*Psychological Medicine* (2016), **46**, 2025–2039. © Cambridge University Press 2016
doi:10.1017/S0033291716000520

REVIEW ARTICLE

# Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges

**S. P. Reise* and A. Rodriguez**

*University of California, Los Angeles, USA*

Item response theory (IRT) measurement models are now commonly used in educational, psychological, and health-outcomes measurement, but their impact in the evaluation of measures of psychiatric constructs remains limited. Herein we present two, somewhat contradictory, theses. The first is that, when skillfully applied, IRT has much to offer psychiatric measurement in terms of scale development, psychometric analysis, and scoring. The second argument, however, is that psychiatric measurement presents some unique challenges to the application of IRT – challenges that may not be easily addressed by application of conventional IRT models and methods. These challenges include, but are not limited to, the modeling of conceptually narrow constructs and their associated limited item pools, and unipolar constructs where the expected latent trait distribution is highly skewed.

*Received 23 July 2015; Revised 22 February 2016; Accepted 23 February 2016; First published online 8 April 2016*

**Key words**: Item response theory, log-logistic model, psychopathology measurement, skewed latent trait, 2-parameter logistic model

## Introduction

Item response theory (IRT) measurement models and associated methods hold much promise for the development, psychometric analysis, refinement, and scoring of psychiatric measures (Reise & Waller, 2009). Many of the advantages of IRT modeling have been realized by large-scale federally funded projects such as the Patient Reported Outcomes and Measurement Information Systems (PROMIS; Cella *et al.* 2010; see http://www.nihpromis.org/science/PublicationsYears) and NIH Toolbox (Gershon *et al.* 2013; see http://www.nihtoolbox.org/Publications/Pages/Articles1.aspx).

In psychiatric measurement, arguably, many of the attractive features of IRT have been severely underutilized (Thomas, 2011; Yang & Kao, 2014). This state of affairs is unfortunate because as the ability to measure neurobiological variables increases, our ability to meaningfully link genetic variation and brain functioning with behavioral phenotypes (e.g. impulsivity, obsessive compulsive, attention deficit) critically depends on having strong, well justified latent variable measurement models for the phenotypes. In this regard, IRT measurement models might be critical to advancing the field (e.g. see Mungas *et al.* 2000; Tavares *et al.* 2004; Xu *et al.* 2015).

In this research, we review IRT measurement models, their assumptions, and applications to solving applied measurement problems. Analysis of data collected on 1101 healthy adults who responded to the dichotomously scored 19-item Eysenck Impulsivity Inventory (EY19; Eysenck & Eysenck, 1978) will be used to illustrate key concepts (see Table 1 for item content). The EY19 was selected as it is a widely used impulsivity instrument and represents a commonly implemented response structure (dichotomous). The main objective of this first section is to demonstrate that IRT modeling has much to offer psychiatric researchers. Our treatment of the technical details of IRT (e.g. parameter estimation, scale-linking methods) and associated statistical methods is necessarily sparse.

We also argue that, although IRT methods hold much promise, many researchers fail to recognize the limits of IRT when applied to non-educational/achievement constructs in general and psychiatric constructs in particular. These unique challenges include, but are not limited to, the modeling of conceptually narrow constructs and their associated limited item pools, and unipolar constructs where the expected latent trait distribution is highly skewed. Some of these issues have been raised and discussed previously in the domain of patient-reported health outcomes measurement (e.g. Reise & Revicki, 2015).

## What are IRT models?

The foundation of (unidimensional) IRT models[1]† is the assumption that a 'causal' common latent variable

* Address for correspondence: S. P. Reise, Ph.D., Department of Psychology, UCLA, Franz Hall, Los Angeles, CA 90095, USA.
(Email: reise@psych.ucla.edu)

† The notes appear after the main text.

**Table 1.** *Eysenck Impulsivity subscale item content*

| Item no. | Eysenck item no. | Item content |
|---|---|---|
| Item 1 | 7 | Do you often buy things on impulse? |
| Item 2 | 9 | Do you generally do and say things without stopping to think? |
| Item 3 | 11 | Do you often get into a jam because you do things without thinking? |
| Item 4 | 16 | Are you an impulsive person? |
| Item 5 | 19 | Do you usually think carefully before doing anything?* |
| Item 6 | 22 | Do you often do things on the spur of the moment |
| Item 7 | 25 | Do you mostly speak without thinking things out? |
| Item 8 | 26 | Do you often get involved in things you later wish you could get out of? |
| Item 9 | 27 | Do you get so 'carried away' by new and exciting ideas that you never think of possible snags? |
| Item 10 | 31 | Do you need to use a lot of self-control to keep out of trouble? |
| Item 11 | 33 | Would you agree that almost everything enjoyable is illegal or immoral? |
| Item 12 | 35 | Are you often surprised at people's reactions to what you do or say? |
| Item 13 | 38 | Do you think an evening out is more successful if it is unplanned or arranged at the last moment? |
| Item 14 | 42 | Do you usually work quickly, without bothering to check? |
| Item 15 | 43 | Do you often change your interests? |
| Item 16 | 44 | Before making up your mind, do you consider all the advantages and disadvantages?* |
| Item 17 | 48 | Do you prefer to 'sleep on it' before making decisions?* |
| Item 18 | 49 | When people shout at you, do you shout back? |
| Item 19 | 52 | Do you usually make up your mind quickly? |

Reverse-coded items denoted by (*).

underlies the responses to a set of scale items; thus, item responses are diagnostic of an individual's position on an underlying continuous latent variable. Having made that basic assumption, the next step in fitting an IRT model is to estimate the functional form linking levels on the latent variable to the probability of endorsing the item in the keyed direction (or with polytomous response formats, linking the latent variable with the probability of responding in each category). This process requires specification of a formal 'measurement model' of item responding. One common model applied to dichotomously scored personality items, such as the EY19, is the two-parameter logistic model (2PL) shown in equation (1).

$$P(x = 1|\theta) = \frac{\exp(\alpha(\theta - \beta))}{1 + \exp(\alpha(\theta - \beta))}. \quad (1)$$

In the above, individual differences on a continuous latent variable are denoted by theta ($\theta$). For purposes of statistical identification, this latent variable is typically defined to have a mean of 0 and variance of 1 in the calibration population. In turn, each item is characterized by two properties: the slope of the item response curve (IRC) at the inflection point ($\alpha$), and the location on the latent variable where the probability of responding to an item is 0.50 ($\beta$). Slopes in this model typically range between 0.7 and 2.0 with higher values (i.e. steeper IRCs) indicating more discriminating items. Location parameters, on the same metric as $z$ scores, *typically* range between −2 and 2, with positive values indicating that higher levels of the latent variable are required to endorse the item content and negative values indicating that the item is relatively easy to endorse, even for individuals low on the latent variable.

