American Psychological Association
DIVISION 10
Psychology of Aesthetics, Creativity, and the Arts

© 2019 American Psychological Association
ISSN: 1931-3896
http://dx.doi.org/10.1037/aca0000287

Development of the R Library "jrt": Automated Item Response Theory Procedures for Judgment Data and Their Application With the Consensual Assessment Technique

Nils Myszkowski
Pace University

Although the Consensual Assessment Technique (CAT; Amabile, 1982) is considered a gold standard in the measurement of product attributes, including creativity (Baer & McKool, 2009), considerations on how to improve its scoring and psychometric modeling are rare. Recently, it was advanced (Myszkowski & Storme, 2019) that the framework of Item Response Theory (IRT) is appropriate for CAT data and would provide several practical and conceptual benefits to both the psychometric investigation of the CAT and the scoring of creativity. However, the packages recommended for IRT modeling of ordinal data are hardly accessible for researchers nonfamiliar with IRT and offer minimal possibility for adaptation of outputs to judgment data. Thus, the package "jrt" was developed for the open source programming language R and is available on the Comprehensive R-Archive Network (CRAN). Its main aim is to make IRT analyses easily applicable to CAT data by automating model selections, diagnosing and dealing with issues related to model-data incompatibilities; providing quick, customizable, and publication-ready outputs for communication; and guiding researchers new to IRT through the different available methods. We provide brief tutorials and examples for the main functions, which are further detailed in the online vignette and documentation on CRAN. We finally discuss the current limitations and anticipated extensions of the jrt package and invite researchers to take advantage of its practicality.

Keywords: item response theory, Consensual Assessment Technique, creativity judgment, creativity assessment, measurement

A considerable range of methods have been developed to assess creativity, yet the Consensual Assessment Technique (CAT; Amabile, 1982) is often considered a gold standard to measure product creativity because it is both widely used and heavily discussed in creativity research (e.g., Baer, Kaufman, & Gentile, 2004; Baer & McKool, 2009; Kaufman, Baer, & Cole, 2009). The CAT allows the measure of product creativity—as well as its facets (e.g., product novelty, originality, usefulness, elaboration, aesthetic quality)—by collecting ratings from several judges. Beyond its flexibility in the attribute being measured, a reason for the success of the CAT is that it does not rely on a theory of creative thinking for its validity (Baer et al., 2004); it essentially relies on the assumption that the attribute can be validly rated by judges. This assumption is based on the use of expert judges—or novice judges that are shown to match expert judges—to rate the products (Baer, 2008).

Recently, Myszkowski and Storme (2019) discussed how CAT research could reap substantial benefits from using an approach of judgment data based on Item Response Theory (IRT), as opposed to more traditional approaches (such as the traditionally used Cronbach's α and sum/average scoring). While that paper discussed the shortcomings of these traditional approaches and the advantages of the IRT framework over them, the present paper—which could be considered its companion paper—acknowledges that a central reason for the little use of IRT in CAT research is the lack of user-friendly software focused on its application to CAT data. This paper also presents a statistical package that aims to address this issue.

The Limitations of Classical Test Theory-Based Procedures

While there have been substantial debates surrounding the data collection stage, especially regarding the required degree of expertise of judges (e.g., Kaufman et al., 2009; Kaufman, Baer, Cole, & Sexton, 2008) and the prospect of efficiently training novice raters (e.g., Storme, Myszkowski, Çelik, & Lubart, 2014), much less work has questioned later stages of measurement: the statistical procedures used to score creativity and to analyze the reliability of the ratings. Traditionally, sum/average scores are used to quantify the attribute being measured, while Cronbach's α—and in some cases, intraclass correlation coefficients—are used to assess reliability (see Cseh & Jeffries, 2019 for a review and description of the different indices). This statistical treatment mimics the

Correspondence concerning this article should be addressed to Nils Myszkowski, Department of Psychology, Pace University, Room 1315, 41 Park Row, New York, NY 10038. E-mail: nmyszkowski@pace.edu

traditional set of analyses often executed with data obtained through psychometric tests, except for the judges replacing the items. Yet, however popular, this set of tools rooted in classical test theory (CTT) has been recently pointed out as limited in many aspects—in creativity research (Myszkowski & Storme, 2019) as well as in psychological measurement in general (Borsboom, 2006; Borsboom & Mellenbergh, 2002). For CAT data, it was noted (Myszkowski & Storme, 2019) that using Cronbach's α as a reliability measure and sum scores as trait estimates formulates assumptions—notably unidimensionality and τ-equivalence—that are both potentially unrealistic and rarely tested, in addition to overlooking important extensions of CTT, such as generalizability theory (Brennan, 1992) and factor analysis.

It was thus proposed that researchers use the framework and set of analyses of Judge Response Theory (JRT; Myszkowski & Storme, 2019), an adaptation of IRT models to multiple judges (instead of multiple items) situations. JRT is defined as "a psychometrical framework that simultaneously uses latent characteristics—trait(s) and/or class(es)—of the judged product or idea (creativity, originality, novelty, etc.) and of the judge (severity, discrimination, etc.) as predictors of observed judgements" (Myszkowski & Storme, 2019, p. 170). In comparison with a traditional set of analyses consisting of using sum/average scoring and Cronbach's α, the JRT approach allows researchers to investigate the variability between judges and account for such variability in the scoring of the product attribute. Notably, polytomous IRT/JRT models can account for the possibility that judges may vary in their severity, in their ability to discriminate between products—in other words, their expertise—as well as in their tendency to use the ordinal response scale in various ways (e.g., preferring central modalities or extreme ones). Although a thorough description of the large (and growing) number of polytomous IRT models is impossible and beyond the scope of this paper, for researchers interested in the specificities of such models, we could point to resources such as Thissen and Steinberg's (1986) taxonomy of polytomous IRT models, Ostini and Nering's (2006) textbook on the subject of polytomous IRT, and the most current general IRT textbooks (e.g., de Ayala, 2013; van der Linden, 2016).

Why IRT Is Underused in CAT

With some exceptions (Akbari Chermahini, Hickendorff, & Hommel, 2012; Barbot, Tan, Randi, Santa-Donato, & Grigorenko, 2012; Forthmann, Celik, Holling, Storme, & Lubart, 2018; Forthmann et al., 2016; Myszkowski, 2019; Myszkowski & Storme, 2017; Sen, 2016; Silvia et al., 2008; Wang, Ho, Cheng, & Cheng, 2014), IRT procedures are rarely used in creativity research. Although the reason for this is unclear, it was advanced (Myszkowski & Storme, 2019) that a central reason behind the underuse of such models in creativity research, like in psychological research in general, is that training in IRT is rarely found in psychology education and that IRT modeling software is often less easily available than traditional CTT applications (Borsboom, 2006). However, as we will briefly discuss here, there are several ways that one can use IRT modeling.

Using IRT Software

The quantity and quality of software available directly to IRT models is rapidly growing, particularly for the statistical environment R (Chalmers, 2012; George & Robitzsch, 2015; Liu & Chalmers, 2018; Mair & Hatzinger, 2007; Robitzsch, Kiefer, & Wu, 2017). However, these software packages remain primarily targeted for educational measurement psychometricians, who are already versed in both IRT and programming. One indication of this is that most IRT packages do not guide the researcher as to what model to use and require the researcher to first identify the available models and then write several lines of code to fit and compare models. Further, some models may have to be manually tweaked—through freeing or constraining parameters or manually specifying response models—to be fitted. For example, the package "mirt" (Chalmers, 2012) technically allows for fitting generalized rating scale models, but it (currently) requires users to fit a generalized partial credit model where category threshold parameters are constrained to be equal across items/judges, while location parameters are freed, which has to be coded by the end-user. While this is a manageable procedure for researchers used to IRT and R, it can be discouraging for others.

In addition, IRT packages often do not offer to compute psychometric information that does not belong in the IRT literature or tradition, such as intraclass correlation coefficients, Cronbach's α, or Fleiss' κ. Meanwhile, to comply with publication standards or demands, researchers often have to use and report on such methods; thus, researchers interested in using IRT models still have to use other packages. Therefore, IRT modeling procedures are often not used *in lieu of*, but *in addition to* other procedures.

Finally, another inconvenience is that IRT software packages are primarily conceptualized for multiple items (not judges); thus, most plots of IRT packages must be tweaked—through additional coding or manual modification—by CAT researchers who would want to rename "items" to "judges" (or "raters," "experts," etc.) for publication or presentation. This point could appear inconsequential, but plotting category curves and information functions is in fact seen as "essential for interpretation of the parameters of polytomous item response models" (Muraki, 1993, p. 354). This implies that, for publication, users not familiar with IRT would *also* need to write their own plotting functions, to ignore this issue (which may require additional justifications in text), or to tweak output plots manually by using graphics software or the plotting syntax, if accessible.

