of classical test theory is that aggregating a series of items will lead to greater measurement precision.

In contrast to the relatively simple view of measurement in classical test theory, the measurements of the longitude of ships in the first example includes an elaborate model and knowledge on the precise position of the stars, the rotation of the earth, and the accuracy of building precise marine chronometers. As with the constructs measured in organizational research, longitude is not directly observable for the human eye and can only indirectly be inferred from the configuration of different measurements. However, the accuracy of the approach can be checked indirectly by comparing the relative position of a series of measurements.

Eddington's approach for testing Einstein's theory also includes a relatively complex underlying model. As with the measurement of longitude, the measurement of gravitational force through the light that is reflected by stars during a solar eclipse is quite indirect. The gravitational force itself is, of course, neither visible to the human eye nor can it be directly measured. The presence of the gravitational force can only be inferred indirectly by recording the exact positions of the stars on the photographic plates. However, as with the approach for longitude measurement, surprisingly complex physical processes also can be understood indirectly by carefully recording the exact positions of stars.

What can organizational researchers potentially learn from the two examples we described? One insight is that patterns of measurements about a phenomenon can enable researchers to build and test an elaborate model about a process even when these measurements are relatively indirect. Carefully studying configurations or positions of objects in space enabled researchers to measure the underlying physical properties that led to these configurations. The condition for this to happen, however, is an elaborate and accurate theory. Although organizational research may never develop measurement theories that are as precise as measurement in physics and more specifically in the two described examples, the two examples nevertheless demonstrate the value of gaining additional insights about the nature of an underlying phenomenon.

A second insight is that a complex measurement model may be useful in practice. Longitude navigation enabled a new age of discovery and global trade (but also colonialism), and the general theory of relativity ultimately made GPS navigation possible.

A third insight that is closely related to the two earlier insights is that applying a simplistic model that ignores the nature of the underlying processes has no guarantee of being useful at all. Imagine that a researcher attempts to study the physical property of longitude or the gravitational force of the sun by simply aggregating light intensity measurements for different stars.

WHAT IS ITEM RESPONSE THEORY?

IRT models are by far not as elaborate as the measurement examples from other research areas we discussed in the previous section. Yet, in many cases, modern IRT approaches can still provide insights that can go beyond more simple measurement approaches and help in understanding the underlying latent processes. The core motivation behind this article is, therefore, to describe how, why, and when IRT can advance measurement and especially theory and practice in organizations beyond simplistic approaches such as classical test theory. We seek to present the material in a nontechnical way and describe the practical and theoretical ideas behind the use of IRT. In a nutshell, IRT can, in many cases, be to measurement in organizational research what solar eclipse measurement is in the physical sciences: an important measurement device and theory testing tool.

At its core, IRT can be understood as a modeling approach that aims to describe relationships between responses to test items and underlying latent constructs using formalized statistical models. In the broadest sense, IRT is a framework that enables researchers to test theories about configurations of item responses in a somewhat similar way to how Eddington used configurations of stars to test Einstein and Newton's theories. As with most theories that describe mechanisms or relationships, IRT can frequently be useful for research and practice. However, the specific benefits of IRT vary somewhat depending on the IRT approach that is employed, the research context, and the philosophical perspective of the researcher.

Basic Item Response Theory Models

Historically, IRT started as a formalized mathematical model for ability or skill measurement in testing (Bock 1997, Drasgow & Hulin 1990, Hambleton et al. 1991, Lord & Novick 1968, van der Linden & Hambleton 1997). The simplest IRT model for this purpose is the 1PL or Rasch model (named after the Danish mathematician Georg Rasch). This model describes the probability P_i that a respondent with a given ability or characteristic theta (θ) solves a given dichotomous (correct/incorrect) item i. This probability is a function of the item easiness b_i, and follows a logistic function with e [approximately equal to 2.718 and accessible by typing exp(1) in most statistics programs] as the logarithmic constant, $P_i(\theta) = \frac{1}{1+e^{-(b_i+\theta)}}$ when the parameters are in the modern intercept-slope metric that is now most commonly used.[1] The model can be fitted to a set of item responses by an optimization algorithm. This procedure yields both the θ and b_i estimates that are easy to interpret because they are placed on a common continuum. For instance, a person with an ability of $\theta = 1$ that works on an item with an easiness of $b_i = -2$ has a probability to solve the

[1] An alternative for writing the model is the more traditional $P_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}}$. This version of the model also includes a scaling factor D that is 1 in the original version of the model but frequently also 1.7 in modern applications. 1.7 transforms the θ and b_i estimates such that the metric of the estimates approximately corresponds to a normal distribution. In this parametrization of the model, the easiness parameters become difficulty parameters (i.e., the higher the values the more difficult the item is).

items correctly of $\frac{1}{1+e^{-(-2+1)}} = 0.27$, or approximately 27%. In practice, the 1PL/Rasch model is similar to classical test theory in the sense that the items do not have different weights and thus the estimated θ values are identical to the simple sum score of all items from classical test theory. On the surface, the model does not add anything new in the way test scores are estimated. However, the model still has several theoretical advantages for scaling and test development.

One important advantage of the 1PL and other IRT models over classical test theory approaches is that the fitting of a logistic function also solves an important problem with measurement error in test scores. In classical test theory, measurement error is construed as a characteristic of the test—the so-called test reliability. Test reliability is typically defined as the squared correlation between test scores and the true underlying characteristics, or alternatively, as the correlation between two perfectly parallel tests. Test reliability is a property of samples and not a property of persons. In research that focuses on samples, this is rarely a problem and test reliability provides useful information. However, when test scores of individuals are of interest as with in most applications of IRT in practice (e.g., cognitive ability testing of applicants), classical test theory needs to resort to a makeshift assumption. The classic test theory framework then simply assumes that the amount of measurement error at the level of the sample directly generalizes to the test score of each respondent. In other words, it assumes that the amount of measurement error is the same for each respondent. IRT overcomes this quite problematic idea and enables researchers to estimate a specific standard error (SE_θ) for each response pattern. This standard error varies across the continuum of ability scores, depending on the difficulties of the items included in the test. One reason is that items provide the most information about the true underlying ability when the item difficulty is close to the θ of the respondent. Accordingly, a test with few items close to the ability of the respondent provides less measurement accuracy than a test with many items around the respondent's ability level. This property of test scores in IRT is actually what one would intuitively assume to be the case. Most readers may be familiar with the phenomenon of being confronted with a test that is way too easy for them so that the test does not provide much information beyond showing that the ability level of the respondent is obviously higher than most of the test items. Another reason is that response patterns that are inconsistent (e.g., high cognitive ability individual getting many simple items incorrect but hard items correct) have low measurement accuracy as opposed to response patterns that are consistent (e.g., high cognitive ability individual getting some hard items correct and most simple items correct). IRT still makes the estimation of a test reliability parameter possible even though such a parameter is traditionally not directly a part of the IRT framework. In the IRT context, the reliability is simply the average amount of measurement error in a sample of respondents. The most intuitive form of this test reliability—empirical IRT reliability—therefore simply relies on the average of the standard errors of the persons in a sample of respondents (Kim 2012, Lang 2014).

Another advantage of the 1PL model is that it provides a testing framework for checking whether responses to the items follow the logistic curve that the model suggests. When this is the case, a couple of desirable properties of the model apply. For instance, the model assumes that the logistic curve has the same slope for all items in the test. As a result of this property, Rasch himself used the term specific objectivity—the item parameters and the ability parameters become independent of each other. For instance, measuring two individuals with another set of items should lead to the same θ parameters for the individuals. Although these properties are desirable, it can be difficult in practice to find test items that fulfill the a priori properties of the model.

A frequently used alternative to the 1PL model is the 2PL model (Birnbaum 1968, Hambleton et al. 1991). The 2PL adds a slope parameter for each item a_i in addition to the easiness parameter, $P_i(\theta) = \frac{1}{1+e^{-(b_i+a_i\theta)}}$. The introduction of this slope parameter makes the model far more flexible and

Table 1 Item parameter estimates from a basic item response theory analysis using the 2PL model and a dataset on the Law School Admission Test from Bock & Lieberman (1970)[a]

	a_i	b_i
Item 1	0.99	1.86
Item 2	0.81	0.81
Item 3	1.71	1.80
Item 4	0.77	0.49
Item 5	0.74	1.85

[a]Empirical reliability is 0.45, a_i is the slope parameter, and b_i indicates item easiness.

Table 2 Model fit statistics from a basic item response theory analysis using the 2PL model and a dataset on the Law School Admission Test from Bock & Lieberman (1970)

Fit statistic	2PL model
Akaike information criterion	5,337.61
Akaike information criterion corrected	5,337.83
Sample-size adjusted Bayesian information criterion	5,354.93
Hannan-Quinn criterion	5,356.26
Bayesian information criterion	5,386.69
loglikelihood	−2,658.81

allows it to differentiate between items that are more successful in capturing the characteristic of interest and those that are less successful. The a_i is conceptually similar to a factor loading in the context of exploratory or confirmatory factor analysis. The person parameter estimates (θ) are thus not identical to sum scores and instead present a weighted average of the item responses. A disadvantage of the 2PL model is the fact that it loses some of the desirable properties of the 1PL surrounding specific objectivity (see Irtel 1995 for a discussion). However, the model still preserves many of the core advantages of IRT such as placing both items and persons on a common continuum.