Equation (1) makes clear that the probability of responding to an item in the keyed direction, the IRC, is determined by both the individual's level on the latent variable ($\theta$) and item properties of discrimination and location ($\alpha$,$\beta$). Item parameter estimates, along with traditional item statistics for the EY19, are shown in Table 2. These parameters were estimated using marginal maximum likelihood with the *mirt* library (Chalmers *et al.* 2015) available in the R program (R Development Core Team, 2015). While *mirt* was used for these analyses, other R libraries (i.e. *irtoys*: Partchev, 2015; *ltm*: Rizopoulos, 2015) and commercial software (i.e. EQSIRT: Wu & Bentler, 2011; FlexMIRT: Cai, 2013; IRTPRO: Cai *et al.* 2011; Mplus: Muthén & Muthén, 2012) are capable of performing the same analyses. For a review of commercial IRT software see Han & Paek (2014).

First, observe in Table 2 that, generally speaking, IRT parameters roughly correspond to their traditional counterparts. That is, items with higher item-test correlations correspond to higher slopes (or discriminations), and items with lower endorsement rates have more positive location parameters. Second, there is a lot of variation in the slope parameters indicating that the items differ importantly in their relation with the latent

**Table 2.** *Classical test statistic, factor loadings, and 2PL model item parameters*

| Item no. | Item-test correlation ($r$) | $\lambda$ | Proportion endorsed | Slope | Location |
|---|---|---|---|---|---|
| Item 1 | 0.50 | 0.61 | 0.23 | 1.41 | 1.18 |
| Item 2 | 0.53 | 0.77 | 0.14 | 1.99 | 1.45 |
| Item 3 | 0.57 | 0.85 | 0.11 | 2.72 | 1.44 |
| Item 4 | 0.64 | 0.75 | 0.30 | 2.33 | 0.65 |
| Item 5 | 0.47 | 0.57 | 0.23 | 1.14 | 1.33 |
| Item 6 | 0.59 | 0.67 | 0.47 | 1.78 | 0.12 |
| Item 7 | 0.51 | 0.76 | 0.12 | 2.02 | 1.59 |
| Item 8 | 0.45 | 0.54 | 0.26 | 1.01 | 1.24 |
| Item 9 | 0.50 | 0.59 | 0.23 | 1.26 | 1.23 |
| Item 10 | 0.46 | 0.61 | 0.15 | 1.34 | 1.69 |
| Item 11 | 0.23 | 0.36 | 0.05 | 0.83 | 4.04 |
| Item 12 | 0.39 | 0.44 | 0.27 | 0.77 | 1.43 |
| Item 13 | 0.37 | 0.33 | 0.43 | 0.64 | 0.45 |
| Item 14 | 0.52 | 0.60 | 0.24 | 1.31 | 1.18 |
| Item 15 | 0.45 | 0.51 | 0.25 | 1.03 | 1.31 |
| Item 16 | 0.38 | 0.44 | 0.17 | 0.90 | 2.01 |
| Item 17 | 0.31 | 0.25 | 0.41 | 0.43 | 0.84 |
| Item 18 | 0.37 | 0.35 | 0.39 | 0.62 | 0.81 |
| Item 19 | 0.39 | 0.35 | 0.51 | 0.69 | −0.07 |
| Range | 0.23 to 0.64 | 0.25 to 0.85 | 0.05 to 0.51 | 0.43 to 2.72 | −0.07 to 4.04 |
| Standardized (raw) $\alpha$ | | 0.79 (0.78) | | | |
| Average inter-item correlation | | 0.16 | | | |
| Mean (S.D.) | | 0.26 (0.19) | | | |

$r$, Raw item-test correlations; $\lambda$, factor loadings.

variable. Third, although not a great deal, there is some variation in the item location parameters with most items having endorsement rates <0.50 and positive IRT location parameters. As will be clear shortly, this importantly affects where on the latent variable continuum the EY19 provides measurement precision, that is, the greatest information.

For didactic purposes, Fig. 1 displays the IRCs for the most (item 3) and least (item 17) discriminating items, and for the item with the lowest (item 19) and highest (item 11) location parameters. Note that, although item 11 has the highest location parameter, it also has the lowest item-test correlation. Based on item content, 'Would you agree that almost everything enjoyable is illegal or immoral?', it is reasonable that factors other than impulsivity are influencing item endorsement. Items 3, 4, and 7 appear to be the most differentiating items. The content for these items involves not thinking before speaking, getting into a 'jam' because of not thinking before speaking, and a directly stated question, 'are you an impulsive person?' The least discriminating items tend to occur toward the end of the scale[2], for example, items 17, 18, and 13. These items may be contaminated by other personality characteristics such as self-esteem (standing up for oneself if someone shouts) and spontaneity (an unplanned evening out).

Having estimated the item parameters, the next step is to consider how well these items measure individual differences on the underlying latent variable. In IRT, there is no notion of 'scale score reliability' or reporting of coefficient alpha internal consistency values, rather, what is critical is the contribution of each item in estimating an individual's position on the latent variable. This can be determined by translating the IRC for each item into an item information curve (IIC). In the case of the 2PL model this is shown in equation (2).

$$\text{info}|\theta = \alpha^2 P|\theta(1 - P|\theta). \quad (2)$$

The IIC indicates how much psychometric information (i.e. reduction in uncertainty) an item provides at each level of the latent variable. Items with larger slopes provide more information or discrimination. The location of the IIC along the latent variable continuum is determined by the item location – items provide the most information at the location parameter. For illustrative purposes, the IIC for each item from Fig. 1 is shown in Fig. 2. Because of the assumption of unidimensionality (and local independence to be described below), IICs are additive. Thus, a test information curve (TIC) can be derived by simply summing the IICs. This is shown in Fig. 3. The amount of test information conditional on the latent variable is