In sum, IRT software packages are overall very well developed and quite versatile, but (a) the learning curve to run and interpret IRT models remains steep without adequate training; (b) IRT software rarely provides non-IRT information (from the interrater reliability literature or CTT literature), although often expected for publication and presentation; and (c) IRT software rarely allows for adapting outputs for CAT contexts. While it is clear that these inconveniences and limitations do not *prevent* creativity for researchers versed in programming and IRT models, the minimal presence of IRT in CAT research clearly demonstrates that these inconveniences still *discourage* many creativity researchers.

Using Alternative Frameworks for IRT/JRT

Popular alternative modeling frameworks may be used for certain IRT/JRT applications (van der Linden, 2016). For example, the framework of generalized linear mixed modeling (GLMM) and its software packages, which are more popular than IRT among most researchers, may be used for IRT/JRT (e.g., Baghaei &

Doebler, 2018). However, under this framework, the availability of response models is limited to models with parallel measurements and Rasch models (Rabe-Hesketh & Skrondal, 2016). Yet in both Rasch models and parallel measurement models, items/judges are not allowed to differ in discrimination/accuracy, and this is especially problematic for the CAT. Indeed, assuming that all judges are equally accurate may be unrealistic, as there often exists a possibility that the panel of judges is heterogeneous in expertise, prior training, attention, or understanding of the judgment instructions.

Similarly, and perhaps more notoriously, the framework of Structural Equation Modeling (SEM), which stems directly from the factor analytic and linear modeling tradition, and its associated software packages are very popular among many psychology researchers. The models allow them to fit item/judge response models with both a varying difficulty/severity parameter and a varying discrimination/accuracy parameter—in the linear framework, typically called intercepts/means and slopes/loadings, respectively (for a discussion of their communalities and differences see Mellenbergh, 1994). However, although a growing number of SEM packages permit fitting models appropriate for ordinal data, the availability of ordinal response models that general SEM packages propose is often limited to few options compared with IRT packages (Chalmers, 2012; Rizopoulos, 2006) and presents limited plotting possibilities.

Thus, both the GLMM and SEM traditions offer possibilities to estimate IRT models, but the choice of models is often limited, leading researchers to make additional, sometimes unrealistic, and often untested, assumptions. Certainly, multiple bridges between statistical frameworks advance the possibilities of fitting IRT models with non-IRT-focused statistical packages; for example, the increased flexibility in SEM-dedicated packages to fit categorical/ordinal responses (generalized SEM) and the development of the generalized linear latent and mixed modeling (GLAMM; Rabe-Hesketh, Skrondal, & Pickles, 2004). Still, the ease of use of these frameworks for studying item/judge characteristics and scoring of individuals/products—two fundamental interests in creativity research and CAT—is often limited and typically necessitates several additional steps and transformations.

The Aim of This Paper

As explained earlier, recent research has indicated that IRT modeling could, in multiple aspects, provide substantial improvements for research using the CAT. Yet we also noted that the current state of IRT-capable packages, whether directly dedicated to IRT or not, does not offer an easy way to perform IRT/JRT analyses and scoring for creativity researchers. In this paper, we discuss the development of a package ("jrt") that combines user-friendly tools for CAT researchers who may be interested in applying IRT/JRT to score and analyze judgment data. Through this development project, our aim is not to provide a replacement of current IRT software but instead to increase the availability and usability of IRT modeling and scoring for creativity researchers, while triggering interest and discussion around IRT applications to CAT data.

The "jrt" Package

The jrt package was created with several primary objectives. First, it aims at maximizing the availability of IRT analysis for CAT researchers. Second, it was developed for use with minimal knowledge of statistical programming. Third, its functions and defaults correspond to typical CAT data situations and needs, with a focus on scoring, reliability, and judge analysis. Fourth, it helps researchers decide between IRT models and make other modeling decisions. Fifth, although it provides defaults, it is sufficiently customizable to adapt to a range of needs and situations. Sixth, its procedures are transparent and didactic. Seventh, it automatically deals with typical issues and software errors arising from the combination of IRT modeling and CAT data—notably, the occurrence of judges with missing response categories. Eighth, it provides publication-ready plots with minimal-to-no fine-tuning necessary. Ninth, it aims at facilitating a CAT researcher's transition to the IRT framework by also allowing for the computation of psychometric information based on the non-IRT (CTT, interrater reliability) literature. Finally, it points researchers to relevant citations for the procedures used, further facilitating publication. In this section, we discuss the choices made in the development of jrt and how they relate to these objectives.

Availability

The jrt package was developed for the statistical programming language R. Although SPSS is certainly more popular currently, its functionalities for IRT modeling remain very limited. In addition, the fact that SPSS is not free limits the availability of jrt, especially for researchers without institutional access to SPSS. In contrast, R is largely favorable to open-source development, and several packages (for IRT estimation, general psychometrics and plotting) already exist that could be used as estimation engines for jrt.

Minimal Programming Skills Necessary

As is explained in the tutorial part of this paper and in the R package vignette, we attempted to minimize the programming skills necessary to fit IRT models as much as possible. Provided that the data set is properly loaded in R and that the jrt package is loaded, one (relatively) simple command will identify the response categories and make decisions if a judge has unobserved categories (which is discussed later), automatically fit IRT models appropriate for ordinal data, compare models, select the best fitting model, report IRT (and non-IRT) reliability estimates, plot all category curves on a graph, and plot the total reliability function plot. This information is output in the console, but it can also be retrieved in a jrt-type object. An additional line of code can be used to retrieve the θ estimates (factor scores) with the standard errors of measurement, appended or not to the original data. As recommended in the IRT literature (e.g., Muraki, 1993), additional plotting functions are provided for more advanced customizations of the judge category curve plots and judge/total information plots (which are used for information, standard errors of measurement, and reliability, three related concepts in IRT).

Availability of Ordinal Models

Most (if not all) research efforts that use the CAT collect judgments at the ordinal, Likert-type level. Even though there are comprehensive taxonomies of polytomous IRT models (Thissen & Steinberg, 1986) and thorough discussions of their specificities (van der Linden, 2016), one objective of jrt is to enable CAT researchers to easily fit a variety of models. Thanks to the mirt estimation package, jrt can currently fit the most popular (Edelen & Reeve, 2007) ordinal/polytomous models: the Rasch rating scale model (Andrich, 1978), the graded rating scale model (Muraki, 1990), the partial credit model (Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the graded response model (GRM; Samejima, 1969). In addition to these models directly available in mirt, jrt also automatically creates parameter constraints to fit additional models; jrt currently fits variants of the GRM with equal item/judge discriminations, which is called the constrained graded response model in jrt, with both equal item/judge discriminations and identical category structures across judges/items, which is called the constrained graded rating scale model in jrt. It also fits a variant of the GPCM with identical category structures across judges/items, often called (including in jrt) the generalized rating scale model. Even though this current eight-model set does not represent every polytomous model possible, it allows one to fit both flexible models that make minimal assumptions about the data (the GPCM and GRM) and more constrained models that assume judges have the same category structure and/or discrimination, which in some cases may be reasonable (or at least testable) constraints. The outputs of the functions of jrt include model comparisons, reliability of the selected model, plots, and factor scores (with conditional standard error estimates). This allows, in one function, to execute a typical analysis and reporting strategy consisting of (a) selecting an appropriate response model, which can be decided a priori or can be done by model comparison, (b) reporting on reliability, (c) discussing diagnostic plots (information/reliability/category curve plots), and (d) exporting/using factor score estimates in subsequent analysis.

Automatic Model Selection

As can be seen in the tutorial, in jrt, researchers can decide manually on the IRT model they would like to fit through an argument in the main function. However, by default, if no model is provided, jrt will compare all models available (listed earlier) based on their fit to the data. Because not all models are nested, this model comparison is based on information criteria (for which a smaller value indicates a better fit); the popular corrected Akaike Information Criterion (AICc) is used by default, but other criteria may be used—namely the Akaike information criterion, Bayesian information criterion, and sample-adjusted Bayesian information criterion. To help with reporting and to make the procedure transparent, jrt outputs the values of the criterion chosen for all models before selecting the best fitting model. In addition to this, as suggested by one of the reviewers, model weights—which are often referred to as Akaike weights, since they are generally applied with AICs—computed from the selected information criterion are reported. The weight of a model ranges between 0 and 1. A value close to 1 indicates that the model fits substantially better than all the other models tested and can be interpreted as the probability that the model is the "best model, given the dataset and the set of candidate models" (Wagenmakers & Farrell, 2004, p. 194). Thus, model weights can be used to compare the relative merits of the tested models and to measure the confidence one can have in favor of the selected model in comparison with the others.