Tables 1, **2**, and **3** and **Figure 1** illustrate the described mechanisms of IRT using a classic example taken from Bock & Lieberman (1970). **Tables 1**, **2**, and **3** report item parameter estimates, typical model fit statistics (De Boeck et al. 2011, Sen & Bradshaw 2017), and response patterns with ability and measurement error estimates, respectively. The example model is a 2PL analysis, and the data include a small set of items from a version of the Law School Admission Test. The analyses

Table 3 Six response patterns with the corresponding θ and SE_θ estimates from a basic item response theory analysis using the 2PL model and a dataset on the Law School Admission Test from Bock & Lieberman (1970)

Person	Item 1	Item 2	Item 3	Item 4	Item 5	θ	SE_θ
Person A	0	0	1	0	0	−1.095	0.665
Person B	0	1	0	0	1	−1.046	0.665
Person C	0	1	1	1	0	−0.243	0.705
Person D	1	0	0	1	1	−0.745	0.692
Person E	1	1	0	0	0	−0.934	0.668
Person F	1	1	1	0	1	0.265	0.665

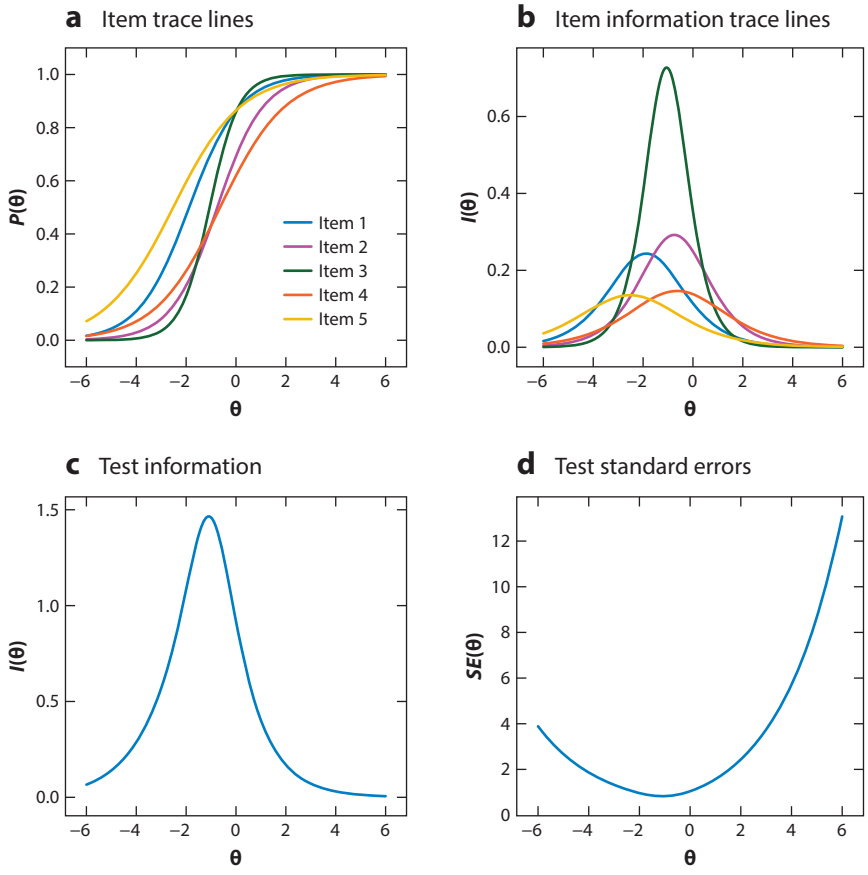

Figure 1

Results from a basic item response theory (IRT) analysis of a dataset on the Law School Admission Test reported in Bock & Lieberman (1970). Panel *a* shows the item characteristic curves from the model. These curves describe the probability (*P*) of a correct response for each item as a function of the ability of the person (θ). Panel *b* illustrates the amount of information (*I*) each of the items provides for different ability levels. Panel *c* plots the test information curve. This curve is a direct result of the amount of information provided by the items shown in *b*. Finally, panel *d* illustrates the amount of measurement error (standard error, *SE*) for ability scores for the test at different ability levels. This curve is a key difference between IRT and classical test theory (for the latter, the curve would be a straight line). The graph was originally generated using the mirt package (Chalmers 2012).

were conducted in the freely available open source environment R using the mirt package.[2] mirt (Chalmers 2012) is one of several available R packages for IRT analyses. Other software that can conduct similar analyses such as the example in **Figure 1** include the R packages ltm (Rizopoulos 2006), sirt (Robitzsch 2020), lme4 (Bates et al. 2015b, Doran et al. 2007), and lavaan (Rosseel 2012) as well as commercial software such as Mplus (Muthén & Muthén 2015).

As shown in **Figure 1*a*,*b*** and in **Table 1**, the items vary in their slope/a_i and easiness/b_i parameters and thus yield different amounts of information across the ability continuum. As a result

Supplemental Material >

[2]The **Supplemental Appendix** provide the code and output for running all examples in this article.

and as illustrated in **Figure 1*c***, the overall information that the test provides across the ability continuum also varies. **Figure 1*d*** shows the consequences of the varying amounts of information across the ability continuum for the standard error of test scores. The test in the example shows the highest measurement precision slightly to the left of the ability continuum.

There are many other IRT models beyond the 1PL and 2PL (e.g., De Boeck et al. 2011, Tay et al. 2015, van der Linden & Hambleton 1997). However, we have restricted our discussion at this point to illustrating and explaining the basic mechanisms and rationales behind IRT as a scale development tool using these basic models. Most other IRT models can be understood as extensions of these models to accommodate specific purposes such as, for instance, polytomous IRT models to accommodate ordinal (Likert-type) rating scales, partial-credit models to deal with tests in which some answers are partly correct, or the 3PL model to account for guessing by respondents.

APPLICATIONS OF ITEM RESPONSE THEORY IN ORGANIZATIONAL RESEARCH

The remainder of this article is organized along practical applications of IRT in four critical areas of organizational research: testing/assessment, questionnaire responding, construct validation, and the group equivalence of scores. As we discuss these key areas, we also introduce the different views of IRT that are relevant for these areas: traditional views of IRT as a scale development tool in large-scale testing and a more modern view of IRT as a set of theory testing tools about measurement. The view of IRT as a theory testing tool also includes analytical frameworks such as explanatory IRT and response-process IRT. Viewing IRT as a theory testing tool is also closely linked to an alternative internal or psychometric view of validity that focuses validation on linking item responses to latent attributes (e.g., Borsboom & Mellenbergh 2007, De Boeck & Wilson 2004, Embretson 1998, Wilson 2005) and contrasts with the traditional view of validity as a property of tests with many different facets that requires a large nomological network of relationships with other tests (Messick 1989a,b).

TESTING/ASSESSMENT

Practical: Item Response Theory as a Technical Tool to Improve Efficiency in Large-Scale Testing

The advantages of IRT for scale development, which we discussed when we described the 1PL and 2PL, may seem somewhat theoretical at first sight. However, these characteristics of IRT models have significant practical advantages and have solved several issues that large-scale testing programs faced since the start of the twentieth century. Large-scale testing programs are used in educational assessment for evaluating ability in students or selecting applicants into educational programs. Several large-scale testing programs have also been implemented in organizational contexts to select future employees or evaluate their progress in training. An early example is the US Army Alpha, which assessed the ability and knowledge of soldiers in World War I (Yoakum & Yerkes 1920).

For large-scale testing, classical test theory has two important limitations (Kolen & Brennan 2004, Lord 1980, Wainer et al. 2007, Weiss 1982). First, using classical test theory requires that all respondents work on all items to make the overall scores easily comparable. Second, with classical test theory, developing a different version of a test requires labor-intensive equivalence research (e.g., present the two test forms to identical groups or let a group of respondents work on both test forms in counterbalanced order). Both limitations can be overcome by IRT. IRT

makes respondents who worked on different sets of items comparable because the item and ability parameters can be brought to a common scale.

Many testing programs use this advantage of IRT to link different forms of a test through test equating procedures. Test equating procedures (Bock et al. 1988, Kim & Cohen 1998, Kolen & Brennan 2004) only require that different forms of a test share some common linking items. It is then possible to bring the metric of the different versions of the test to the same scale even when the samples to which the tests were administered differ in their ability level.

One form of flexible item presentation is computerized adaptive testing (Oswald et al. 2015, Thissen et al. 2007, Weiss 2004). In an adaptive test, the item difficulty is tailored to the respondent. Respondents typically start with an item of average difficulty and the next item is then adaptively selected on the basis of the result of the first item (i.e., correct or incorrect). The idea of an adaptive test was a part of the first intelligence tests developed in 1905 by Binet & Simon (1916) and adapted and further developed by Terman (1916). These tests included several versions of scales with varying difficulty and were administered by a psychologist in one-on-one sessions. The psychologist would choose an easier or more difficult second part of a scale depending on the performance of the respondent in an initial scale. The idea of adaptive testing was subsequently abandoned for a considerable time, mainly because it was not practically feasible for large-scale testing with large numbers of respondents but also because of the difficulties in making scores comparable and conducting scale development without an analytical framework (Yerkes 1917). The development of IRT provided the framework that made adaptive testing feasible. Modern computerized adaptive testing developed from the 1970s (Weiss 2004). In modern adaptive testing, the items are first all administered to one or more reference samples to estimate item parameters for all items (a_i and possibly b_i). These item parameters are then subsequently added to an item bank. The item bank is then used in the computerized adaptive test to estimate ability after the respondent worked on each item. After a correct answer, a more difficult item that is closest to the new ability estimate for the respondent is selected by the computer program. After an incorrect answer, an easier item that is closest to the new ability estimate of the respondent is chosen. The computer program also estimates the measurement error of the ability estimate after each answer by the respondent. The adaptive process of selecting and administering items typically continues until a specific measurement precision is reached. With modern computerized adaptive tests, the length of the test thus varies across respondents and is affected by several factors such as the ability of the respondent, the availability of items at a particular point of the ability continuum, and the consistency of the item responses by the respondent. Not all modern computer-based large-scale tests are fully adaptive. However, many of these tests include adaptive elements and, for instance, start with an initial set of items and then select an easier or more difficult set of items in a second step. The application of IRT-based adaptive testing can also provide a way to enhance test security (i.e., lower item exposure and reduce overlaps between applicants); this is particularly important in unproctored Internet testing, which is becoming more popular among organizations due to convenience and cost-savings (Lievens & Burke 2011).

Item Response Theory as a Tool for Scale Development for Measurement Purists and Statisticians

In addition to the use of IRT in large-scale testing programs, IRT has also been advocated by measurement purists and statisticians to improve tests with a fixed set of items. The motivation in these applications is frequently to improve scientific practice, and the practical advantages for flexible test administration and scaling are less important.

For instance, organizational researchers have conducted IRT analyses of scales developed without IRT to examine to what degree the responses on the scales fit the IRT assumptions and to provide recommendations for future improvements of the scales (DeSimone & James 2015, Hernández et al. 2004, Zickar 1998). One argument for the broader scientific use of IRT is also that IRT guarantees "pure" unidimensional test scores. Measures developed on the basis of unidimensional IRT models, for instance, do not suffer from the common issues with estimating test reliability that result from measures that do not have a clear unidimensional structure (Drasgow & Hulin 1990, Revelle & Zinbarg 2009, Sijtsma 2009). Although the use of IRT to improve scientific practice has gained some traction within the community of organizational researchers (Foster et al. 2017), its use as a scale development tool may be more widely established in other research domains such as, for instance, medical outcome assessment (McHorney & Monahan 2004, Reeve et al. 2007, Smith & Burns 2014) or screening in clinical psychology (Bliese et al. 2008).