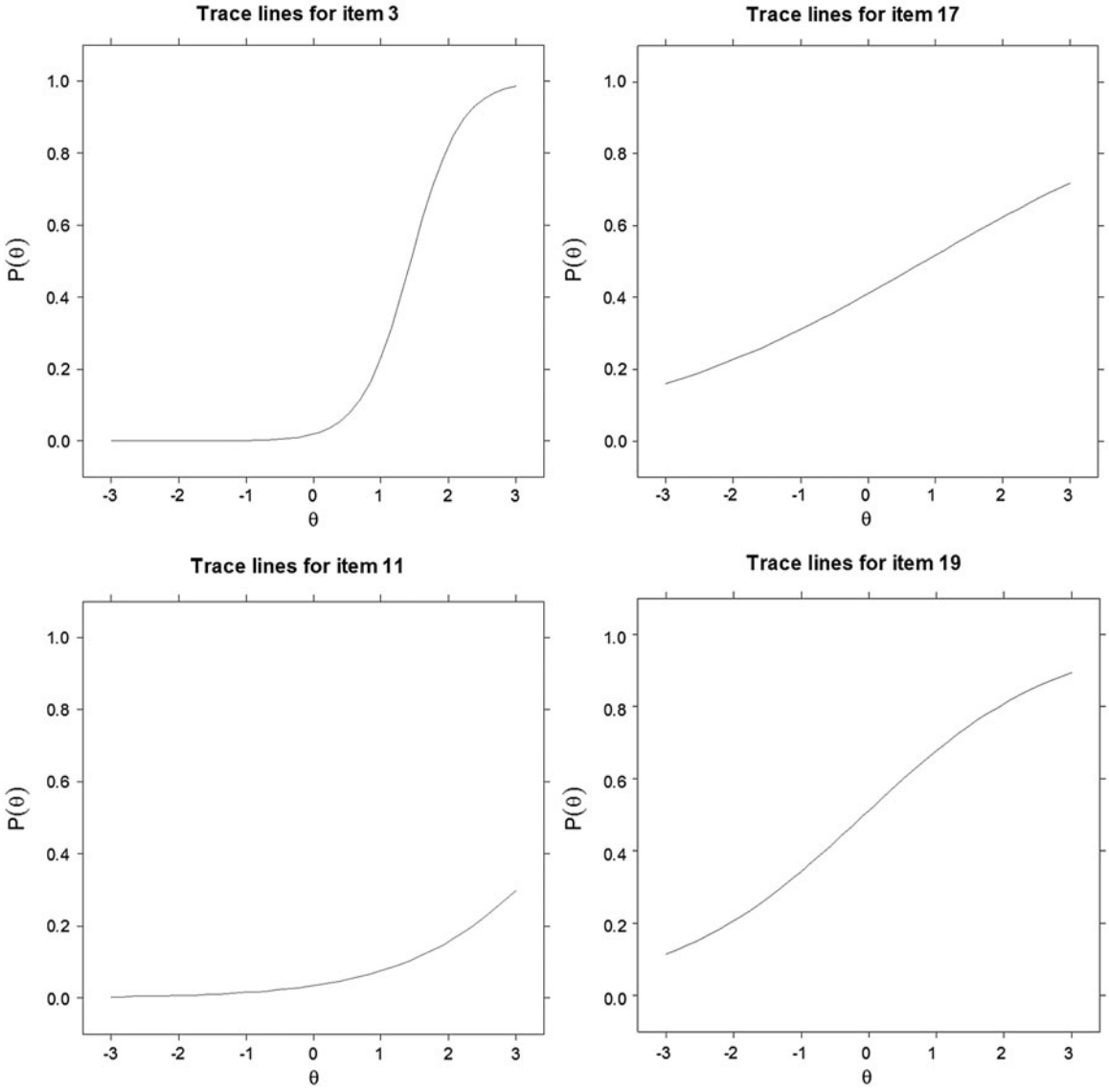


**Fig. 1.** Item response curves for items with the highest and lowest slopes (top two plots) and items with the highest and lowest location (bottom two plots).

important because it is inversely related to the standard error of the maximum-likelihood estimate of the latent variable as shown in equation (3).

$$\text{SEM}|\theta = \frac{1}{\sqrt{\text{TINFO}|\theta}}. \quad (3)$$

This formula demonstrates that the more information at a specific point along the latent variable, the higher the measurement precision (i.e. lower standard error). Thus, a conditional information value of 4 is required to produce a standard error of 0.50, 9 is required for a standard error of 0.33, and 16 is required for a standard error of 0.25. For illustrative purposes, the standard error of measurement curve for the EY19 is also shown in Fig. 3. Clearly, this is a peaked information function suggesting that the items are especially good at differentiating among individuals who are around one standard deviation above the mean on the latent variable. Observe that the measurement precision is much lower below the mean of the latent variable (zero), precisely because there are no items with location parameters in that range. We believe that such findings not only inform about the EY19, but also potentially tell researchers something important about the construct, as we argue below.

### Assumptions of IRT models and assessing fit

An important strength of IRT models, in part, is that they force a researcher to consider the assumptions