Customization Options

We later discuss how to use the most important customization options, or those that we anticipate will be the most used. Although more information is included in the package documentation, we can note that optional arguments allow for change to the estimation algorithm, convergence threshold, number of iterations, estimation method for factor scores, and number of iterations. It can also append factor scores to the original dataset, provide non-IRT reliability statistics, override the automatic detection of response categories, decide on the behavior to adopt in case of unobserved categories among judges, change "judge" to another term ("expert," "rater," etc.) in outputs, and use themes and colors for plots.

Didactic Options

Purposely, jrt is a very verbose package. Throughout all the stages (detection of response categories, model selection, scoring, etc.), jrt will, by default, output console messages indicating the procedure it is completing. These messages may be turned off selectively through function arguments but can also be completely silenced with the "silent = TRUE" argument. In addition, when encountering some issues (e.g., if the model did not converge), jrt will provide advice (as console outputs) as to what choices may be made to overcome them. Finally, as suggested by a reviewer, researchers interested in how the models are specified to the estimation package can see the mirt function calls with the argument "show.calls = TRUE."

Automatic Troubleshooting of Typical Modeling Issues

Using IRT modeling on CAT data is rarely executed (Myszkowski & Storme, 2019); thus, there may be issues in such applications that are not yet anticipated. However, at this stage, a few may be anticipated, and we dedicated specific attention to offering solutions to them directly in jrt. The most notable is certainly that different polytomous IRT models have different requirements. More specifically, models that constrain the response category structure to be identical across items/judges (namely, the rating scale model, generalized rating scale model, graded rating scale model, and constrained graded rating scale model) cannot be estimated when at least one item/judge does not have all response categories observed, which is likely to happen with small samples of judges. Such cases are automatically detected prior to analysis, and the choice is given to the user to either remove judges with unobserved categories prior to fitting or to not fit the rating scale models (default). In addition, and because it is a frequent situation for judgment data, jrt allows for dealing with missing observations. If missing data are detected, the user is warned with the count and percentage of missing observations, after which the models are fit on the incomplete data. The factor scores are estimated from the selected model, and these factor scores are used to compute plausible values (Mislevy, 1991) for the incomplete data. The full data, completed with the imputed plausible values, is also made easily available in the output object.

Providing Publication-Ready Plots

As was mentioned earlier, IRT software often provides outputs that are not necessarily easy to modify. Yet plots are very central to the understanding and communication of IRT models (Muraki, 1993), and the adaptation of IRT plots to CAT data requires changing the text in the plots. In addition, it is often desirable for researchers to change plots to present them in greyscale rather than color, which further requires not only changing the color but also the style of the lines (dashed, dotted, etc.). Frequently, graphical packages will also require separately applying these changes to the legend, requiring further modifications. The jrt package includes its own plotting functions, based on the general plotting package “ggplot2” (Wickham, 2009), which facilitates customization and adaptation to CAT data.

Facilitating Transition to IRT Models

In order to facilitate transition to IRT modeling for CAT researchers, we decided to include some non-IRT procedures within the jrt package. By doing so, our hope is that jrt can be used as a go-to statistical package for CAT data and that using IRT models instead of these classical analyses will be an easy step. In addition, it may be that the amount of data is insufficient for IRT calibration; in these cases, researchers may want to revert to more classical approaches. In jrt, one can thus output sum scores (along with the factor scores) as well as compute the indices of reliability that are most frequently used in CAT research, from both the classical test theory literature (especially Cronbach’s α), which are computed with the “psych” package (Revelle, 2017), and from the interrater reliability literature (Fleiss’ κ, Fleiss-Conger exact κ, as well as one- and two-way consistency and agreement model intraclass correlation coefficients), computed with the package “irr” (Gamer, Lemon, Fellows, & Singh, 2012).

In addition to this, jrt facilitates an understanding of IRT in a way that is close to traditional approaches, most notably by presenting the concept of reliability in IRT through the reliability metric, while a number of IRT software use information only. In IRT/JRT, the concept of information refers to how much an item/judge (or several; even an entire set) contributes to the estimation of the latent attribute for a person or product. A key feature of information is that it is *conditional* upon the latent attribute itself—in other words, different individuals/products may be measured with different accuracy by different items/judges. Although the advantages of using IRT-based information over CTT-based indices of reliability—especially Cronbach’s α—are often discussed, what is lesser known is that information can be rescaled to a (conditional) reliability estimate (Raju, Price, Oshima, & Nering, 2007).

Although conditional reliability is an essential benefit of IRT, estimating reliability at group level (instead of examinee level) is also possible. Currently in jrt, the functions of mirt are used to compute empirical reliability and marginal reliability. Empirical reliability is computed as the variance across the person estimates, divided by the sum of the same variance added to the average of the squared standard errors. Marginal reliability (which, for clarity, we call “expected reliability” in jrt) is similarly computed, but instead of using averaging across sample estimates, it is based on integration using a prior distribution of person estimates (Green, Bock, Humphreys, Linn, & Reckase, 1984)—here a standard normal distribution. We anticipate that alternate computations of IRT-based reliability (e.g., Brown & Maydeu-Olivares, 2011; Culpepper, 2013; Raju et al., 2007), as well as confidence interval computations (e.g., Myszkowski & Storme, 2018), will be added in the future.

In order to make conclusions about the observed reliability of the measure in the data–for example, if one aims to check reliability of the judgments prior to executing further analyses–we recommend to report the empirical reliability in the sample. In contrast, in order to make inferences about the measurement device beyond the sample–for example, if new products are to be judged by the same panel–we recommend to report the expected reliability from a standard normal prior distribution. These estimates of reliability can be interpreted in a similar fashion as non-IRT-based estimates of reliability. In fact, if an IRT model is used for scoring and/or judge analysis, we recommend, for coherence between the scoring procedure and the psychometric analysis, to prefer IRT-based estimates of marginal reliability.

Pointing to Relevant Resources and Citations

Finally, because jrt is aimed at facilitating transition to IRT modeling, and because jrt stands on the shoulders of years of IRT modeling developments, it appeared important that jrt would point researchers to relevant resources and citations to allow researchers who use it to easily find and cite relevant information. Notably, references with digital object identifiers are provided for the IRT estimation engine, the estimation algorithm, and the IRT model used. We recommend, and it is common practice, to cite them. For researchers using jrt for their analysis, we especially recommend to the estimation engine used—currently mirt (Chalmers, 2012)—as well as jrt itself–through this paper–if they made a use of it that is a specific to it (automated model selection, plots, etc.).

Using jrt

In this section, we want to provide a quick introduction to the use of the jrt package. More (updated) information on the different options of the package is available in the documentation and vignette of jrt, which are freely accessible on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=jrt. Updates of the functions and their options can be accessed there.

The R Environment

R is a statistical programming language and environment that presents multiple advantages for researchers. Although our point is not to defend it, we want to stress some important features of R that are relevant to using jrt. First, as we mentioned earlier, R is free and available for the most common operating systems (e.g., PC, Mac, and Linux). The current version of R can be downloaded from CRAN at https://cran.r-project.org/. Although R has base packages automatically installed, a key advantage of R is its focus on community contributions. R easily allows researchers to share their code (usually primarily composed of statistical functions) through the submission of packages to CRAN, which users can easily download and install. All packages on CRAN go through checks to ensure that the packages can be installed on the current

R version. A first step of using jrt is to download and install R on a computer.

Installing and Loading jrt

The jrt package is hosted in CRAN and can be downloaded for free by typing the command “install.packages(jrt)” in the R console. jrt depends on other R packages, which will be installed by default. It mainly depends on the mirt package (Chalmers, 2012) as its estimation engine and on ggplot2 (Wickham, 2009) for its plots. For non-IRT reliability estimates, it also uses the psych (Revelle, 2017) and irr (Gamer et al., 2012) packages. The package “directlabels” (Hocking, 2017) is used to add labels on plots. All dependencies are currently maintained and have been for several years. Once installed, loading jrt only requires calling “library(jrt)” in the R console.

Importing and Preparing a Dataset

We will not extensively discuss how to import a dataset in R because it depends on several parameters, especially what the base file is. However, users beginning in R may be especially interested in the “rio” package (Chan, Chan, Leeper, & Becker, 2018) available from CRAN, which combines the possibilities of many data importing packages and allows for importing data from a large number of sources (including Excel and SPSS formats) with a simple “import(file)” call. It should also be noted that the R Studio Integrated Development Environment offers a convenient point-and-click solution to import from text, Excel, SPSS, SAS, and Stata data files.