Critical readers may also note that the applications of IRT that we reviewed in this section offer clear advantages but seem more like a nuisance than a fundamental paradigm shift. Applying a 1PL to a scale of homogeneous items yields a score that is perfectly correlated with a simple sum score, and the item difficulty parameters from the 1PL analysis are also very similar to parameters from test analysis using classical test theory. Simply using IRT thus does not necessarily allow the organizational researcher to gain important theoretical insights into the nature of the underlying mechanisms. However, IRT as a scientific framework also has important potential advantages beyond its use as a scale development tool that many organizational researchers may not be fully aware of. In the next section, we describe how researchers can gain new insights from IRT that are interesting, helpful, and go beyond what is available from classical test theory approaches.

Item Response Theory as a Tool to Test Theories About Test Items: Explanatory Item Response Modeling

In our earlier definition of IRT in the section titled What Is Item Response Theory?, we described IRT as a modeling approach that aims to describe relationships between responses to test items and underlying latent constructs using formalized statistical models. IRT is thus fundamentally about modeling and mapping responses of the latent trait(s) to the observed responses. In the scale development examples we discussed in the previous section, each item receives its own difficulty parameter estimate. Although this approach is useful for maximizing the information that can be gained from an item, especially in a flexible model such as the 2PL, these models do not provide insights into why a particular item is difficult or easy for respondents or discover relationships that are not easily apparent from descriptive examinations of the data. In other words, these models do not provide insights into the drivers behind the responses. IRT approaches that use IRT as a theory testing tool seek to open up this black box and utilize IRT as a theory testing tool to understand the drivers of human responses to stimuli or uncover processes behind responses that are not easily apparent through descriptive examinations of the data.

The first researcher who proposed that IRT can be useful to test theory was possibly Austrian Gerhard Fischer (Fischer 1973, Hornke 2002, van der Linden & Hambleton 1997). Fischer suggested the use of item features (e.g., item generation principle, similarity of content, wording, style) as predictors of test responses instead of a specific item difficulty parameter for each item. The approach can be illustrated by thinking of the traditional 1PL/Rasch model as a logistic regression model with dummy codes for items and persons as predictors as well as a dichotomous dependent variable. Readers familiar with regression analysis will likely agree that such a model is not a very parsimonious model. Fischer suggested the linear logistic test model (LLTM) as an alternative. In the LLTM, the dummy codes for items are exchanged with predictors of item

characteristics; statistically, this also reduces the number of dummy codes and makes the model more parsimonious. For instance, a researcher may use a set of ethical dilemmas about behavior in organizations and may systematically vary the characteristics of these dilemmas (e.g., young versus old person in the scenario, presence of observers or not) to study when respondents decide to violate organizational policies. The characteristics of the dilemmas will probably not explain all of the differences in the difficulty of the scenarios and thus an LLTM will typically provide a less good fit to the data than a 1PL/Rasch model. However, the use of an LLTM can provide useful insights in this case because through the use of item features as predictors it explains why respondents react to the items in a certain way. An LLTM approach has many parallels to a repeated-measures experiment, and thus it is frequently feasible to interpret the estimates causally. Within the LLTM framework, it is possible to vary certain item characteristics between persons while ensuring that there are items common across conditions so that all items and ability estimates can be placed on a common scale. For instance, between-person manipulations can be useful when a researcher suspects that order effects may distort estimates.

The original LLTM had several limitations, including the fact that the approach was not capable of incorporating unexplained variance in the difficulty of the item parameters. The LLTM framework was therefore rarely used in practice until the early 2000s when a group of researchers showed how software originally designed to estimate multilevel models can be used to estimate the LLTM and several extensions of it (De Boeck & Wilson 2004, Embretson 1998, Wilson 2005, Wilson et al. 2008). These insights led to the development of an analytical framework that is today most widely known as explanatory item response modeling. Within the explanatory framework, researchers can fit models such as the LLTM plus error (LLTMe)—a model that accounts for remaining variation in the item difficulties after predictors are added, the multilevel extensions of IRT models, or models that add person predictors (e.g., De Boeck et al. 2011, Tay et al. 2016)—so-called latent regression IRT models. The use of software to estimate multilevel models allowed new flexibility by allowing multiple dimensions, overlapping dimensions, and random items. To understand the concept of random items (De Boeck 2008), it is frequently helpful to reconsider our earlier example of a basic IRT model as a logistic regression model with dummy codes for both items and persons. This example is accurate for some early forms of IRT models. More modern versions of IRT models additionally assume that the person parameters are sampled from a normal distribution. This assumption adds additional model assumptions but also makes the estimation of the model simpler and faster. Random items in the explanatory IRT framework additionally assume that the items come from a random distribution and thereby make models even simpler (but also adding the additional assumption that the items are sampled from a normal distribution).

The explanatory IRT approach includes ideas that are in line with some recent development in selection and assessment (Lievens & Sackett 2017, Sackett et al. 2017). These researchers have issued calls to go beyond the typical emphasis on researching selection and measurement procedures as holistic entities to examine their inner workings (i.e., a modular approach to selection). A broader use of explanatory IRT could be one way to accomplish such a novel approach to test development, theory development, and theory validation.

Example of an Explanatory Item Response Theory Analysis

Suppose an organizational researcher has argued that a new construct of sensitive thinking is important in social situations and also predicts leadership of virtual teams. The researcher also argues that the construct includes a total of three suboperations: (*a*) detecting sensitive situations, (*b*) applying sensitive reasoning, and (*c*) balancing emotions and thoughts. The researcher also reasons that suboperations *b* (applying sensitive reasoning) and *c* (balancing emotions and thoughts) would

Table 4 Model fits for a 1PL model, an explanatory item response theory (IRT) analysis using the linear logistic test model plus error (LLTMe), and a multidimensional (bifactor) model in the same dataset

Fit statistic	1PL	LLTMe	Multidimensional
Akaike information criterion	14,510.96	14,628.24	14,491.51
Akaike information criterion corrected	14,513.70	14,628.54	14,516.13
Sample-size adjusted Bayesian information criterion	14,536,97	14,636.57	14,566.43
Hannan-Quinn criterion	14,552.30	14,641.48	14,610.59
Bayesian information criterion	14,616.33	14,661.96	14,794.97
loglikelihood	−7,230.48	−7,306.12	−7,173.76

be easier when the item includes a hypothetical social situation with a colleague than when the item is a social situation with a follower. To study sensitive thinking, the researcher decides that he/she wants to develop a new test that measures sensitive thinking using hypothetical scenarios. Each scenario will include a couple of response options and one of the options is the correct answer.

In the traditional validity framework, the researcher would develop items, and then use classical test theory or a basic IRT model such as the 1PL/Rasch model to study the properties of the scale. In the next step, the researcher would use the items to predict leadership outcomes in virtual teams. In contrast, developing a measure using an explanatory IRT framework by varying core elements of the new construct and examining whether item responses reflect features of the hypothetical scenarios allow the researcher to gain insights about the mechanisms underlying the test items. **Table 4** provides model fit statistics for both analyses. The first column of **Table 5** provides the easiness parameter estimates from the simple 1PL/Rasch model. As **Table 5** shows, the 1PL analysis provides the expected item difficulties for each item. However, it is difficult to see a pattern in these estimated difficulties. In contrast, **Table 6** shows the estimates for the explanatory LLTMe analysis. The LLTMe analysis provides important additional insights. As expected, the suboperations contribute to the difficulty of the item. Furthermore, the second (applying sensitive reasoning) and third (balancing emotions and thoughts) suboperations are indeed relatively easier when the scenario is with a colleague instead of a follower. The model fit information in **Table 4** also reveals that the LLTMe provides a slightly less good fit compared to the 1PL model as indicated by the smaller values of the fit criteria. The reason is that the LLTMe is a simpler model because the random item parameters are not technically a part of the model. A caveat with model comparisons is that the two models are not nested and represent two different ways to think about test items. The important take-away message is that the explanatory approach provides important insights into how the items function psychologically.

Multidimensional Item Response Theory

Another development that we highlight in this section on IRT in testing is multidimensional IRT (Ackerman 1989, De Boeck et al. 2011, De Boeck & Wilson 2004, Reckase 2009). Most IRT approaches in practice are still unidimensional because unidimensional models are familiar to most assessment practitioners and can easily be interpreted. Multidimensional IRT models, in contrast, are relatively complex to estimate. However, the availability of faster computers, new optimization procedures, and more efficient estimation (Bates et al. 2015b, Cai 2010) have made the use of multidimensional IRT much more readily available for researchers.

One important advantage of multidimensional IRT is that the approach can deal with tests with a complex dimensional structure. It is quite common that test items load on a main dimension and additionally on a narrower side dimension and thus violate the assumptions of unidimensionality

Table 5 Item parameter estimates for a 1PL model and a multidimensional (bifactor) model fitted on the same data[a]

	1PL	Multidimensional model				
Item	b_i	a_{1i}	a_{2i}	a_{3i}	a_{4i}	b_i
Item 1	−0.25	0.55	0.11	—	—	−0.24
Item 2	−0.44	0.99	−0.03	—	—	−0.47
Item 3	0.22	1.18	0.64	—	—	0.27
Item 4	−0.29	0.29	0.21	—	—	−0.27
Item 5	0.18	1.04	—	−0.65	—	0.21
Item 6	0.03	1.12	—	0.23	—	0.04
Item 7	0.81	0.94	—	0.45	—	0.88
Item 8	0.06	0.77	—	0.57	—	0.07
Item 9	0.45	0.43	—	—	0.20	0.43
Item 10	0.51	0.59	—	—	0.43	0.51
Item 11	0.24	0.63	—	—	0.40	0.25
Item 12	0.31	0.47	—	—	−0.24	0.30
Item 13	−0.64	0.67	0.30	—	—	−0.64
Item 14	−0.24	0.83	0.08	—	—	−0.25
Item 15	0.05	1.12	−0.29	—	—	0.07
Item 16	0.03	0.65	0.60	—	—	0.04
Item 17	1.40	0.44	—	0.26	—	1.34
Item 18	0.87	0.60	—	0.30	—	0.86
Item 19	1.04	0.49	—	0.06	—	0.99
Item 20	0.84	1.58	—	0.42	—	1.09
Item 21	1.91	0.98	—	—	0.07	2.04
Item 22	2.01	0.61	—	—	0.12	1.97
Item 23	1.64	0.91	—	—	1.50	2.22
Item 24	1.69	0.98	—	—	−0.24	1.82

[a] a_{1i}, a_{2i}, a_{3i}, and a_{4i} are slope parameters; b_i indicates item easiness.