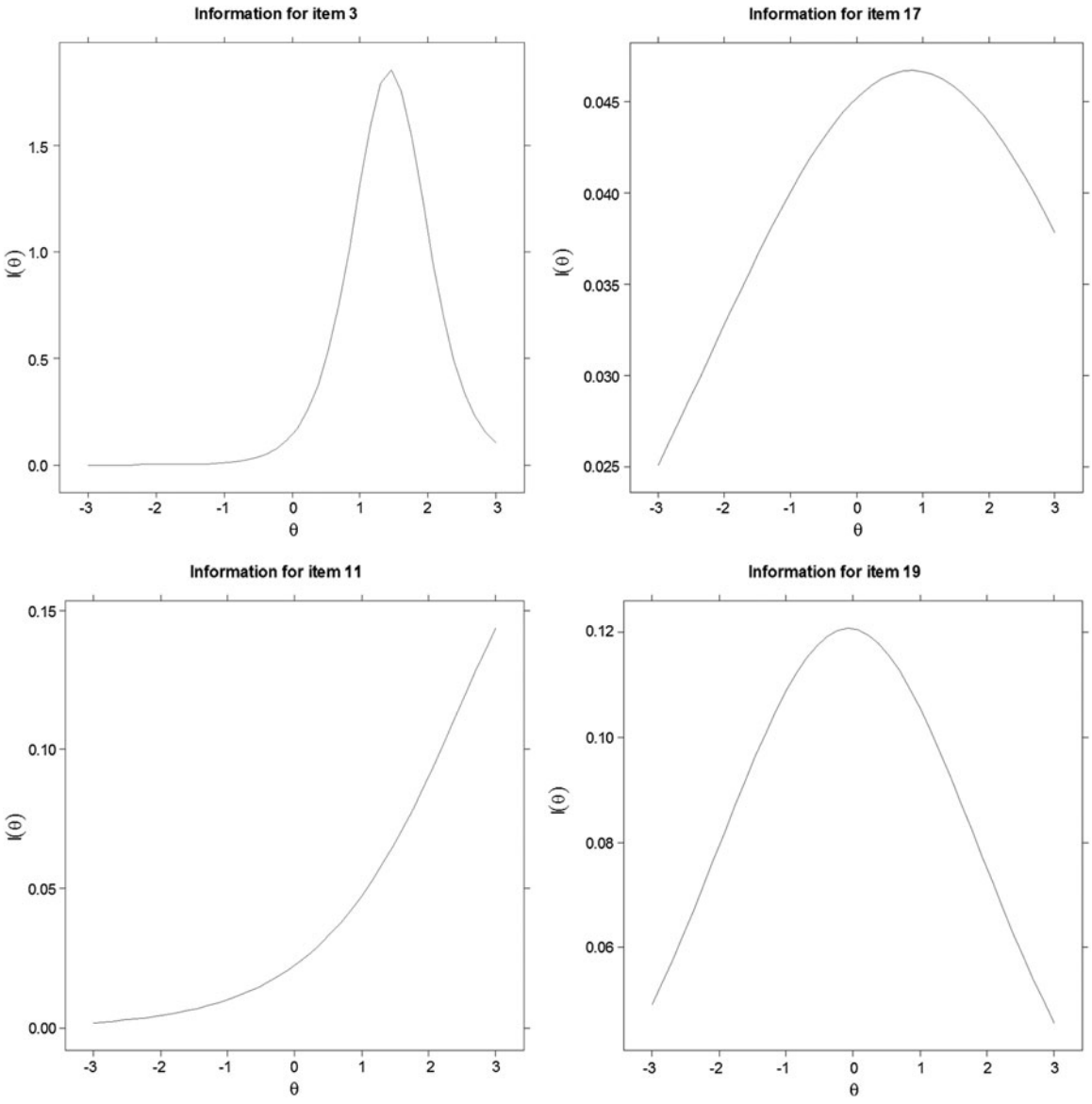


**Fig. 2.** Item information curves for items with highest and lowest discrimination and location.

underlying their application and to evaluate the degree to which the estimated model (i.e. the estimated item parameters) accounts for, reproduces, or 'fits' the data. In other words, IRT modeling forces researchers to study in detail how well the items serve as indicators of a latent variable and the overall quality of measurement across the latent variable continuum. This process of assumption checking and model-fit evaluation often reveals troubling concerns with an instrument, concerns that are easily overlooked in traditional scale analyses.

IRT models make three fundamental assumptions (Embretson & Reise, 2000). Evaluating these assumptions and judging the consequences of their violation, relies on application of complex statistics. We do not have the space here to elaborate on the plethora of available methods and their technical details but instead we hope to convey the major ideas.

First, because IRT models (i.e. the IRC) force a monotonically increasing relation between the latent variable and the probability of endorsing the item [see equation (1) and Fig. 1], the item response data must be monotonically increasing; as trait levels increase, so should the item endorsement rates. This assumption can be checked through the examination of rest-score functions (Meijer *et al.* 2015). Rest-score functions are simply plots of the raw total score (minus the item score) and the item proportion endorsed within a given rest-score grouping.

In Fig. 4 we display the rest-score functions (and confidence bands) for all 19 items. These were computed using the *mokken* library in R (Van der Ark,

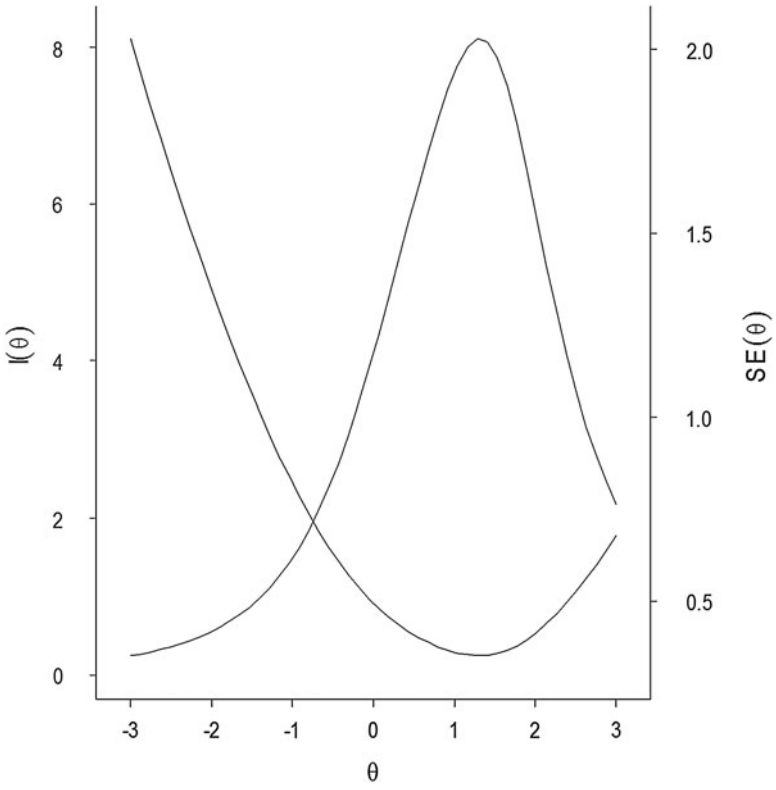


**Fig. 3.** Test information curve and standard error for Eysenck Impulsiveness subscale.

2014). These plots reveal that the rest-score functions for items 4 and 6 are 'ideal' in terms of applying a 2PL model – as rest-scores increase, the proportions endorsed start near zero, then increase in a systematic way across the entire rest-score range, and end up near 1.0. In contrast, the rest-score functions for item 11, appears troubling because response rates do not increase much as trait levels increase. Recall that this is the item with only 5% endorsement and a location above 4.0. In short, this item may be too extreme to be of any value in the measurement of impulsivity.