Once a dataset is loaded in the R environment, the jrt user will need to subset it to only keep the variables that contain the judgments. There are different ways to do this in R. One easy way is to use the “select” function of the “dplyr” package (Wickham, François, Henry, & Müller, 2018), which is already installed as a dependency of jrt. For convenience, the argument “select.variables.that.contain” in the jrt function will filter in only the variables that contain a specific character string, meaning that, if one’s judgment data contains “judge” or “rater,” for example, this can be passed to jrt, which will do the variable filtering prior to analysis. For the rest of this brief tutorial, we will assume that a data frame named “data” is loaded in the R environment and contains the judgment data—in numeric format—only.

Fitting Models: The jrt Function

The jrt function is the main function of the jrt package. It takes a data frame containing the judgment data as its only required argument. The default use of jrt is very simple and consists of typing “fit <- jrt(data)” in the R console. From this call, jrt will (a) detect and deal with any possible issue of missing data or judges with unobserved categories by selecting only available models, (b) fit all available response models (a progress bar is displayed, as this process can be long in some data sets), (c) compare models, (d) select the best fitting model, (e) provide a summary of the model with estimates of the reliability, (f) output a total information plot, (g) output a faceted plot with judge category curves of all judges with an overlay of their reliability functions, and (h) store several outputs into an object (which in this tutorial we named “fit”) of class jrt, which may be further used.

Although the class-jrt object contains additional results, the package is meant to be didactic; thus, it directly presents a summary printed in the console. In Figure 1, we present an example of what is printed in the console from the function “jrt(data)” (on randomly simulated data). Along with this console output is also printed a layered judge category curve plot with reliability overlay and a total information/reliability plot.

The jrt function includes several other arguments for customization. They may evolve in the future; therefore, it appeared more appropriate that they appear on the vignette and documentation of the package. However, at this stage, we can note that they currently include (a) bypassing the automatic selection of models by providing a model selected a priori (provided in the “irt.model” argument), (b) getting traditional (non-IRT) statistics (Cronbach’s α, intraclass correlation coefficients, etc.) as supplementary output (provided with the “additional.stats” argument), (c) setting different names and name prefixes for the output objects, (d) changing the estimation algorithm, convergence tolerance, maximum iterations, and method of factor scoring, and (e) changing the criterion for the automatic model selection.

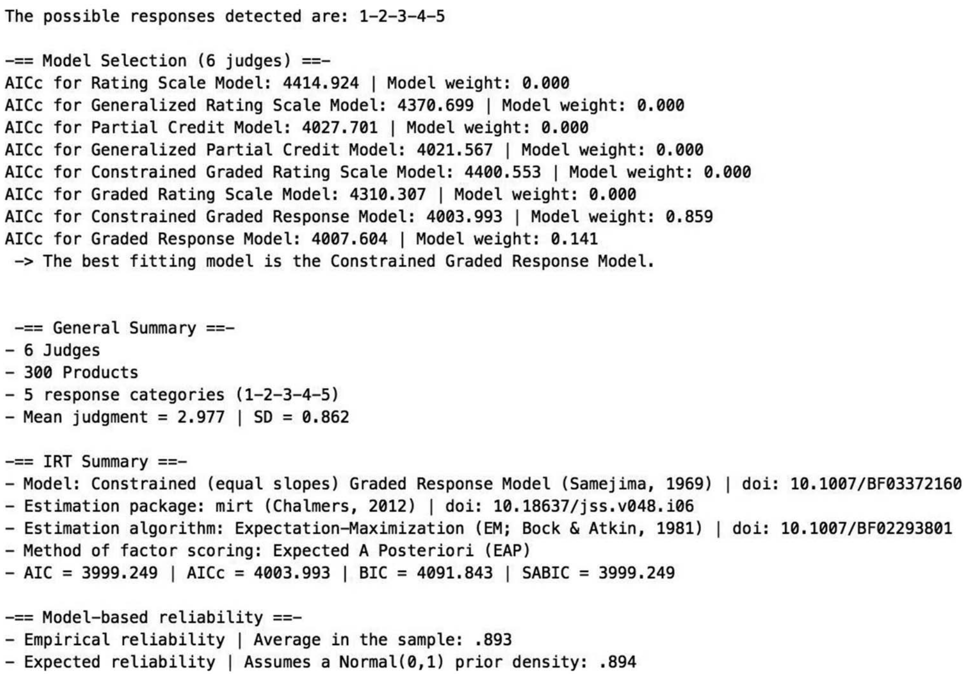
Getting Factor Scores and Standard Errors From the jrt Object

The created object of class jrt (fit) offers several outputs that are easily accessible. Of special interest to CAT researchers are the estimates of the latent attribute (e.g., creativity), which, when obtained through modeling procedures, are often named *factor scores*. Along with factor scores, it may be useful for researchers to append factor scores to the judgment data and to also retrieve standard errors of measurement (which, in IRT, are conditional upon the attribute measured and, thus, different per product judged). The factor scores can be retrieved in a dataframe with “fit@factor.scores” (which includes the factor scores, standard errors, and mean scores) and can be appended to the data with “fit@output.data.” For convenience, the vector of factor scores is also directly available in “fit@factor.scores.vector.”

Imputing Missing Values

If there were missing data in the input data set, jrt will warn the user that this is the case and automatically complete the dataset with plausible values imputation, which is based on the IRT model selected and factor scores and is provided through the mirt package. The data frame completed with plausible values imputation can be retrieved with “fit@complete.data.” The factor scores (based on the incomplete data) and original (incomplete) dataset are still retrievable with the same calls as described earlier.

Contrary to mirt, jrt does not require an input dataset in which all cases have at least one observed item/judgment score. In jrt, observations that are completely missing are only skipped in analysis, and their factor scores (and standard errors and means scores) are output as “NA.” Keeping them as NA instead of removing them prior to analysis allows the output factor scores and standard error vectors to have the same length as the original data, thereby allowing users to append factor scores to the original dataset more easily.



```
The possible responses detected are: 1-2-3-4-5

-== Model Selection (6 judges) ==-
AICc for Rating Scale Model: 4414.924 | Model weight: 0.000
AICc for Generalized Rating Scale Model: 4370.699 | Model weight: 0.000
AICc for Partial Credit Model: 4027.701 | Model weight: 0.000
AICc for Generalized Partial Credit Model: 4021.567 | Model weight: 0.000
AICc for Constrained Graded Rating Scale Model: 4400.553 | Model weight: 0.000
AICc for Graded Rating Scale Model: 4310.307 | Model weight: 0.000
AICc for Constrained Graded Response Model: 4003.993 | Model weight: 0.859
AICc for Graded Response Model: 4007.604 | Model weight: 0.141
 -> The best fitting model is the Constrained Graded Response Model.


 -== General Summary ==-
- 6 Judges
- 300 Products
- 5 response categories (1-2-3-4-5)
- Mean judgment = 2.977 | SD = 0.862

-== IRT Summary ==-
- Model: Constrained (equal slopes) Graded Response Model (Samejima, 1969) | doi: 10.1007/BF03372160
- Estimation package: mirt (Chalmers, 2012) | doi: 10.18637/jss.v048.i06
- Estimation algorithm: Expectation-Maximization (EM; Bock & Atkin, 1981) | doi: 10.1007/BF02293801
- Method of factor scoring: Expected A Posteriori (EAP)
- AIC = 3999.249 | AICc = 4003.993 | BIC = 4091.843 | SABIC = 3999.249

-== Model-based reliability ==-
- Empirical reliability | Average in the sample: .893
- Expected reliability | Assumes a Normal(0,1) prior density: .894
```

Figure 1. Statistics output in the R console from “jrt(data).”

Plotting Judge Category Curves: The “jcc.plot” Function

Plots are central in IRT modeling; thus, in jrt, it appeared important to include plotting functions that could output graphs that are publication-ready without necessitating a lot of tweaking. Because tweaking the outputs from existing IRT packages did not offer enough customization possibilities, dedicated plotting functions were developed for jrt: “jcc.plot” and “info.plot.” The former function allows users to easily plot item/judge category curves (JCC), which are central in polytomous IRT (Muraki, 1993). The JCC plot presents the predicted probabilities of each response category as a function of the latent trait. In the context of CAT, it presents the predicted probabilities that a given judge will attribute a specific score to a product as a function of that product’s attribute (e.g., creativity). Examining JCC plots allows for better understanding of judge variability and how it is accounted for by the IRT model chosen. By simply passing the jrt-class object created by the jrt function (fit) and a judge number (e.g., “3” for Judge 3), the function “jcc.plot(fit, 3)” will output the corresponding JCC plot. In Figure 2, we present the output of jcc.plot(fit, 3). Here and throughout, the plots were adapted to greyscale in the print version by simply adding the argument “greyscale = TRUE.” As can be seen, labels—provided with the directlabels package (Hocking, 2017)—are by default used to distinguish the response categories. If there are a lot of response categories, it may be more readable to use a regular legend by passing the argument “labeled = FALSE” (we present an example of this later).