Table 6 Item parameter estimates for an LLTMe (linear logistic test model plus error)

Parameter	Estimate	*SE*	*z*
Fixed effects			
Sub 0 (detecting sensitive situations)	−0.302	0.136	−2.215*
Sub 1 (applying sensitive reasoning)	0.141	0.085	1.661
Sub 2 (balancing emotions and thoughts)	0.464	0.185	2.511*
Context 1 (colleague)	−0.003	0.209	−0.013
Sub 1 × context 1 (colleague)	0.863	0.392	2.199*
Sub 2 × context 1 (colleague)	1.193	0.227	5.264**
Random effects variances			
θ (person)	0.518		
Item	0.059		

*$p < 0.05$; **$p < 0.01$.

that unidimensional IRT models make. Unidimensionality, in practice, means that the items are linked only by the common trait that they measure. However, within a single test, it is common to find sets of items that share a common question format, context, or content area. In cases of this type, it makes sense to estimate models that allow items to have two loadings—one loading on the primary factor of the test and a second loading on a subfacet to reflect similar formats or content in sets of items. There are several different approaches for implementing multidimensional models. One popular model is the bifactor approach, which can be understood as a 2PL model with two loadings and uncorrelated factors. The third column in **Table 4** and the second column in **Table 5** provide such an analysis of the dataset we discussed in the previous section on explanatory item response modeling using the mirt package. As shown in **Table 5**, the item difficulties for the multidimensional model are very similar to the item difficulties in the first column of **Table 5** for the 1PL that we already discussed earlier. However, each item has two loadings that are freely estimated. In contrast to the LLTMe analysis in the previous section that assumes that a unidimensional latent dimension exists that involves several different mechanisms or operations that cause the difficulties, the multidimensional model is quite flexible but also attempts to explain what causes the difficulty of the items. One important limitation of the bifactor approach is that it can lead to counterintuitive findings in rare situations (van Rijn & Rijmen 2012). Specifically, a response to an item can also lead to a lower score on the overall dimension. The reason is that the model seeks to estimate all components as accurately as possible and thereby maximizes the use of information. Scores on subdimensions can therefore affect scores on main dimensions and vice versa. This sound methodological idea can sometimes go against the intuitive interpretation of tests as competitions. An alternative noncompensatory model that addresses this situation is the testlet model (Wainer et al. 2007). In the testlet model, the loadings of the items on the main dimension and the subfacet for each item are set equal, and the variance of the subfacets is freely estimated (typically set to 1 in the bifactor model). This model specification is less flexible but ensures that counterintuitive findings cannot occur.

Another frequent application of multidimensional IRT is in the assessment of constructs that have a natural multidimensional structure. For instance, IRT models can be used to model responses in construct-driven situational judgment tests (Lievens 2017, Lievens & Sackett 2017). A variant of the multidimensional approach that has recently received attention in the literature is cognitive skill modeling procedures (e.g., DINA models; see de la Torre 2009). These procedures were originally developed in the context of educational research to better understand necessary skills in complex items but have recently been used also to model the presence of knowledge in situational judgment tests (Sorrel et al. 2016). Fundamentally, these models are multidimensional models with discrete latent variables (i.e., the discrete latent variables capture whether a skill is present or not).

QUESTIONNAIRE RESPONDING

In organizational research, questionnaires are perhaps the most widely used tool to capture information about workers and their workplaces (e.g., Podsakoff & Organ 1986, Rogelberg et al. 2008). Given that most questionnaire instruments are quantitatively scored (e.g., Likert-type responses, categorical responses), IRT can readily be used to study the response processes behind questionnaire responses.

The use of IRT frequently has the potential to advance research using questionnaires in important ways. In the remainder of this section, we highlight some core principles and novel developments in using IRT for questionnaire responses.

Item Response Theory as a Tool to Study the Assumptions of Rating Scales

IRT has historically been used for modeling test responses to deal with categorical right-wrong scoring using models such as the 1PL or 2PL. However, as we mentioned in our initial discussion of IRT, models that can deal with polytomous, ordered, or categorical responses are straightforward extensions of basic dichotomous models. For models with multiple response categories, IRT has distinct advantages because it allows researchers to systematically study assumptions about the scaling of responses. For example, the Job Descriptive Index (Smith et al. 1969), a widely used job satisfaction measure, has response options of "Yes," "?," and "No." IRT modeling revealed that individuals respond to the category "?" in incommensurate ways; a majority of individuals do not use the "?" category, whereas others use it as a middle category between "Yes" and "No" (Hernández et al. 2004).

Item Response Theory as a Tool to Test Theories: Response Process Item Response Theory

The use of IRT for modeling survey questionnaires has increased in appeal over the course of the last two decades because researchers have realized that IRT not only can serve measurement purposes [e.g., ensuring measurement equivalence for between-group comparisons (Tay et al. 2015)] but also can make substantive contributions. This perspective represents a shift from viewing IRT methods merely as tools for substantive research to applying IRT methods to operationalize, detect, and understand substantive phenomena that would previously have been inaccessible (Greenwald 2012). These advances are recognized as methodological-substantive synergies for the scientific process (Marsh & Hau 2007). One burgeoning area of IRT research is in modeling the response processes individuals use to respond to self-reported typical behaviors such as attitudes, affect, personality, and interests, which are key constructs in organizational research (Böckenholt 2012, De Boeck & Partchev 2012, Lang 2014, Tay et al. 2009, Tay & Ng 2018).

In the earlier section on testing, we already described explanatory IRT modeling as one important way in which IRT can be used for theory testing. The use of IRT for modeling response processes is a second way that IRT can contribute to advancing theory. We use the term process to indicate that these models do not simply assume a likelihood of some sort of reaction at a certain threshold such as in basic IRT models but a more complex series of decisions or comparisons by the respondent.

Process IRT models have developed gradually and over the course of many decades. The broad use and availability of these methods, however, are relatively recent. Starting in the late 1990s and 2000s, psychometricians became increasingly aware that some of the more complex decision-making models in economics, psychophysics, or marketing (e.g., McFadden 2001, Thurstone 1927) that were typically fitted to large groups of participants could be transformed into IRT models by adding person-specific parameters. Another key development was that the availability of advanced software made revisiting concepts from the early twentieth century possible because item parameter estimation became suddenly feasible (e.g., Böckenholt 2001, De Boeck & Wilson 2004, Roberts 2001).

A novel element in process IRT models is that IRT response processes are not restricted to specific items. The IRT processes can span several items or response options. Another core novel element in process IRT models is the fact that these models can extract information in a way that classical test theory–based tests fundamentally cannot. In other words, the information that these models can extract from data can reveal important novel information and can provide insights into underlying response processes that would not be available without these models.

Response process IRT includes several different types and classes of models. We start by highlighting two response process approaches that may be particularly promising and are easy to use using available packages in R: ideal point models and response tree models.

Ideal point models. The development of an ideal point IRT model (Roberts 2001, Thurstone 1928) has enabled the detection of ideal point responding to self-reported typical behavior; individuals respond most positively to items closest to their latent trait standing [e.g., endorsing "happy" when one is "happy" but not when one is "extremely happy" or "somewhat happy" (Tay & Kuykendall 2017)]. It had previously been assumed that individuals used dominance responding; individuals respond most positively the higher they are on the latent trait compared to the item location (e.g., a greater probability of endorsing "happy" when one is "extremely happy" compared to "happy"). This assumption resulted from past research that had exclusively relied on dominance measurement models (e.g., sum-score, factor analysis). Dominance models had originally been developed to be applied to maximal performance constructs (e.g., cognitive ability) and were then simply used for self-reported typical behaviors.

Figure 2 illustrates the conceptual differences between ideal point and dominance models. In **Figure 2*a***, the ideal point item characteristic functions of nine dichotomous extraversion items are shown. These item characteristic functions come from the most widely known ideal point model, the generalized graded unfolding model (Roberts et al. 2000), and were fitted using the mirt package in R (code available in the **Supplemental Appendix**). Although most modern ideal point approaches are designed for data with multiple categories, we use dichotomous items here for illustrative purposes. The shape of the response function in **Figure 2*a*** markedly differs from the more commonly used form of the response function in **Figure 2*b***. The model underlying the item characteristics in **Figure 2*b*** are from the standard dominance IRT model (2PL or graded response model with several categories).

Supplemental Material >

The use of ideal point IRT models has led to substantive advances on researchers' understanding of some self-reported typical behaviors. For instance, ideal point IRT models can test the empirical viability of core affect theory [where positive and negative affect are in opposition (Tay & Kuykendall 2017)], which is widely used in organizational research (Lord & Kanfer 2002).

Research has also studied what happens when dominance models are incorrectly applied to ideal point data. One important finding is that ideal point responding on a bipolar dimension (positive-negative affect) can lead to ostensible orthogonal unipolar dimensions (e.g., positive affect and negative affect) when dominance models are applied (Davison 1977). Personnel selection using instruments such as personality or vocational interests also found that the incorrect use of dominance models for ideal point data can lead to inaccurate rank ordering of individuals at the high end of the continuum (Stark et al. 2006, Tay et al. 2009). Moreover, the misapplication of dominance models on ideal point data can lead to problems in detecting curvilinear effects (Carter et al. 2014). Accordingly, it seems important for researchers either to make a theoretically informed choice between ideal point and dominance models when they develop new measures, or alternatively, to compare both models on their datasets (Nye et al. 2019).