Also troubling, in terms of applying the 2PL model, is that some items appear to require non-zero lower asymptote parameters, for example, items 13, 17, 18, and 19. These are all items with relatively low slopes and relatively high endorsement rates. On the other hand, many items, such as 1, 5, and 7, appear to require a non-one upper asymptote to model appropriately. For those items, regardless of how high an individual scores on the EY19, the probability of endorsing a statement appears to have an upper bound <1.0. These examples represent the problem of model 'under-parameterization' – the 2PL model does not properly capture the response process.

More complex models such as a 3PL (a model with a non-zero positive lower asymptote), a 3PLU (a model with a non-one upper asymptote), or 4PL (a model where both a lower and upper asymptote parameter are estimated) may be required (see Reise & Waller, 2003, for discussion within the context of psychopathology measurement). The 4PL model is shown in equation (4).

$$P(x = 1|\theta) = c + (d - c)\ \frac{\exp(\alpha(\theta - \beta))}{1 + \exp(\alpha(\theta - \beta))}, \tag{4}$$

where $c$ is a lower asymptote parameter and $d$ is an upper asymptote parameter setting limits on the lower and upper tails of the IRC, respectively. The 3PLU model fixes all the $c$ parameters to zero and estimates $d$. The 3PL fixes all the $d$ parameters to 1.0 and estimates $c$. If $c$ is fixed to 0 and $d$ is fixed to 1.0, this model reduces to the 2PL in equation (1).

For illustration purposes, we estimated 3PLU and 3PL models on the present data and the results are shown in Table 3. Clearly, for several items, estimating additional parameters changes the slope and location[3] of the IRCs, and thus the IRC and IIC, and ultimately the test information and standard error. However, note that $\chi^2$ tests based on comparison of log-likelihoods indicated that the 2PL is not a significant decrement in fit compared to the 3PL ($\chi^2 = 12.61$, df = 19, $p = 0.86$) or 3PLU ($\chi^2 = 11.54$, df = 19, $p = 0.90$) and thus, the 2PL should be preferred due to parsimony. Moreover, the *mirt* program yielded a warning that parameter

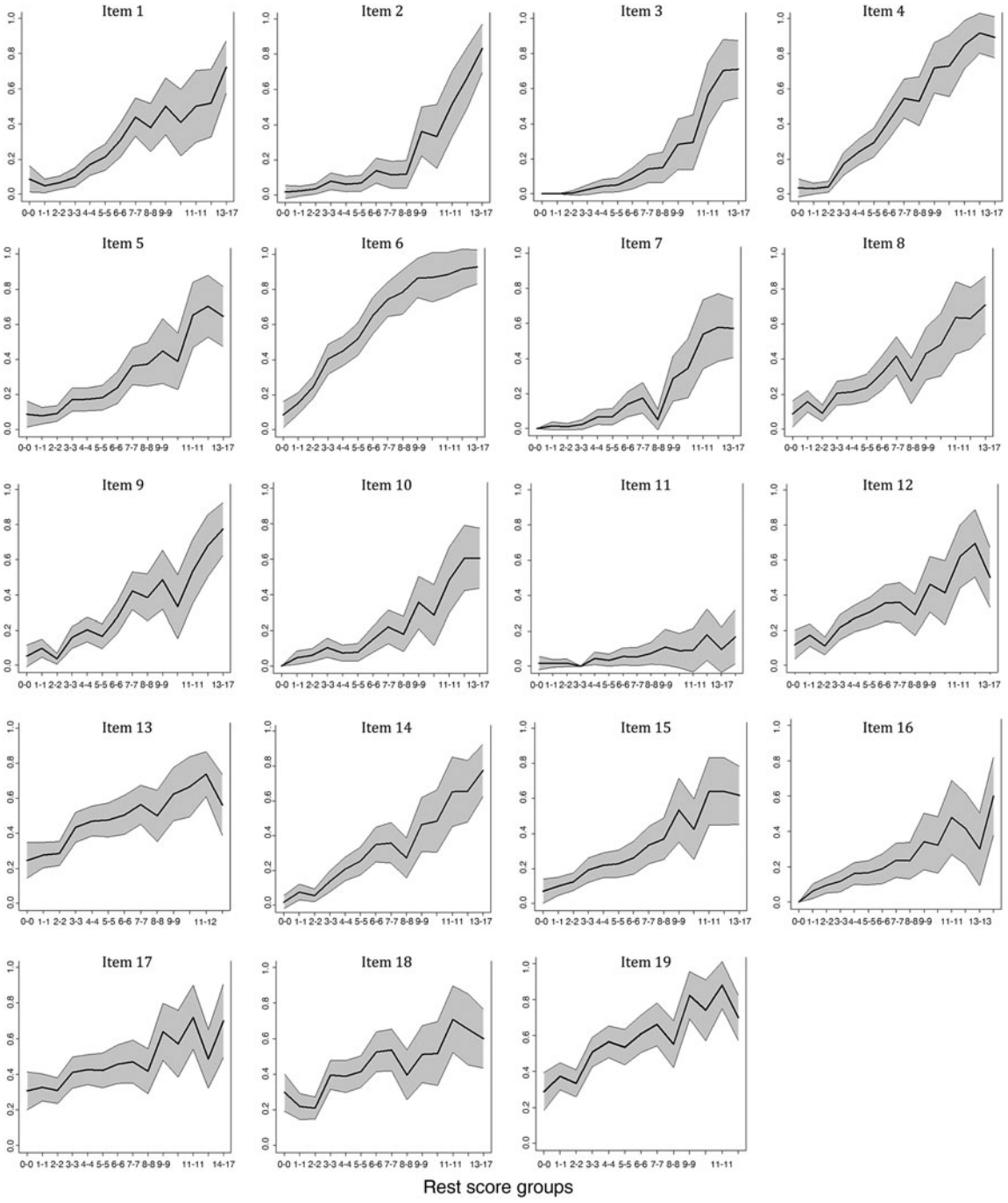


**Fig. 4.** Rest-score functions for all 19 items. Items in order (1–19) from left to right.

estimates in the 3PL and 3PLU models may be unstable, and suggested imposing priors. In other words, for accurate parameter estimation these models may require larger sample sizes than are presently available.