The jcc.plot function allows for many customization options. A notable one is the possibility to output a faceted plot with several judges plotted on the same graph. To obtain a faceted plot, one can simply pass a vector of judge numbers (instead of a single value) to the “judge” argument (which is the second argument of the function). For example, to plot the JCC of Judges 1 through 3, one can call “jcc.plot(fit, 1:3)”; or if one wants to compare Judge 1 with Judge 4, one can call “jcc.plot(fit, c(1,4)).” Alternatively, by not providing anything to the judge argument by calling “jcc.plot(fit),” the function will plot all judges, which is presented in Figure 3.

Although we aimed for the plots to be publication-ready without tweaking, several elements may be customized. They are further discussed in the documentation and vignette of the package, but one can notably overlay the conditional reliability function with “overlay.reliability = TRUE,” use a legend instead of labels placed on the curves with “labeled = FALSE,” change/remove the title with the “title” argument, change the range of the latent attribute with “theta.span,” or replace the name of “judge” with something else with “preferred.name.for.judge.” Regarding aesthetic considerations, a layout theme may be used with the argument “theme,” the font family and size may be modified with “font.family” and “font,” respectively, the color palette may be changed with the “color.palette” argument, and the width of the lines may be changed with “line.width.” In Figure 4, we present an example of a few of these customizations with “jcc.plot(fit, 2, overlay.reliability = TRUE, theme = classic, theta.span = 5, title = ““, labeled = FALSE, legend.position = bottom).”

Plotting Judge Information Plots: The “info.plot” Function

An important advantage of IRT models is that, in the IRT framework, the accuracy of estimation of a latent attribute depends on the attribute measured. This is represented graphically in information plots, which show the precision of measurement provided

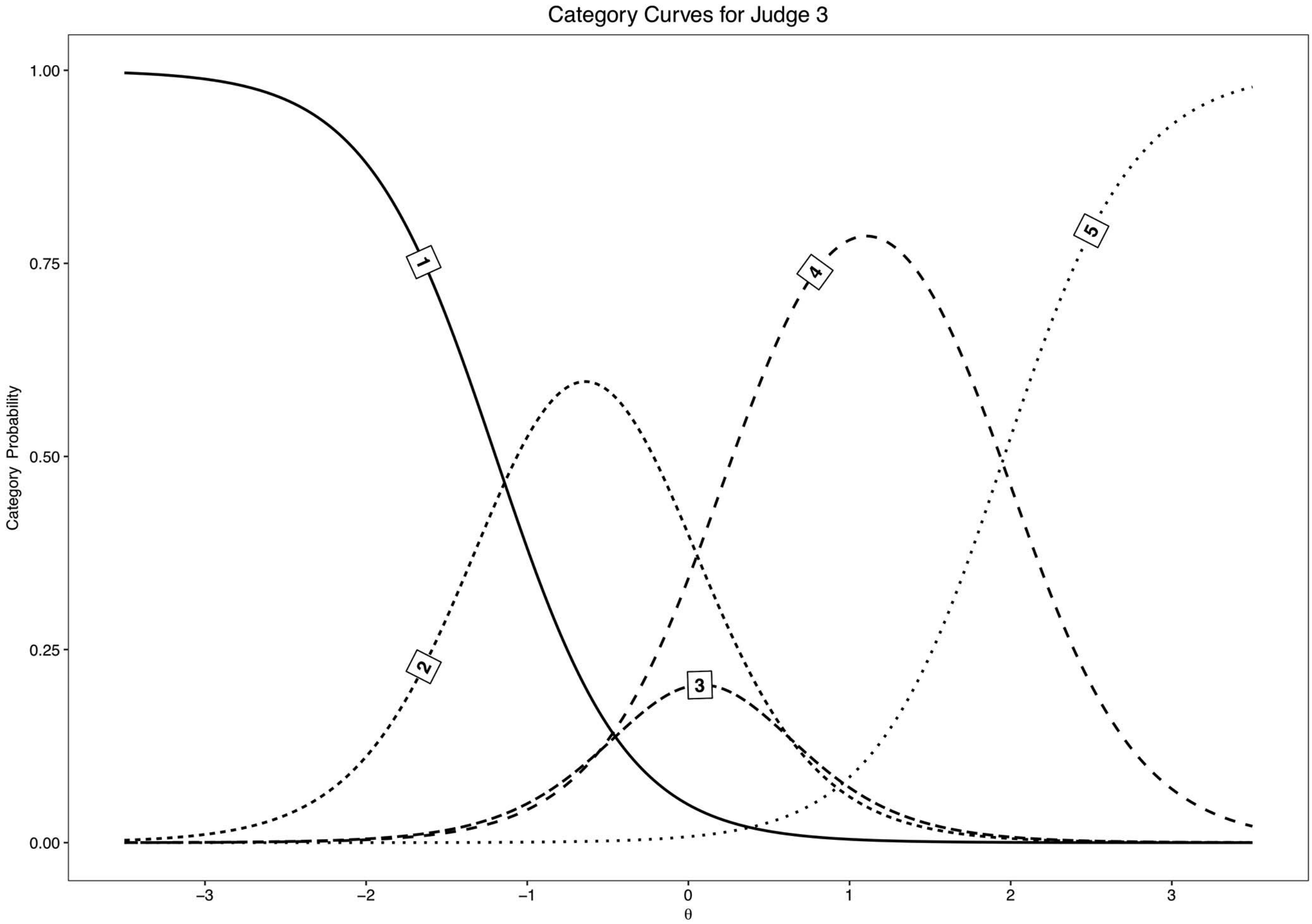


Figure 2. Example judge category curve plot from the "jcc.plot()" function.

by a judge, item, or test as a function of the latent attribute, and are essential outputs in polytomous IRT (Muraki, 1993). The metric used for information plots in IRT is traditionally the Fisher information (the inverse of the squared standard error of the estimator), which is used as an indicator of measurement precision and, thus, "replaces" the concept of reliability used in CTT. However, transformations of Fisher information may be used to express precision in a standard error of measurement metric or a reliability metric (Raju et al., 2007). The reliability metric is more rarely used than the information metric, but it presents the advantage of ranging between 0 and 1; thus, it is graphically convenient when one wants to overlay reliability with category probability curves (which is an option of the jcc.plot function).

Information plots thus represent measurement precision as a function of the latent attribute, and the jrt package can easily plot them with the info.plot function. The info.plot function has only one required argument, which is the object of class jrt (fit). When only providing this, as presented in Figure 5, "info.plot(fit)" will plot the total information (of the entire set of judges) as a function of the latent attribute (typically, creativity).

If one wants to express measurement precision using the reliability or standard error metric, one can pass this to the "type" argument of the function. It is also possible to plot information with standard error (or reliability) as a secondary *y*-axis, which presents redundant information but may be useful for didactic purposes. Finally, the information, reliability, or standard error function of each judge can be plotted by passing a number for a judge in the judge argument. Graphical options similar to jcc.plot (and with consistent names) are available for this function, which are more detailed in the documentation materials.

Discussion

The Consensual Assessment Technique is a central paradigm in the measurement of creativity and creativity-related product attributes, such as aesthetic quality, elaboration, appropriateness, novelty, and originality. Although there is a substantial amount of research on this technique, it rarely concerns its psychometric modeling aspects and more often concerns the expertise (or lack thereof) of the judges. Yet even if one is not interested in how the relation between creativity (or other attributes) and the judgment scores is structured, nor in understanding judge variability, one must at least be interested in how creativity is estimated. After all, estimating creativity is the final step in achieving its measurement, which is the main purpose of the CAT.