Tree models. Another growing area of research is in the use of IRT tree models (Böckenholt 2012, Böckenholt & Meiser 2017, De Boeck & Partchev 2012, LaHuis et al. 2019, Plieninger 2020). Tree models assume that a response tree that includes several stacked decisions can frequently underlie items with several response categories. For instance, individuals responding to Likert-type responses (e.g., strongly disagree, strongly agree) use multiple judgment points to reach a final decision as to which option to choose. For example (Böckenholt 2012), individuals may choose first between an indifferent (i.e., neutral) or directed response (i.e., strongly disagree,

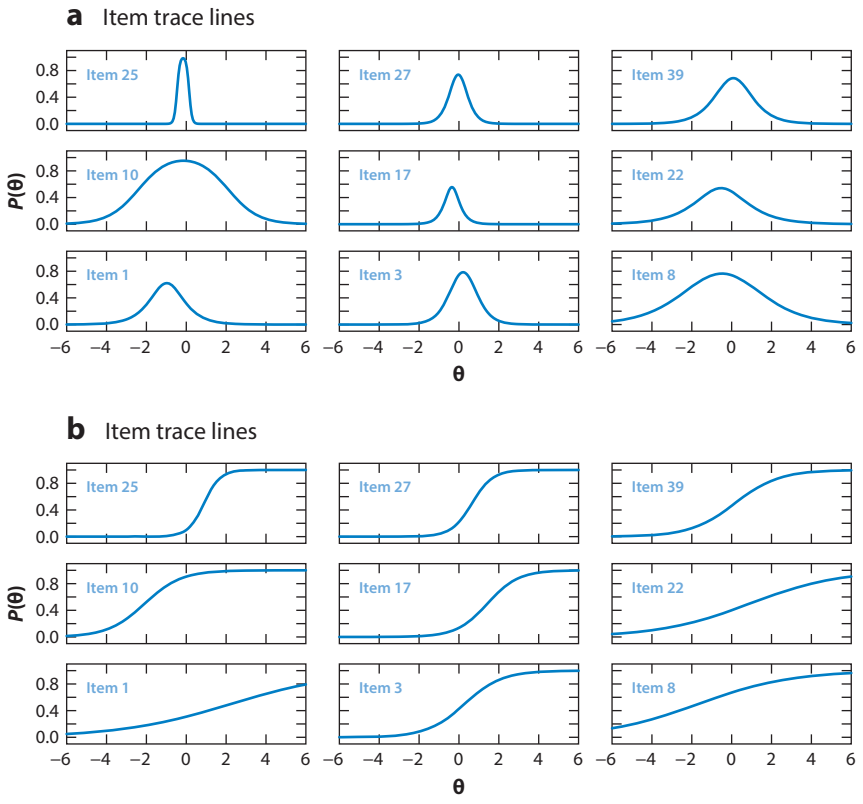

Figure 2

Ideal point (*a*) and graded response (*b*) model item characteristic curves for nine extraversion items from the Eysenck Personality Inventory (Eysenck & Eysenck 1968) in a dataset included in the psych package in R (Revelle 2020). These curves describe the probability (*P*) of a correct response for each item as a function of the standing of the person on the extraversion continuum (θ). The graph was originally generated using the mirt package (Chalmers 2012).

disagree, agree, and strongly agree), followed by the direction of the response (i.e., strongly disagree and disagree versus agree and strongly agree), followed by a magnitude of the response (i.e., strongly or not). These decision points—akin to decision trees—can be examined to understand how individuals choose specific response options while modeling latent trait standing. An attractive characteristic of tree models is that they can be used to systematically study the degree to which decisions in a response process are related to each other. A common expectation in the literature, for instance, is that respondents view Likert scales as a continuous ordinal scale. Tree models, however, suggest that this may frequently not be the case. Empirical research found that different decisions on the same response scale can sometimes be surprisingly independent of each other (Zettler et al. 2016) and have different correlates. For instance, the occurrence of response styles (e.g., extreme responding versus neutral and midpoint responding) has been linked to person characteristics, whereas directional responding (i.e., agree or disagree) is linked to item content. Practically, it has also been shown that the decision on the direction of the response (i.e., agree or disagree) in self-reported personality predicted job performance better than the other decision processes (LaHuis et al. 2019). Tree models can also be used to address a longstanding discussion within personality research on the usefulness of variability traits. For instance, a specific type of

tree model—the trait variability tree model—can be used to extract a situational variability trait from construct-driven situational judgment tests (Lievens et al. 2018) or from personality inventories (Lang et al. 2019).

Other response process item response theory approaches. Thurstonian IRT models have developed from Louis Thurstone's work on scaling (Thurstone 1927) and are based on the idea that respondents compare the utility of different choice alternatives on a latent utility scale. These models have recently been proven useful in scoring forced-choice questionnaires (Brown 2016; also see Drasgow et al. 2010, Maydeu-Olivares & Brown 2010, Stark et al. 2006) and also in recovering latent motive scores from implicit motive measures (Lang 2014, Runge & Lang 2019). Dynamic IRT models (De Boeck et al. 2011, Verhelst & Glas 1993) assume that responses to earlier items affect subsequent items and have been useful in IRT modeling of implicit motives (Lang 2014, Runge & Lang 2019) and have also successfully been utilized to study item-order effects (Debeer & Janssen 2013, Verhelst & Glas 1993, Wang et al. 2013) and systematic missing observations (Debeer et al. 2017).

Further Applications of Item Response Theory in Studying Questionnaires

In closing this section, we highlight additional applications of IRT for understanding questionnaires beyond the scope of this article. For instance, extreme responding or faking has been studied using tree models in the recent literature but there are also other approaches to detect potential faking or unusual responding such as person-fit indices (LaHuis & Copeland 2009, Meijer & Sijtsma 2001). Furthermore, the methods we highlighted in the earlier section on testing/assessment can frequently also be readily used for questionnaires. For instance, researchers can use IRT to model the multidimensional structure of emotions (Tay et al. 2011).

CONSTRUCT VALIDITY/VALIDATION

An important theoretical idea related to the explanatory IRT and the response process IRT approaches discussed earlier is a novel view of test validity (Borsboom & Mellenbergh 2007, Borsboom et al. 2004, De Boeck & Wilson 2004, Doran et al. 2007, Embretson 1998, Wilson 2005). Test validation was long viewed as a complex process, in which a researcher should examine a variety of different forms of validity and especially put the test into a larger framework of relationships to outside criteria (Messick 1989a,b). Psychometricians in the 2000s began to advocate an alternative and more straightforward view of validity. An important starting point for this alternative view was a paper by Borsboom et al. (2004). These authors argued that a test is valid for measuring an attribute "if (*a*) the attribute exists and (*b*) variations in the attribute causally produce variation in the measurement outcomes" (p. 1061). This internal or psychometric definition of validity shifted the focus on test validation from placing a test into a larger nomological network to developing an IRT response model and studying it psychometrically through tools such as explanatory IRT or response process IRT. An important scientific advantage of this revised view of validity is the fact that it is possible to have a valid test that does not predict an outcome of theoretical interest. The core limitation of earlier validity theory with its emphasis on examining a variety of forms of validity and placing a measurement instrument into a larger nomological net of constructs is the fact that a failure of a measurement instrument in showing correlations with an outcome or correlate of theoretical interest also automatically raises questions about the validity of the instrument. In shifting the focus in test validation away from outside criteria and to the inner mechanisms of how responses react to a posited attribute, this modern perspective of validity addresses this fundamental limitation of earlier validity theory.

Several developments in organizational research are related to this modern view of validity. In our earlier discussion of explanatory IRT, we mentioned recent development in selection and assessment (Lievens & Sackett 2017, Sackett et al. 2017) suggesting that researchers should not only focus on selection and measurement procedures as holistic entities and instead try to examine their inner workings.

Recently, researchers have also extended the psychometric view of validity by suggesting shifting the focus in test development. When researchers develop a new test, they typically start by defining a research area and developing items. An alternative approach is to shift the focus around and start with the continuum (Tay & Jebb 2018). In continuum specification, researchers carefully define the construct continua and its operationalizations (e.g., response options). In determining whether construct continua overlap on a continuum, IRT can be applied to examine the extent item locations fall along a common continuum (Wilson 2005). Moreover, because people use ideal point response processes for self-reported typical behaviors (e.g., personality, attitudes, interest, attitudes), relying on ideal point IRT is necessary for such modeling given that there is no ideal point factor analytic comparison (Tay & Ng 2018). Continuum specification seems particularly relevant with the proliferation of organizational constructs in both the light [e.g., character strength (Peterson & Park 2004)] and dark sides [e.g., dark triad (Paulhus & Williams 2002)]. There are questions as to the extent these constructs are continuous or distinct from prior constructs (e.g., normal personality).

The implications of the psychometric view of validity can be illustrated by reconsidering the earlier example of a novel test on sensitive thinking (see **Table 6**). Imagine, the researcher finds that the newly developed sensitive thinking test has no relationship with the outcome of interest. In the traditional framework, a researcher would not know whether the new test does not measure the new construct adequately or whether the new construct just has no relationship with the outcome of interest. In the psychometric view of validity, it is possible that a valid measure of sensitive thinking exists and that the construct simply does not predict leadership outcomes in virtual teams. This example shows that the explanatory item response framework and the psychometric view of validity have some resemblance to the innovative measurement approaches in other areas of science discussed at the start of this article. As in the examples estimating the longitude of ships or indirect measurements of gravity at the start of this article, the sensitive thinking example in **Table 6** uses configurations of responses to make inferences about the way an underlying latent construct causes responses.

MEASUREMENT EQUIVALENCE OF SCORES

Over the decades, organizational psychologists have recognized the importance of measurement equivalence when making group comparisons or comparisons between measurements over time (Vandenberg & Lance 2000). Measurement equivalence especially forms the basis of whether mean-level comparisons between groups represent true latent differences or measurement artifacts [e.g., differences in how measures are interpreted and used (Drasgow 1982)]. Organizational researchers pay close attention to this not only because of concerns about validity of findings but also because of legal ramifications, as measurement bias can unfairly disadvantage certain groups. Although measurement equivalence was popularized in organizational psychology research using factor analytic models (Vandenberg & Lance 2000), measurement equivalence was also widely tested using IRT methods in employment and educational testing (Bock 1997, Drasgow 1987). One major reason for the use of IRT was the dichotomous response options in test items (i.e., right versus wrong). More importantly, many tests have a multiple-choice format that allows for guessing the right answer. For example, with four multiple-choice options, there is a

0.25 probability of obtaining the right answer by chance. The traditional three-parameter logistic IRT model can account for the possibility of guessing on an item by incorporating a lower asymptote in the item response curve (Birnbaum 1968). In general, IRT can enable flexible assessment of measurement invariance for psychological tests and categorical items (Tay et al. 2015). The IRT methods can serve to pinpoint which items lack measurement equivalence [i.e., differential item functioning (DIF)] on such tests so that noninvariant items can be excluded. For example, past research showed that on the SAT, a verbal item containing the term regatta was biased against blacks as compared to whites; for the same level of ability, blacks were less likely to get the item correct (Weiss 1987). These methods have been extended to an IRT-covariate approach so that multiple covariates (e.g., race, gender, age) can be modeled simultaneously to determine the key predictors of DIF (Tay et al. 2011, 2016).