The second and third assumptions are highly interrelated, but can be considered as distinct concepts and empirically evaluated separately. Specifically, item response models assume unidimensionality and local independence. Unidimensionality means that the correlations among the items can be explained by a single common factor that represents the target construct a researcher designed the instrument to assess (impulsivity in the present case). The problem is that no psychological data that contains meaningful content heterogeneity will strictly meet this assumption. Thus, researchers typically examine item response data for evidence of 'essential' unidimensionality – the data

**Table 3.** *Item parameter estimates for the 3PLU, 3PL, and log-logistic models*

| | 3PLU | | | 3PL | | | Log-logistic | |
|---|---|---|---|---|---|---|---|---|
| Item no. | $\alpha$ | $\beta$ | $d$ | $\alpha$ | $\beta$ | $c$ | $\eta$ | $\lambda$ |
| 1 | 1.60 | 0.89 | 0.83 | 1.43 | 1.18 | 0.00 | 1.41 | 0.19 |
| 2 | 1.95 | 1.46 | 1.00 | 3.75 | 1.40 | 0.04 | 1.99 | 0.06 |
| 3 | 2.73 | 1.44 | 1.00 | 2.78 | 1.43 | 0.00 | 2.72 | 0.02 |
| 4 | 2.94 | 0.48 | 0.89 | 2.32 | 0.66 | 0.00 | 2.33 | 0.22 |
| 5 | 1.22 | 1.09 | 0.88 | 1.44 | 1.36 | 0.05 | 1.14 | 0.22 |
| 6 | 1.91 | 0.07 | 0.97 | 1.81 | 0.13 | 0.00 | 1.78 | 0.81 |
| 7 | 1.99 | 1.59 | 0.99 | 2.28 | 1.55 | 0.01 | 2.02 | 0.04 |
| 8 | 1,00 | 1.25 | 1.00 | 1.41 | 1.34 | 0.08 | 1.01 | 0.29 |
| 9 | 1.24 | 1.23 | 1.00 | 1.32 | 1.24 | 0.01 | 1.26 | 0.21 |
| 10 | 1.34 | 1.69 | 1.00 | 1.55 | 1.66 | 0.02 | 1.34 | 0.10 |
| 11 | 1.05 | 1.99 | 0.31 | 0.85 | 3.96 | 0.00 | 0.83 | 0.03 |
| 12 | 0.77 | 1.41 | 0.99 | 0.86 | 1.48 | 0.03 | 0.77 | 0.33 |
| 13 | 1.23 | −0.58 | 0.68 | 0.65 | 0.45 | 0.00 | 0.64 | 0.75 |
| 14 | 1.3 | 1.17 | 1.00 | 1.32 | 1.18 | 0.00 | 1.31 | 0.21 |
| 15 | 1.02 | 1.31 | 1.00 | 1.13 | 1.33 | 0.02 | 1.03 | 0.26 |
| 16 | 1.07 | 1.22 | 0.68 | 0.93 | 1.96 | 0.00 | 0.90 | 0.16 |
| 17 | 0.45 | 0.62 | 0.95 | 0.60 | 1.46 | 0.15 | 0.43 | 0.70 |
| 18 | 0.63 | 0.71 | 0.97 | 0.63 | 0.82 | 0.00 | 0.62 | 0.61 |
| 19 | 0.99 | −0.59 | 0.82 | 0.72 | −0.06 | 0.00 | 0.69 | 1.05 |

3PLU is a three-parameter logistic model with upper asymptote; 3PL is a three-parameter logistic with lower asymptote; $\alpha$, item discrimination; $\beta$, item location; $c$, lower asymptote; $d$, upper asymptote; $\eta = \alpha$; $\lambda = \exp(-\alpha\beta)$.

are close enough to unidimensionality such that the item parameters are relatively unbiased reflections of the relation between the target latent variable and the item responses, and examinee's theta estimates reflect individual differences on the intended latent variable (i.e. are not overly biased by multidimensionality).

One of many ways to assess essential unidimensionality is the evaluation of the relative size of eigenvalues. In the present data, for example, the first five eigenvalues are 6.2, 1.2, 0.6, 0.4, and 0.3 suggesting that the first factor is much stronger than the additional common factors by a factor of roughly 6:1. Moreover, when the data were factor analyzed, no interpretable 2-, 3-, or 4-factor models could be identified. Alternatively, IRT estimation programs such as *mirt* also report more familiar model fit indices for the unidimensional model that are similar to those routinely used in structural equation modeling (SEM; see Maydeu-Olivares *et al.* 2011). In the current data, we found: RMSEA = 0.056 (0.052–0.061), CFI = 0.917, and SRMSR = 0.055. Judging by conventional standards, collectively, these vales indicate a good fit of a one-dimension model. Note, however, that the generalizability of SEM fit benchmarks to an IRT context is questionable.

Finally, local independence (LI) is the assumption that, after controlling for the latent variable, the correlations among items are zero. This indeed is also the technical definition of unidimensionality (i.e. when responses are locally independent after controlling for a single factor). As noted above, however, it is still possible for an instrument to be essentially unidimensional but the single latent factor does not reproduce the item responses exactly. This occurs commonly in personality and psychopathology scales due to redundant item content. When item content is too redundant, the correlation among those items may be inflated (due to sharing variance from a general and item specific factors). In turn, this may 'pull' the latent variable toward the item pair with the inflated correlation resulting in overestimated slope parameters (see also Steinberg & Thissen, 1996).

In the present data, we searched for local independence violations by computing the Chen & Thissen (1997) index, again using *mirt*. This index compares the observed response proportions in each 2 × 2 contingency table between item pairs, with those predicted based on the estimated model parameters. The values can be interpreted like $z$ scores where higher positive values indicate large positive residuals and large negative values indicate a large negative residual. We will only be concerned here with the large positive values, say, when the index is >10.

By this criterion, the most problematic item pair is items 16 and 5. A glance at Table 1 reveals these to be essentially the same question asked in slightly

different ways (basically, '*do you think carefully*' before acting). Other potential problem pairs are items 1 & 4, 2 & 7, 6 & 4, 12 & 8, 17 & 16, and 17 & 19. Although not obviously overly content redundant, there may be contingencies between these item pairs that are not accounted for by the unidimensional IRT model. These LI violations are important because if an item from one of these pairs were to be removed from the scale, the item parameters for the other items may change. If there were no LI violations, not only would the data be unidimensional in a strict statistical sense, the item parameter estimates would be invariant, that is, their values would not depend on what other items are included in the scale.