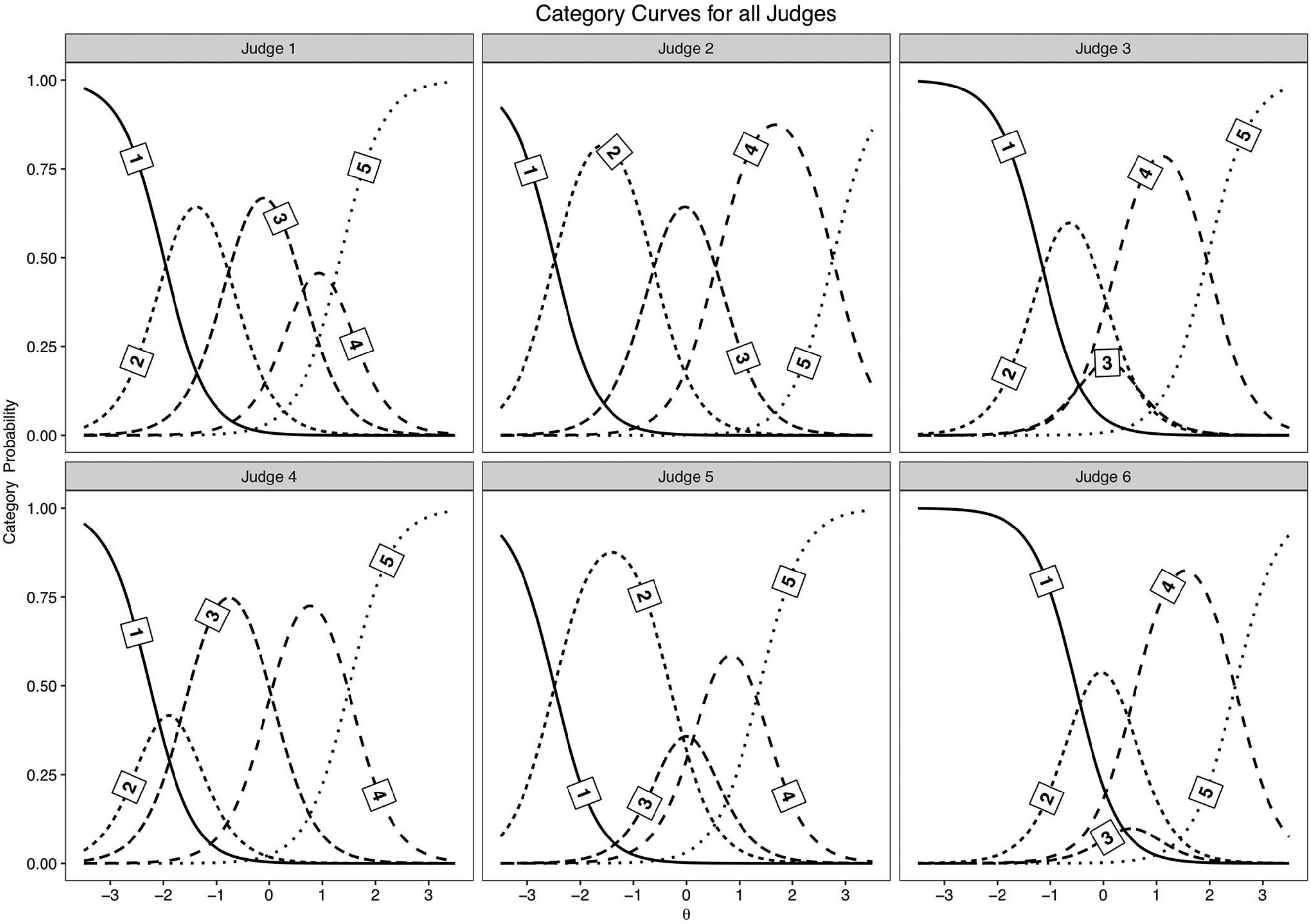


Figure 3. Example faceted judge category plot from the "jcc.plot()" function.

Recent perspectives on psychological measurement in general (Borsboom, 2006), and on the CAT specifically (Myszkowski & Storme, 2019), have extensively discussed the multiple conceptual and practical advantages of using measurement models, whether they take the form and name of item response models, latent variable models, or factor analysis models (Mellenbergh, 1994), over the regular practice of sum scoring. Sum scores are certainly more popular than estimates of attributes achieved with measurement models. They also may, in some cases and for certain uses, be relatively accurate proxies for creativity in CAT contexts. However, assuming—potentially falsely—that this is the case poses direct threats to the validity of the interpretations made with sum scores (Borsboom, 2006).

Although various reasons have been advanced for the prominence of classical test theory-based analyses, a traditionally invoked (Borsboom, 2006; Myszkowski & Storme, 2019) reason is that measurement models are not always accessible. In contrast with linear models, which are appropriate for continuous/normal data, measurement models for polytomous/ordinal data—which primarily consist of item response theory models—are perhaps even less accessible. As we noted, on one hand, there are several shortcomings to estimating polytomous IRT models with more popular modeling frameworks (such as structural equation modeling) and limitations to the user-friendliness of most IRT packages, especially when it comes to accommodating the specificities of judgment data. Noting this, the jrt package was created to provide an easy interface for researchers interested in using polytomous/ordinal IRT common factor models on judgment data.

Thanks to advances in IRT packages in R, especially mirt (Chalmers, 2012), jrt automatically fits an extensive set of ordinal response models on CAT data, detects potential modeling issues, compares models, produces model-based estimates of reliability, examines and accounts for judge variability using plots, and, most importantly, achieves model-based measurement through factor scores—all with simple calls. jrt directly encourages citing relevant statistical packages and procedures used and provides researchers with publication-ready outputs that do not necessitate tweaking. To sum up, while jrt does not hope to replace existing IRT packages and certainly does not offer the flexibility of some of these packages, it hopes to at least facilitate the transition of CAT researchers toward using IRT models.

As noted by one of the reviewers, beyond its use for judgments of one product per examinee, jrt may be used in virtually any situation where subjective ratings are used, such as subjectively rated divergent thinking task ideas. In addition, although it is not its primary purpose, we can anticipate that jrt could be used in

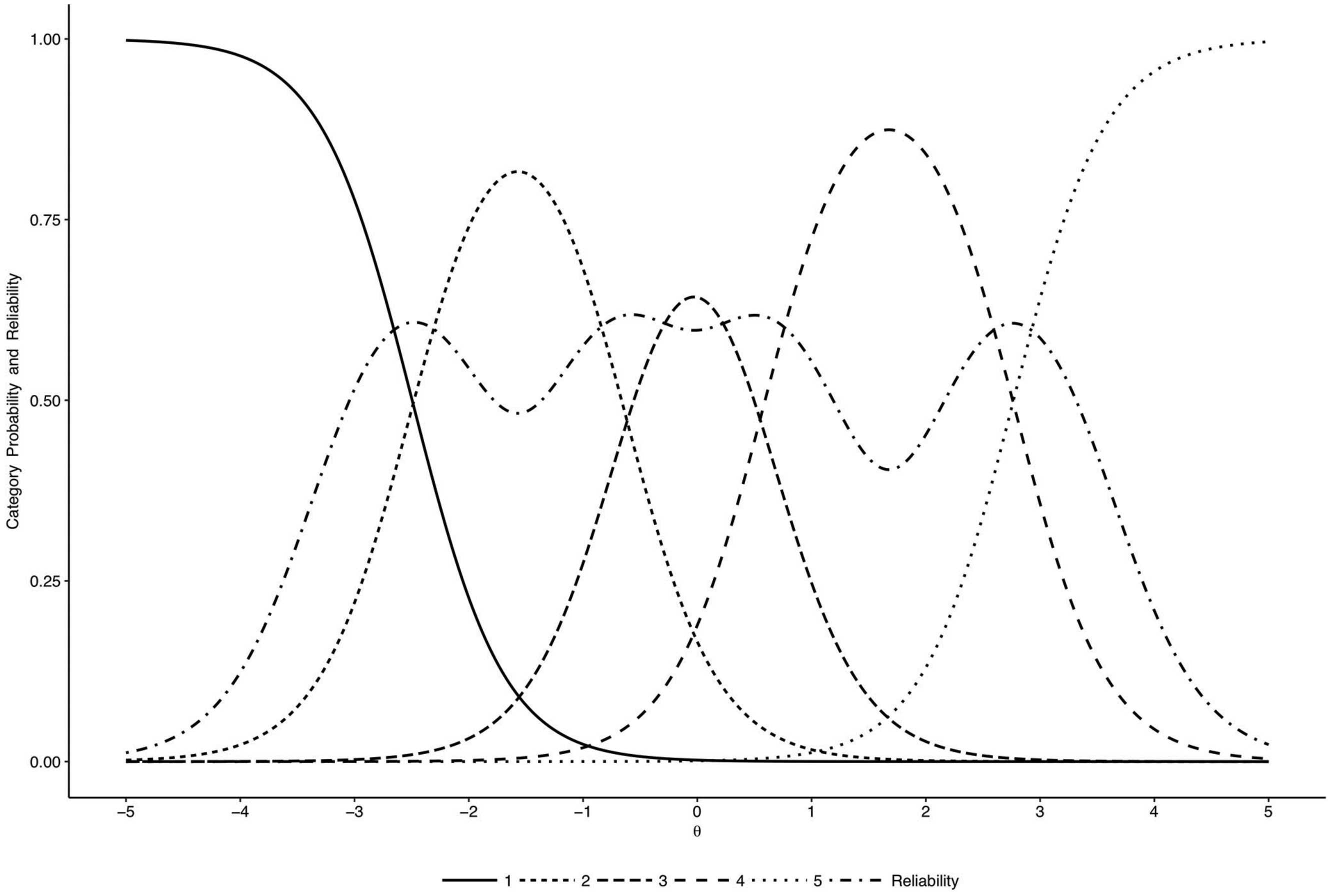


Figure 4. Example judge category plot with conditional reliability overlay with "jcc.plot()." Reliability = TRUE; Theme = classic; Theta.span = 5; title = ""; Labeled = FALSE; Legend.position = bottom.