Most advanced IRT approaches can be relatively straightforwardly extended to allow researchers to conduct DIF analyses (e.g., De Boeck 2008, De Boeck et al. 2011 extended to Runge et al. 2019). For instance, earlier in this article, we discussed the LLTMe model. This model includes a random item parameter (De Boeck 2008, De Boeck et al. 2011). Removing the item predictors from the LLTMe leads to the basic random item IRT model. An important assumption in random item IRT models is that test or survey items are assumed to be drawn from a pool of items (and set of item parameters). As applied to measurement invariance, unlike typical measurement invariance approaches where items are assumed to be fixed parameters, random item IRT models can allow random item effects to vary across different groups (e.g., countries), and it is not necessary to establish full measurement invariance for group comparisons (De Boeck 2008, Fox & Verhagen 2010). This is a critical advancement given that there are always concerns about whether anchor items used to assess measurement invariance of other items are themselves invariant (Meade & Wright 2012, Stark et al. 2006). For example, past research found that a substantial number of biodata employment items exhibited race DIF (Whitney & Schmitt 1997), and it is known that a greater proportion of DIF items in a scale can lead to problems in detecting DIF and suitable anchor items (Gierl et al. 2004).

Another extension of common IRT models that is relevant for studying measurement invariance is multilevel IRT models (Doran et al. 2007). Multilevel IRT models are basically standard measurement invariance models with a larger number of groups. Especially when group membership has implications on item functioning, it can make sense to include the multilevel nesting into the IRT model. For instance, Doran et al. (2007) used organizational data from a survey of military personnel and showed that group membership had an impact on parameters in the IRT model. Multilevel IRT models are conceptually related to models that include predictors at the person or at the group level into the IRT model (De Boeck et al. 2011, De Boeck & Wilson 2004, Tay et al. 2016).

CHALLENGES AND LIMITATIONS

Over the course of the past two decades, the field of psychometrics has made significant progress. Nevertheless, several challenges for researchers, who use or are interested in using IRT and especially novel developments in IRT, remain. One challenge for researchers is the fact that it can be hard to figure out how different IRT approaches and software solutions (*a*) differ from each other and (*b*) differ from other analytical frameworks for testing and measurement. Other common analytical frameworks for testing and measurement include classical test theory, structural equation modeling, confirmatory factor analysis, exploratory factor analysis methodology, or multilevel models (generalized linear mixed-effects models). In many cases, different IRT frameworks and

other analytical frameworks are closely related to each other but use different terminology to refer to the same things. This sometimes confusing situation is the result of the fact that researchers have increasingly identified links between different analytical frameworks in the past two decades. On the one hand, the identification of these links has led to a wealth of innovation in IRT modeling and the emergence of new ways to view IRT models and measurement. Users now routinely use software developed for other analytical frameworks to fit their IRT models. On the other hand, this new flexibility can be confusing because the terminology in the theoretical model that researchers seek to estimate and the terminology that the software that he/she uses for the same parameters frequently differ from each other. This situation has made it considerably harder for non-methods experts to navigate this research area and communicate with other researchers such that both parties actually talk about the same thing. Misunderstandings can easily emerge because different terms can refer to the same phenomenon, and, vice versa, similar terms can refer to very different phenomena. Differences in the use of terminology between different social science disciplines may further increase these complexities. For instance, ideal point in political science refers to 1PL/Rasch type models (Bafumi et al. 2005) and not the type of ideal point models used in organizational research discussed earlier in this article. Furthermore, within organizational research, the term unfolding model can be used to refer to ideal point IRT models (Roberts et al. 2000), a process of responding to multidimensional forced-choice items (McCloy et al. 2005), or a theory of voluntary turnover (Lee & Mitchell 1994).

Another challenge for researchers and practitioners who seek to use IRT in their work is to adequately balance model complexity versus parsimony in developing their research models. Issues around overfitting versus underfitting exist in many research areas and even in statistics (Barr et al. 2013, Bates et al. 2015a, Matuschek et al. 2017). Model evaluation in IRT has made progress in recent years (Foster et al. 2017, Nye et al. 2019). However, it can still be hard for researchers to conclusively evaluate model fit and decide between alternative modeling approaches. Most importantly, the fact that an IRT model fits the data well does not automatically guarantee that the parameters that are being estimated are also sufficiently free of bias and accurate. Especially for multidimensional and other complex models, it is frequently advisable that researchers conduct a model parameter recovery study (Brown 2016, Lang 2014, Reise & Yu 1990, Weiss 1982). In a model parameter recovery study, researchers first simulate data with known properties. In the second step, the IRT model is estimated, and the estimated parameters from the IRT model are compared with the true parameters. This procedure is typically repeated many times, and the characteristics of the simulated data can be varied across different conditions to develop a systematic understanding of how well the model recovers the true parameters under the assumption that the model would actually be true. When the procedure suggests that significant bias may exist in the item parameters, person parameters, or reliability estimates, it may be advisable to correct the estimates for bias, or alternatively, to consider a simpler model.

Readers who are interested in developing an understanding of how a parameter recovery study works can use the code (see the **Supplemental Appendix**) we provided to simulate the data used for the analyses in **Tables 4**, **5**, and **6**. It can be insightful, for instance, to alter the parameters in the simulation code and rerun the simulation. It can also be of interest to run the simulation and analyses multiple times to see natural sample variation. Finally, it is possible to compare the squared correlation between the true theta values and the estimated theta values from the models (the real reliability) with the estimated empirical reliability using the code we provided (see the **Supplemental Appendix**). As shown in the **Supplemental Appendix**, these estimated values are close to each other for the models we fitted earlier in this article (see **Tables 4**, **5**, and **6**).

Supplemental Material >

Table 7 Checklist: scientific and practical advantages of using item response theory (IRT)

Checklist question	Key points	Section
Is accurate information on measurement error of individual test scores (not just sample-based) needed?	IRT provides response-pattern-specific estimates of measurement error.	Basic IRT models
Is there interest in the degree to which a measure or item response follows an idealized measurement model?	IRT can be used to study these properties.	Basic IRT models, IRT as a tool for scale development for measurement purists and statisticians
Are many different test forms needed (while having a way to compare people and items)?	IRT makes the development of multiple versions of tests more flexible because item and person parameters can be brought on a common scale for comparison.	Practical: IRT as a technical tool to improve efficiency in large-scale testing
Should theories about what affects responses to the test items be tested?	Explanatory item response models enable researchers to test theories on the role of item features, person predictors, and random error in test items.	IRT as a tool to test theories about test items: explanatory item response modeling
Are unidimensional scores needed, but the test violates the assumption of unidimensionality?	Multidimensional IRT provides a framework to model tests with a complex dimensional structure (e.g., items loading on a main dimension and additionally a narrower dimension, or several equally important dimensions). Multidimensional IRT is efficient (quick estimation with modern algorithms). Multidimensional IRT is quite flexible (can deal with missing data and allows for multiple constraints).	Multidimensional IRT
Should assumptions about the scaling of the response options in rating scales be tested?	IRT can model dichotomous, polytomous, ordered, or categorical responses. IRT can be used to study characteristics of specific response options within rating scales.	IRT as a tool to study the assumptions of rating scales
Should a complex (not simply additive) decision-making process in responding to items be modeled?	Process IRT models such as ideal point models, tree models, Thurstonian IRT models, or dynamic IRT models allow researchers to test theories about how respondents make decisions when they respond to test items that could otherwise not be uncovered.	IRT as a tool to test theories: response process IRT
Is a modern approach to test validation of interest?	The internal or psychometric view of validity focuses on linking variations in attributes to variations in the measurement outcomes—typically by developing an IRT measurement approach.	Construct validity/validation
Are mean differences between groups true latent differences or measurement artifacts?	IRT is a flexible framework for modeling measurement equivalence (e.g., accounting for guessing, incorporating covariates, partial invariance with random items).	Measurement equivalence of scores
Does IRT offer a unique advantage over other modeling frameworks and not merely a re-expression of a comparable model in a particular context/application?	Some IRT models and estimates can be highly similar or even identical to models fitted with tools such as classical test theory, structural equation modeling, confirmatory factor analysis, exploratory factor analysis, or multilevel models.	Challenges and limitations
Is the IRT model adequately balancing complexity versus parsimony?	Although many IRT models are parsimonious, modern software also allows researchers to specify overly complex models potentially. Tools such as fit statistics and parameter recovery studies can help researchers examine the implications of increasing or decreasing model complexity.	Challenges and limitations

SUMMARY AND CHECKLIST

The degree to which different social science fields use IRT models markedly differs. Whereas some fields have enthusiastically embraced IRT methods, other fields have been critical, and organizational researchers typically fall in between these extreme positions. However, IRT is not a homogeneous method, and the usefulness of IRT depends on the context and purpose of the application. When IRT is used as a scale development tool for fixed-item tests, it frequently has few clear benefits over other methods. A traditional strength of IRT is its use as a scale development tool in large-scale testing that makes more flexible item presentation and test development possible. We also highlighted the other strengths and unique features of IRT for theory testing and test validation purposes such as explanatory and process IRT methods.

To highlight how and when IRT can provide benefits for science and practice, we have assembled a checklist. **Table 7** presents this checklist and summarizes the key concepts and ideas in this article. Both researchers and practitioners can linearly go through this checklist to make an informed decision on whether and how IRT can provide benefits in a particular measurement situation.