The above analyses are typically conducted either prior to estimating an IRT model (e.g. monotonicity assessment) or immediately after fitting a hypothesized model (LI evaluation). After a model has been fitted, however, a researcher must also conduct some empirical investigations of item fit. The evaluation of item fit remains a contentious issue in the literature and there is no current consensus on how best to proceed. Nevertheless, for illustrative purposes here we review a standard $\chi^2$ and graphical approaches for judging item fit.

To judge statistical fit, we reviewed chi-square item-fit values output from *mirt* based on a formula presented in Orlando & Thissen (2000, 2003). The computation of these indices is complex because it involves an iterative or recursive formula. Simply stated, these indices are based on comparison of the number of individuals endorsing each item conditional on the overall composite raw score, compared with the proportion predicted based on the estimated IRT model parameters. These $\chi^2$ indices are notoriously powerful, but in the present data, only items 1 and 2 produced the warning flags of statistical significance below $p < 0.05$.

To follow-up on these results, Fig. 5 presents 'fit plots'. Specifically, for each individual, their location on the latent trait is estimated. Individuals are then grouped into, say 10, intervals along the theta continuum. For each interval, the observed response proportion is computed and compared to the estimated IRC. In the left panel is the fit plot for item 4 which had the best fit as judged by $\chi^2$. Clearly, the IRC does an excellent job of representing observed response proportions – all the confidence bands contain the IRC. Interestingly, the right panel displays the fit plot for the worst fitting item (item 1). Here, the problem seems to lie in the tendency for the estimated IRC to overestimate the observed response proportions around the mean theta, and then underestimate them for trait levels above the mean. Again though, despite statistical significance, the empirical fit is certainly 'in the ballpark' and we would not be overly concerned that the estimated item parameters for item 1 are greatly in error.

## The applications of IRT

The above establishes that, for the most part, the EY19 is amenable to a 2PL model. The three concerns are: (*a*) some items do not contribute meaningfully to the measurement of the latent variable, (*b*) several item pairs may display LI violations, possibly inflating the correlation among those item pairs and thus distorting model parameter estimates (and thus test information and standard error), and (*c*) many of the items appear to require a more complicated IRT model, such as a 3PLU model where the upper asymptote of the IRC is estimated, rather than assumed to be 1.0 as in the 2PL model. What exactly to do about these concerns would require extended discussion and possibly some simulations to judge the impact of any model misspecification.

There have been many texts and articles written comparing IRT psychometrics with traditional approaches. These articles typically emphasize four virtues of IRT modeling: (*a*) scale analysis and construction, (*b*) linking the scales from multiple measures of the same construct, (*c*) assessing differential item functioning (measurement invariance), and (*d*) administering tests via computerized adaptive testing (CAT). Each of these topics has its own large literature and we can only try to convey the major ideas here.

### *Psychometric analysis*

Several of the advantages of IRT modeling, in terms of scale construction and psychometric analyses, are obvious from the above application. Specifically, because IRT simultaneously considers both person properties and item properties, it provides a framework for clearly understanding how each item contributed to measurement of a latent variable. For example, we can compute a test information curve for any subset of EY19 items and know exactly what the standard errors of measurement would be for examinees at different trait levels. In turn, that information function makes clear where along the latent variable continuum measurement precision is high or relatively low, perhaps unacceptably low. This, of course, is critical information in terms of scale revision.

### *Linking scales*

Another chief advantage of IRT modeling is that, assuming two measures are assessing the same latent variable, it is relatively easy to use IRT 'linking' methods to place the item parameter estimates (and scale