other fields, especially in personality research, where Likert-type ordinal scales and common factor models are also heavily used. Further, the plotting functions of jrt can take mirt objects as inputs; thus, jrt can be used as a complement of mirt for its plotting capabilities fairly easily. mirt already has comprehensive plotting functions based on "lattice" (Sarkar, 2008), while the plotting functions of jrt are based on ggplot2 (Wickham, 2009). jrt thus offers an alternate plotting solution, which may be useful to some mirt users.

Before discussing the limitations of jrt, we should note that there are several questions that are not tackled here. An important one is sample size requirements of the different IRT models. These requirements depend on many factors—notably the model considered, the degree of accuracy desired, the parameter(s) under study, the estimation algorithm, the priors used for the parameters, and so forth; in general, IRT models are notorious for requiring large sample sizes. However, it was noted that recent advances in estimation algorithms have reduced sample size requirements without it appearing in psychometric trainings (Myszkowski & Storme, 2019; Zickar & Broadfoot, 2009). For example, Myszkowski and Storme (2019, pp. 173–174) have noted, based on a previous simulation study (Kieftenbeld & Natesan, 2012), that, for graded response models fitted with marginal maximum likelihood estimation, "as few as five items (judges) could already provide ability (creativity) estimates that are on average correlated at .84 with true attribute levels—with a negligible effect of how many products are judged." More research is evidently called to investigate the applicability and appropriateness of IRT models in realistic (or real) CAT situations.

Even though jrt permits executing the most classical ordinal IRT analyses and customizing them with very minimal coding, at this stage, the package itself presents several limitations, which we aim at circumventing in future iterations. To enumerate a few, first, jrt does not currently allow the fitting of multidimensional (or exploratory) IRT models. Although unidimensional models are implied in most CAT data collection efforts, where only one attribute is supposed to be measured, this may be a limitation for some research. Second, jrt only accommodates ordinal/polytomous data. Regarding this, most—if not all—CAT data collection efforts imply Likert-type ordinal scales, so this should be an irrelevant inconvenience in most cases.

Third, jrt assumes the latent attribute to be a continuous trait normally distributed in the population. It is important to note that this does not imply that the judgments themselves are assumed normal in the sample or population. It does imply, however, that researchers who want to formulate a different distributional assumption about the attribute—for example, the assumption that it is a class rather than a continuous trait—would currently need to

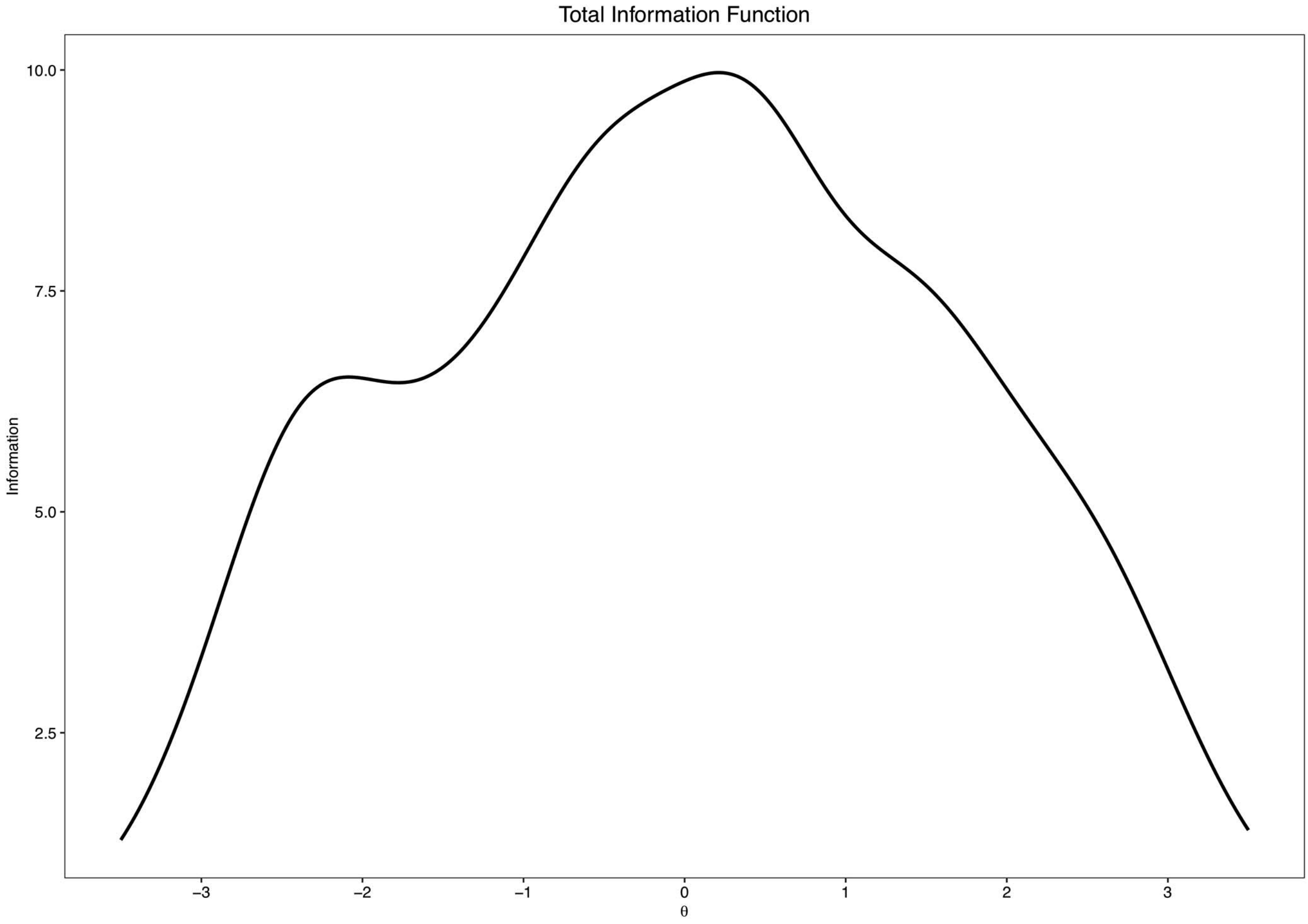


Figure 5. Example total information plot from the "info.plot()" function.

turn to other packages capable of handling latent class or mixture response models.

Fourth, jrt does not include models that accommodate for data sets with *simultaneously* multiple items (multiple products per individual) and judges. At its current stage, to deal with such multifaceted data sets, researchers would have to either estimate the attribute for each product separately (and then aggregate them) or use another package capable of fitting rater-mediated models, such as "TAM" (Robitzsch et al., 2017). The bifactor models included in mirt (Chalmers, 2012) may be a way that jrt could accommodate this in the future. Fifth, jrt should not be confused with an IRT estimation engine, in that it instead depends on an existing IRT estimation package, mirt (Chalmers, 2012). Although mirt is actively and successfully developed (Liu & Chalmers, 2018) and we could generally anticipate IRT estimation packages to grow in quantity and in features, jrt will have to adapt to the evolution of its dependencies.

Sixth, all the models currently implemented in jrt assume that the common product attribute (e.g., creativity) is the only explanation for the commonalities between the judgments of a product—an assumption referred to as *conditional* (or *local*) *independence* in IRT. This implies that two (or more) judgments, above and beyond being caused by the same latent attribute, are assumed independent. This assumption, in some cases, may be unrealistic. For example, if the judges are selected from distinct backgrounds (e.g., art students, art critics, or product designers), it may be suspected that their ratings of the creativity of the same products may be related within backgrounds over and beyond the product's creativity, thereby violating the assumption of conditional independence. If such phenomena are suspected, we would encourage turning to tests of conditional dependence or bifactor IRT models, which are not currently available in jrt.