CONCLUSION

In this article, we put the emphasis on how IRT can contribute something novel to research and practice. We have described a set of emerging views in the literature that emphasize that IRT can not only be understood as a scale development tool but has developed into a solid framework for testing important theories about measurement. These developments may form a foundation for IRT to develop from a niche application for scale development purists and statisticians into a more commonly used research tool in organizational research.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Malte Runge for comments on an earlier version of this article as well as *Annual Review of Organizational Psychology and Organizational Behavior* Editor Frederick Morgeson and Committee Member Filip Lievens for helpful feedback during the review process.

LITERATURE CITED

Ackerman TA. 1989. Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Appl. Psychol. Meas.* 13:113–27

Ashby N. 2003. Relativity in the Global Positioning System. *Living Rev. Relativ.* 6(1):1

Bafumi J, Gelman A, Park DK, Kaplan N. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Anal.* 13(2):171–87

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68:255–78

Bates D, Kliegl R, Vasishth S, Baayen H. 2015a. *Parsimonious mixed models*. arXiv:1506.04967 [stat.ME]

Bates D, Mächler M, Bolker B, Walker S. 2015b. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1). **https://doi.org/10.18637/jss.v067.i01**

Binet A, Simon T. 1916. *New Methods for the Diagnosis of the Intellectual Level of Subnormals. (L'Année Psych., 1905, pp. 191–244)*, transl. A Binet, T Simon, ES Kite, in *The Development of Intelligence in Children (The*

Binet-Simon Scale), pp. 37–90. Philadelphia: Williams & Wilkins Co (from French). **https://doi.org/10.1037/11069-002**

Birnbaum A. 1968. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, ed. FM Lord, MR Novick, pp. 17–20. Reading, MA: Addison-Wesley

Bliese PD, Wright KM, Adler AB, Cabrera O, Castro CA, Hoge CW. 2008. Validating the primary care posttraumatic stress disorder screen and the posttraumatic stress disorder checklist with soldiers returning from combat. *J. Consult. Clin. Psychol.* 76(2):272–81

Bock RD. 1997. A brief history of item theory response. *Educ. Meas. Issues Pract.* 16(4):21–33

Bock RD, Lieberman M. 1970. Fitting a response model for *n* dichotomously scored items. *Psychometrika* 35(2):179–97

Bock RD, Murakl E, Pfeiffenberger W. 1988. Item pool maintenance in the presence of item parameter drift. *J. Educ. Meas.* 25:275–85

Böckenholt U. 2001. Hierarchical modeling of paired comparison data. *Psychol. Methods* 6(1):49–66

Böckenholt U. 2012. Modeling multiple response processes in judgment and choice. *Psychol. Methods* 17:665–78

Böckenholt U, Meiser T. 2017. Response style analysis with threshold and multi-process IRT models: a review and tutorial. *Br. J. Math. Stat. Psychol.* 70:159–81

Borsboom D. 2006. The attack of the psychometricians. *Psychometrika* 71:425–40

Borsboom D, Mellenbergh GJ. 2007. Test validity in cognitive assessment. In *Cognitive Diagnostic Assessment for Education*, ed. J Leighton, M Gierl, pp. 85–116. Cambridge, UK: Cambridge Univ. Press

Borsboom D, Mellenbergh GJ, van Heerden J. 2004. The concept of validity. *Psychol. Rev.* 111:1061–71

Brown A. 2016. Item response models for forced-choice questionnaires: a common framework. *Psychometrika* 81(1):135–60

Cai L. 2010. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *J. Educ. Behav. Stat.* 35(3):307–35

Carter NT, Dalal DK, Boyce AS, O'Connell MS, Kung M-C, Delgado KM. 2014. Uncovering curvilinear relationships between conscientiousness and job performance: how theoretically appropriate measurement makes an empirical difference. *J. Appl. Psychol.* 99(4):564–86

Chalmers RP. 2012. mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48(6). **https://doi.org/10.18637/jss.v048.i06**

Coles P. 2019. Relativity revealed. *Nature* 568(7752):306–7

Davison ML. 1977. On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika* 42(4):523–48

De Boeck P. 2008. Random item IRT models. *Psychometrika* 73:533–59

De Boeck P, Bakker M, Zwitser R, Nivard M, Hofman A, et al. 2011. The estimation of item response models with the lmer function from the lme4 package in R. *J. Stat. Softw.* 39(12). **https://doi.org/10.18637/jss.v039.i12**

De Boeck P, Partchev I. 2012. IRTrees: tree-based item response models of the GLMM family. *J. Stat. Softw.* 48(Code Snippet 1). **https://doi.org/10.18637/jss.v048.c01**

De Boeck P, Wilson M. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer

de la Torre J. 2009. DINA model and parameter estimation: a didactic. *J. Educ. Behav. Stat.* 34:115–30

Debeer D, Janssen R. 2013. Modeling item-position effects within an IRT framework. *J. Educ. Meas.* 50(2):164–85

Debeer D, Janssen R, De Boeck P. 2017. Modeling skipped and not-reached items using IRTrees. *J. Educ. Meas.* 54(3):333–63

DeSimone JA, James LR. 2015. An item analysis of the Conditional Reasoning Test of Aggression. *J. Appl. Psychol.* 100(6):1872–86

Doran H, Bates D, Bliese P, Dowling M. 2007. Estimating the multilevel Rasch model: with the lme4 package. *J. Stat. Softw.* 20(2). **https://doi.org/10.18637/jss.v020.i02**

Drasgow F. 1982. Biased test items and differential validity. *Psychol. Bull.* 92(2):526–31

Drasgow F. 1987. Study of the measurement bias of two standardized psychological tests. *J. Appl. Psychol.* 72(1):19–29

Drasgow F, Chernyshenko OS, Stark S. 2010. Improving the measurement of psychological variables: Ideal point models rock! *Ind. Organ. Psychol.* 3:515–20
Drasgow F, Hulin CL. 1990. Item response theory. In *Handbook of Industrial and Organizational Psychology*, ed. MD Dunnette, LM Hough, pp. 577–636. Palo Alto, CA: Consult. Psychol. Press
Embretson SE. 1998. A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol. Methods* 3:380–96
Eysenck HJ, Eysenck SBG. 1968. *Manual for the Eysenck Personality Inventory*. San Diego, CA: Educ. Ind. Test. Serv.
Fischer GH. 1973. The linear logistic test model as an instrument in educational research. *Acta Psychol.* 37(6):359–74
Foster GC, Min H, Zickar MJ. 2017. Review of item response theory practices in organizational research. *Organ. Res. Methods* 20:465–86
Fox J-P, Verhagen AJ. 2010. Random item effects modeling for cross-national survey data. In *Cross-Cultural Analysis: Methods and Applications*, ed. E Davidov, P Schmidt, J Billiet, pp. 461–82. New York: Routledge
Gierl MJ, Gotzmann A, Boughton KA. 2004. Performance of SIBTEST when the percentage of DIF items is large. *Appl. Meas. Educ.* 17(3):241–64
Greenwald AG. 2012. There is nothing so theoretical as a good method. *Perspect. Psychol. Sci.* 7(2):99–108
Gulliksen H. 1950. *Theory of Mental Tests*. New York: Wiley
Hambleton RK, Swaminathan H, Rogers HJ. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage
Hernández A, Drasgow F, González-Romá V. 2004. Investigating the functioning of a middle category by means of a mixed-measurement model. *J. Appl. Psychol.* 89:687–99
Hornke LF. 2002. Item-generation models for higher-order cognitive functions. In *Item Generation and Test Development*, ed. SH Irvine, PC Kyllonen, pp. 159–78. Mahwah, NJ: Erlbaum
Irtel H. 1995. An extension of the concept of specific objectivity. *Psychometrika* 60(1):115–18
Johnston AK, Connor RD, Stephens CE, Ceruzzi PE. 2015. *Time and Navigation: The Untold Story of Getting from Here to There*. Washington, DC: Smithson. Books
Kim S. 2012. A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika* 77(1):153–62
Kim S-H, Cohen AS. 1998. A comparison of linking and concurrent calibration under item response theory. *Appl. Psychol. Meas.* 22:131–43
Kolen MJ, Brennan RL. 2004. *Test Equating, Scaling, and Linking: Methods and Practice*. New York: Springer. 2nd ed.
Kubinger KD. 2009. Applications of the linear logistic test model in psychometric research. *Educ. Psychol. Meas.* 69(2):232–44
LaHuis DM, Blackmore CE, Bryant-Lees KB, Delgado K. 2019. Applying item response trees to personality data in the selection context. *Organ. Res. Methods* 22(4):1007–18
LaHuis DM, Copeland D. 2009. Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organ. Res. Methods* 12(2):296–319
Lang JWB. 2014. A dynamic Thurstonian item response theory of motive expression in the picture story exercise: solving the internal consistency paradox of the PSE. *Psychol. Rev.* 121:481–500
Lang JWB, Lievens F, De Fruyt F, Zettler I, Tackett JL. 2019. Assessing meaningful within-person variability in Likert-scale rated personality descriptions: an IRT tree approach. *Psychol. Assess.* 31(4):474–87
Lee TW, Mitchell TR. 1994. An alternative approach: the unfolding model of voluntary employee turnover. *Acad. Manag. Rev.* 19(1):51–89
Lievens F. 2017. Construct-driven SJTs: toward an agenda for future research. *Int. J. Test.* 17(3):269–76
Lievens F, Burke E. 2011. Dealing with the threats inherent in unproctored Internet testing of cognitive ability: results from a large-scale operational test program. *J. Occup. Organ. Psychol.* 84(4):817–24
Lievens F, Lang JWB, De Fruyt F, Corstjens J, Van De Vijver M, Bledow R. 2018. The predictive power of people's intra-individual variability across situations: implementing whole trait theory in assessment. *J. Appl. Psychol.* 103:753–71
Lievens F, Sackett PR. 2017. The effects of predictor method factors on selection outcomes: a modular approach to personnel selection procedures. *J. Appl. Psychol.* 102:43–66