Seventh, there is a growing body of research in IRT about how to integrate and use collateral information in the measurement of attributes. As proposed for aesthetic judgment tasks (Myszkowski, 2019), we can suggest that judgment time could prove invaluable in the detection of hasty (and thus potentially invalid) judgments. Eighth, there are still improvements to the didactic aspect of the package that could be made, such as pointing to more didactic resources (e.g., textbook chapters) and providing guidance in the interpretation of the parameters (e.g., with category response function plots annotated with the parameter estimates). Finally, jrt only requires very minimal coding compared with other packages, but it is still based on R and thus requires some familiarity (that we hope to be minimal) with this programming environment. Further improvements of jrt may include ways to improve its ease of use for non-R users.

There are certainly other limitations that we did not point out here. However, our priority in the development of jrt was to provide a usable toolbox for typical CAT data sets. We aim to clear these limitations by providing new features, following the evolution of the literature on the statistical treatment of CAT data.

Conclusion

Latent variable measurement models present substantial practical and conceptual advantages over classical test theory approaches (Borsboom, 2006). However, ordinal response models are hardly accessible for researchers nonversed in item response theory. Through jrt, we intend to make ordinal item response theory modeling, which is currently an exception in scoring and investigating CAT data, easily accessible for CAT researchers, in the hope that it will encourage them see the forest of benefits for the trees of statistical programming.

References

Akbari Chermahini, S., Hickendorff, M., & Hommel, B. (2012). Development and validity of a Dutch version of the Remote Associates Task: An item-response theory approach. *Thinking Skills and Creativity, 7,* 177–186. http://dx.doi.org/10.1016/j.tsc.2012.02.003

Amabile, T. M. (1982). The social psychology of creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology, 43,* 997–1013. Retrieved from http://www.hbs.edu/faculty/Pages/item.aspx?num=7355

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573. http://dx.doi.org/10.1007/BF02293814

Baer, J. (2008). Commentary: Divergent thinking tests have problems, but this is not the solution. *Psychology of Aesthetics, Creativity, and the Arts, 2,* 89–92. http://dx.doi.org/10.1037/1931-3896.2.2.89

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the Consensual Assessment Technique to nonparallel creative products. *Creativity Research Journal, 16,* 113–117. http://dx.doi.org/10.1207/s15326934crj1601_11

Baer, J., & McKool, S. S. (2009). Assessing creativity using the Consensual Assessment Technique. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). Hershey, PA: IGI Global. http://dx.doi.org/10.4018/978-1-60566-667-9.ch004

Baghaei, P., & Doebler, P. (2018). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports, 122,* 1967–1994. http://dx.doi.org/10.1177/0033294118797577

Barbot, B., Tan, M., Randi, J., Santa-Donato, G., & Grigorenko, E. L. (2012). Essential skills for creative writing: Integrating multiple domain-specific perspectives. *Thinking Skills and Creativity, 7,* 209–223. http://dx.doi.org/10.1016/j.tsc.2012.04.006

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. http://dx.doi.org/10.1007/s11336-006-1447-6

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30,* 505–514. http://dx.doi.org/10.1016/S0160-2896(02)00082-X

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11,* 27–34. http://dx.doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71,* 460–502. http://dx.doi.org/10.1177/0013164410375112

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Chan, C., Chan, G. C., Leeper, T. J., & Becker, J. (2018). rio: A Swiss-army knife for data file I/O (Version 0.5.16) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rio/citation.html

Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the Consensual Assessment Technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts, 13,* 159–166. http://dx.doi.org/10.1037/aca0000220

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement, 37,* 201–225. http://dx.doi.org/10.1177/0146621612470210

de Ayala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16,* 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0

Forthmann, B., Celik, P., Holling, H., Storme, M., & Lubart, T. (2018). Item response modeling of divergent-thinking tasks: A comparison of Rasch's Poisson model with a two-dimensional model extension. *The International Journal of Creativity & Problem Solving, 28,* 83–95. Retrieved from https://www.researchgate.net/publication/328290492_Item_response_modeling_of_divergent-thinking_tasks_A_comparison_of_Rasch%27s_Poisson_model_with_a_two-dimensional_model_xtension

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence, 57,* 25–32. http://dx.doi.org/10.1016/j.intell.2016.03.005

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement (Version 0.84.1). Retrieved from https://CRAN.R-project.org/package=irr

George, A., & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology, 11,* 189–205. http://dx.doi.org/10.20982/tqmp.11.3.p189

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01039.x

Hocking, T. D. (2017). directlabels: Direct labels for multicolor plots (Version 2018.05.22). Retrieved from https://CRAN.R-project.org/package=directlabels

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the Consensual Assessment Technique. *The Journal of Creative Behavior, 43,* 223–233. http://dx.doi.org/10.1002/j.2162-6057.2009.tb01316.x

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal, 20,* 171–178. http://dx.doi.org/10.1080/10400410802059929

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36,* 399–419. http://dx.doi.org/10.1177/0146621612446170

Liu, C.-W., & Chalmers, R. P. (2018). Fitting item response unfolding models to Likert-scale data using mirt in R. *PLoS ONE, 13,* e0196292. http://dx.doi.org/10.1371/journal.pone.0196292

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20,* 1–20. http://dx.doi.org/10.18637/jss.v020.i09

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. http://dx.doi.org/10.1007/BF02296272

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307. http://dx.doi.org/10.1037/0033-2909.115.2.300

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196. http://dx.doi.org/10.1007/BF02294457

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71. http://dx.doi.org/10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series, 1992,* i–30. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363. http://dx.doi.org/10.1177/014662169301700403

Myszkowski, N. (2019). The first glance is the weakest: "Tasteful" individuals are slower to judge visual art. *Personality and Individual Differences, 141,* 188–195. http://dx.doi.org/10.1016/j.paid.2019.01.010

Myszkowski, N., & Storme, M. (2017). Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). *Personality and Individual Differences, 117,* 91–100. http://dx.doi.org/10.1016/j.paid.2017.05.041

Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence, 68,* 109–116. http://dx.doi.org/10.1016/j.intell.2018.03.010

Myszkowski, N., & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts, 13,* 167–175. http://dx.doi.org/10.1037/aca0000225

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models*. http://dx.doi.org/10.4135/9781412985413

Rabe-Hesketh, S., & Skrondal, A. (2016). Generalized linear latent and mixed models. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (Vol. 1, pp. 503–526). Boca Raton, FL: CRC Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190. http://dx.doi.org/10.1007/BF02295939

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31,* 169–180. http://dx.doi.org/10.1177/0146621606291569

Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research (Version 1.8.12). Retrieved from https://CRAN.R-project.org/package=psych

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17,* 1–25. http://dx.doi.org/10.18637/jss.v017.i05

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules (Version 3.3–10). Retrieved from https://CRAN.R-project.org/package=TAM

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34,* 1–97. http://dx.doi.org/10.1007/BF03372160

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York, NY: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Sen, S. (2016). Applying the mixed Rasch model to the Runco Ideational Behavior Scale. *Creativity Research Journal, 28,* 426–434. http://dx.doi.org/10.1080/10400419.2016.1229985

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., . . . Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2,* 68–85. http://dx.doi.org/10.1037/1931-3896.2.2.68

Storme, M., Myszkowski, N., Çelik, P., & Lubart, T. (2014). Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges. *Learning and Individual Differences, 32,* 19–25. http://dx.doi.org/10.1016/j.lindif.2014.03.002

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577. http://dx.doi.org/10.1007/BF02295596

van der Linden, W. J. (Ed.) (2016). *Handbook of item response theory: Models* (Vol. 1, 1st ed.). New York, NY: CRC. http://dx.doi.org/10.1201/9781315374512

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. http://dx.doi.org/10.3758/BF03206482

Wang, C.-C., Ho, H.-C., Cheng, C.-L., & Cheng, Y.-Y. (2014). Application of the Rasch model to the measurement of creativity: The Creative Achievement Questionnaire. *Creativity Research Journal, 26,* 62–71. http://dx.doi.org/10.1080/10400419.2013.843347

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Retrieved from www.springer.com/us/book/9780387981413

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A grammar of data manipulation (Version 0.8.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=dplyr

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 37–59). New York, NY: Routledge/Taylor & Francis Group.

Received January 5, 2019
Revision received August 28, 2019
Accepted September 8, 2019 ■