Lord FM. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Erlbaum
Lord FM, Novick MR. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley
Lord RG, Kanfer R. 2002. Emotions and organizational behavior. In *Emotions in the Workplace: Understanding the Structure and Role of Emotions in Organizational Behavior*, ed. RG Lord, RJ Klimoski, R Kanfer, pp. 5–19. San Francisco: Jossey-Bass
Marsh HW, Hau K-T. 2007. Applications of latent-variable models in educational psychology: the need for methodological-substantive synergies. *Contemp. Educ. Psychol.* 32(1):151–70
Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. 2017. Balancing Type I error and power in linear mixed models. *J. Mem. Lang.* 94:305–15
Maydeu-Olivares A, Brown A. 2010. Item response modeling of paired comparison and ranking data. *Multivar. Behav. Res.* 45:935–74
McClimans L, Browne J, Cano S. 2017. Clinical outcome measurement: models, theory, psychometrics and practice. *Stud. Hist. Philos. Sci. A* 65–66:67–73
McCloy RA, Heggestad ED, Reeve CL. 2005. A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. *Organ. Res. Methods* 8(2):222–48
McFadden D. 2001. Economic choices. *Am. Econ. Rev.* 91:351–78
McHorney CA, Monahan PO. 2004. Postscript: applications of Rasch analysis in health care. *Med. Care* 42(Suppl.):1–73
Meade AW, Wright NA. 2012. Solving the measurement invariance anchor item problem in item response theory. *J. Appl. Psychol.* 97(5):1016–31
Meijer RR, Sijtsma K. 2001. Methodology review: evaluating person fit. *Appl. Psychol. Meas.* 25:107–35
Messick S. 1989a. Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18(2):5–11
Messick S. 1989b. Validity. In *Educational Measurement*, ed. RL Linn, pp. 13–103. New York: Am. Counc. Educ./Macmillan Publ.
Michell J. 2015. Measurement theory: history and philosophy. In *International Encyclopedia of the Social & Behavioral Sciences*, ed. JD Wright, pp. 868–72. Amsterdam: Elsevier. 2nd ed.
Mitchell DJ, Tal E, Chang H. 2017. The making of measurement: Editors' introduction. *Stud. Hist. Philos. Sci. A* 65–66:1–7
Muthén LK, Muthén BO. 2015. *Mplus User's Guide*. Los Angeles: Muthén & Muthén. 7th ed.
Nye CD, Joo S-H, Zhang B, Stark S. 2019. Advancing and evaluating IRT model data fit indices in organizational research. *Organ. Res. Methods* 23:457–86
Oswald FL, Shaw A, Farmer WL. 2015. Comparing simple scoring with IRT scoring of personality measures. *Appl. Psychol. Meas.* 39:144–54
Paulhus DL, Williams KM. 2002. The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *J. Res. Personal.* 36(6):556–63
Peterson C, Park N. 2004. Classification and measurement of character strengths: implications for practice. In *Positive Psychology in Practice*, ed. PA Linley, S Joseph, pp. 433–46. Hoboken, NJ: Wiley
Plieninger H. 2020. Developing and applying IR-Tree models: guidelines, caveats, and an extension to multiple groups. *Organ. Res. Methods.* In press. **https://doi.org/10.1177/1094428120911096**
Podsakoff PM, Organ DW. 1986. Self-reports in organizational research: problems and prospects. *J. Manag.* 12(4):531–44
Reckase MD. 2009. *Multidimensional Item Response Theory*. New York: Springer
Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, et al. 2007. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care* 45(5):S22–31
Reise SP, Yu J. 1990. Parameter recovery in the graded response model using MULTILOG. *J. Educ. Meas.* 27(2):133–44
Revelle W. 2020. psych: procedures for psychological, psychometric, and personality research. *R Package*, Version 2.07. **https://CRAN.R-project.org/package=psych**
Revelle W, Zinbarg RE. 2009. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika* 74(1):145–54

Rizopoulos D. 2006. ltm: an R package for latent variable modeling and item response theory analyses. *J. Stat. Softw.* 17(5). **https://doi.org/10.18637/jss.v017.i05**

Roberts JS. 2001. GGUM2000: estimation of parameters in the generalized graded unfolding model. *Appl. Psychol. Meas.* 25(1):38

Roberts JS, Donoghue JR, Laughlin JE. 2000. A general item response theory model for unfolding unidimensional polytomous responses. *Appl. Psychol. Meas.* 24(1):3–32

Robitzsch A. 2020. sirt: supplementary item response theory models. *R Package*, *Version 3.9-4*. **https://cran.r-project.org/package=sirt**

Rogelberg SG, Church AH, Waclawski J, Stanton JM. 2008. Organizational survey research. In *Handbook of Research Methods in Industrial and Organizational Psychology*, ed. SG Rogelberg, pp. 140–60. Hoboken, NJ: Blackwell Publ. **https://doi.org/10.1002/9780470756669.ch7**

Rosseel Y. 2012. lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48(2). **https://doi.org/10.18637/jss.v048.i02**

Runge JM, Lang JWB. 2019. Can people recognize their implicit thoughts? The motive self-categorization test. *Psychol. Assess.* 31(7):939–51

Runge JM, Lang JWB, Chasiotis A, Hofer J. 2019. Improving the assessment of implicit motives using IRT: cultural differences and differential item functioning. *J. Personal. Assess.* 101(4):414–24

Sackett PR, Lievens F, Van Iddekinge CH, Kuncel NR. 2017. Individual differences and their measurement: a review of 100 years of research. *J. Appl. Psychol.* 102(3):254–73

Sen S, Bradshaw L. 2017. Comparison of relative fit indices for diagnostic model selection. *Appl. Psychol. Meas.* 41(6):422–38

Sijtsma K. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74(1):107–20

Smith AG, Burns TM. 2014. Reevaluating clinical measurement tools in therapeutic trials: Time to make a Rasch decision? *Neurology* 83(23):2104–5

Smith PC, Kendall L, Hulin CL. 1969. *The Measurement of Satisfaction in Work and Retirement: A Strategy for the Study of Attitudes*. Chicago: Rand McNally

Sorrel MA, Olea J, Abad FJ, de la Torre J, Aguado D, Lievens F. 2016. Validity and reliability of situational judgement test scores. *Organ. Res. Methods* 19(3):506–32

Stark S, Chernyshenko OS, Drasgow F, Williams BA. 2006. Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *J. Appl. Psychol.* 91(1):25–39

Tal E. 2017. Measurement in science. In *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Stanford, CA: Metaphys. Res. Lab., Stanford Cent. Stud. Lang. Inf. Fall 2017 Ed. **https://plato.stanford.edu/archives/fall2017/entries/measurement-science/**

Tay L, Diener E, Drasgow F, Vermunt JK. 2011. Multilevel mixed-measurement IRT analysis: an explication and application to self-reported emotions across the world. *Organ. Res. Methods* 14(1):177–207

Tay L, Drasgow F, Rounds J, Williams BA. 2009. Fitting measurement models to vocational interest data: Are dominance models ideal? *J. Appl. Psychol.* 94(5):1287–304

Tay L, Huang Q, Vermunt JK. 2016. Item response theory with covariates (IRT-C): assessing item recovery and differential item functioning for the three-parameter logistic model. *Educ. Psychol. Meas.* 76(1):22–42

Tay L, Jebb AT. 2018. Establishing construct continua in construct validation: the process of continuum specification. *Adv. Methods Pract. Psychol. Sci.* 1(3):375–88

Tay L, Kuykendall L. 2017. Why self-reports of happiness and sadness may not necessarily contradict bipolarity: a psychometric review and proposal. *Emot. Rev.* 9(2):146–54

Tay L, Meade AW, Cao M. 2015. An overview and practical guide to IRT measurement equivalence analysis. *Organ. Res. Methods* 18:3–46

Tay L, Newman DA, Vermunt JK. 2011. Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organ. Res. Methods* 14(1):147–76

Tay L, Ng V. 2018. Ideal point modeling of non-cognitive constructs: review and recommendations for research. *Front. Psychol.* 9. **https://doi.org/10.3389/fpsyg.2018.02423**

Terman LM. 1916. *The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale*. Cambridge, MA: Houghton Mifflin

Thissen D, Reeve BB, Bjorner JB, Chang C-H. 2007. Methodological issues for building item banks and computerized adaptive scales. *Qual. Life Res.* 16(S1):109–19

Thurstone LL. 1927. A law of comparative judgment. *Psychol. Rev.* 34(4):273–86

Thurstone LL. 1928. Attitudes can be measured. *Am. J. Sociol.* 33(4):529–54

van der Linden W, Hambleton RK, eds. 1997. *Handbook of Modern Item Response Theory*. New York: Springer

van Rijn PW, Rijmen F. 2012. A note on explaining away and paradoxical results in multidimensional item response theory. *ETS Res. Rep. Ser.* 2012(2):i–10

Vandenberg RJ, Lance CE. 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3(1):4–70

Verhelst ND, Glas CAW. 1993. A dynamic generalization of the Rasch model. *Psychometrika* 58(3):395–415

Wainer H, Bradlow ET, Wang X. 2007. *Testlet Response Theory and Its Applications*. Cambridge, UK: Cambridge Univ. Press

Wang X, Berger JO, Burdick DS. 2013. Bayesian analysis of dynamic item response models in educational testing. *Ann. Appl. Stat.* 7(1):126–53

Weiss DJ. 1982. Improving measurement quality and efficiency with adaptive testing. *Appl. Psychol. Meas.* 6(4):473–92

Weiss DJ. 2004. Computerized adaptive testing for effective and efficient measurement in counseling and education. *Meas. Eval. Couns. Dev.* 37(2):70–84

Weiss J. 1987. The Golden Rule bias reduction principle: a practical reform. *Educ. Meas. Issues Pract.* 6(2):23–25

Whitney DJ, Schmitt N. 1997. Relationship between culture and responses to biodata employment items. *J. Appl. Psychol.* 82(1):113–29

Wilson MR. 2005. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum

Wilson MR, De Boeck P, Carstensen CH. 2008. Explanatory item response models. In *Assessment of Competencies in Educational Contexts*, ed. J Hartig, E Klieme, D Leutner, pp. 83–110. Göttingen, Ger.: Hogrefe & Huber

Yerkes RM. 1917. The Binet versus the Point Scale method of measuring intelligence. *J. Appl. Psychol.* 1(2):111–22

Yoakum CS, Yerkes RM. 1920. *Army Mental Tests*. New York: Henry Holt

Zettler I, Lang JWB, Hülsheger UR, Hilbig BE. 2016. Dissociating indifferent, directional, and extreme responding in personality data: applying the three-process model to self- and observer reports. *J. Personal.* 84:461–72

Zickar MJ. 1998. Modeling item-level data with item response theory. *Curr. Dir. Psychol. Sci.* 7(4):104–9