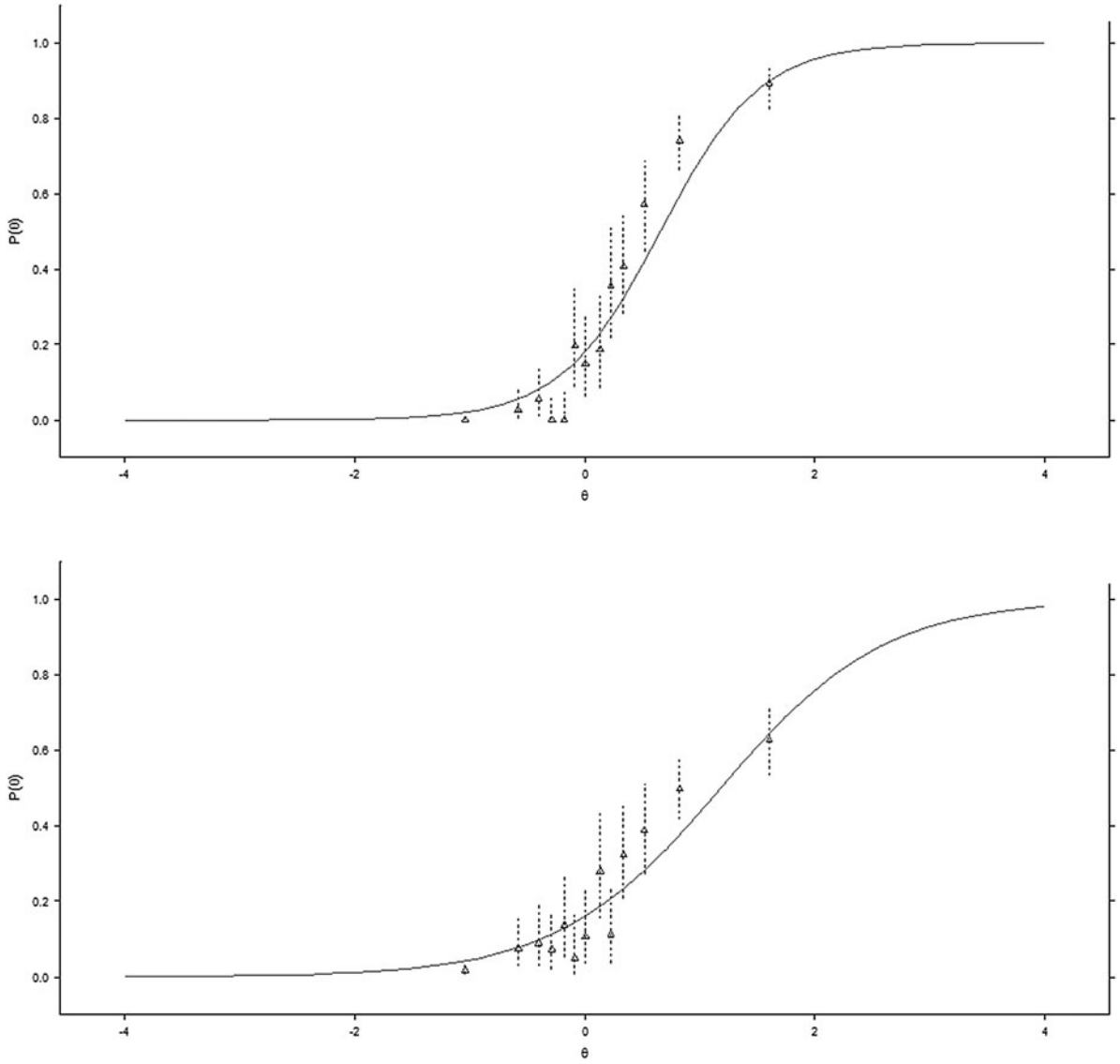


**Fig. 5.** Item fit plots for the best (item 4) and worst (item 1) items.

scores) from one measure on the scale of another measure. This is invaluable and is arguably one of the chief accomplishments of projects such as PROMIS cited earlier. As is well known, in certain health outcome domains, there are dozens of competing measures of similar constructs. This chaotic situation only serves to hamper attempts to progress research in a given substantive domain.

One goal of the PROMIS project was to use IRT linking methods to create a common metric (e.g. depression, anxiety), so that individuals could be compared not only if they were administered PROMIS depression or anxiety items, but also many popular competing depression or anxiety scales (Choi *et al.* 2014; Schalet *et al.* 2014). Put in the present context, the need for such an IRT project in the domain of 'impulsivity' (also known as 'cognitive control', 'response inhibition') measurement is clear and compelling, especially for researchers who hope to link biological parameters to self-reported impulsiveness phenotypes (e.g. Horn *et al.* 2003); however, like us, they are bewildered by the numerous conceptualizations and measures presently available.

### *Differential item functioning*

A further advantage of IRT modeling is that it provides an elegant framework for the evaluation of between-group differences in item or test functioning (i.e. whether an item or test is measuring the same construct in the same way for two or more groups). Simply stated, an item is invariant when IRCs estimated separately in demographic groups are equal.

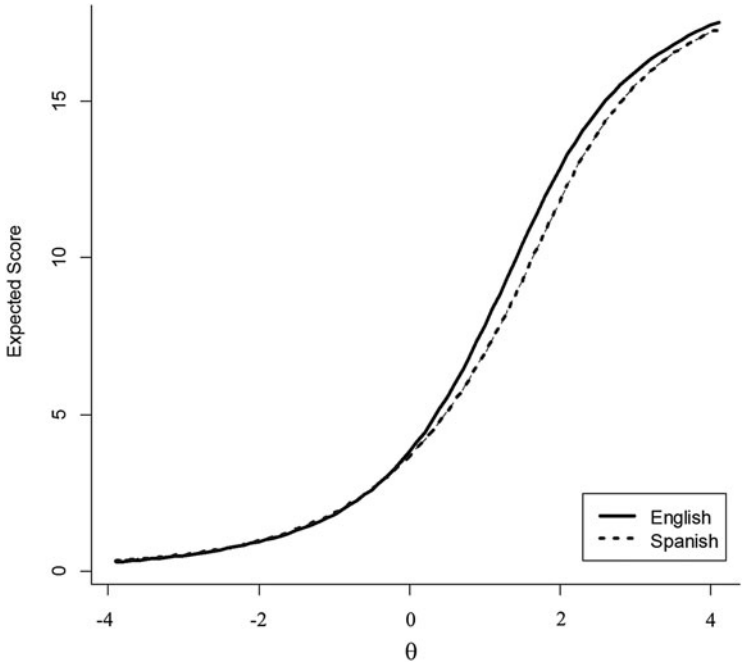


**Fig. 6.** Test response curves for English and Spanish speakers.

A test is considered invariant if the test response curves (TRC) are equivalent when estimated separately by group (with the caveat that the item parameters from the two groups have first been 'linked' to the same scale). A TRC is simply the sum of the IRCs and reflects the relationship between the latent variable and the expected total test score. In the EY19 for example, Fig. 6 shows the TRC when the item parameters are estimated separately for English and Spanish speakers. It is clear that there are no meaningful differences here suggesting that the EY19 is measuring the same construct with equal fidelity in both groups and there is no need to allow for different item parameters for the different populations.

### *Scoring and CAT*

A final advantage of IRT is that, assuming the data fits the model, any item response, or subset of item responses can be used to estimate an individual's position on the latent variable. This feature of IRT has resulted in a revolution in testing such that nowadays many major tests such as the Graduate Record Examination are administered as CATs (computer adaptive tests). A CAT is essentially a short-form tailored to the trait level of the individual examinee.

The basic idea of a CAT (or 'tailored' testing) is that a test begins by assuming, that in general, an individual is at the mean, that is, $\theta = 0$. An item of middle difficulty is administered ($\beta = 0$), a response collected, and then an individual's updated position on theta is estimated (typically using a Bayesian prior). A new item is then administered that maximizes the information at the individual's current trait level estimate, a response collected, and the latent trait score re-calibrated. This process is repeated until either a fixed number of items are answered or an individual's standard error is below a certain value.

In education, much research has demonstrated that CAT reduces the number of items administered by at least 50% with almost no loss in measurement precision. The reason being is that CAT is designed to only administer items that are relevant to an individual's trait level (i.e. hard items to high ability individuals and easy items to low ability individuals). In the EY19, we used the Firestar CAT (Choi, 2009) simulation program to estimate what would happen in our data if, hypothetically, the examinees were evaluated with CATs of 5, 10, and 15 items (of 19 possible items).

We found that trait level estimates based on CAT were correlated with theta estimates based on all 19 items ($r = 0.90$, 0.97, and 0.99, respectively). In other words, we can do a good job of recovering the relative ordering of individual differences with as few as 5 items, but with 10 (half the scale length) we recover full-scale theta almost perfectly. Interestingly, an examination of the items administered frequencies in the 5-item fixed length CAT showed that only items, 1, 4, 6, 14, and 19, were used with high frequency. This suggests that perhaps a short form should be created using just these five items.

**What are the challenges to applying IRT models?**

IRT models are attractive for many reasons as described above. In previous writings, however, we have urged caution in applying latent variable modeling techniques in general, and IRT modeling in particular in personality, psychopathology, and health outcomes domains (Reise & Waller, 2009). It should be noted that IRT methods were developed almost entirely within the context of large-scale educational/achievement assessment to meet the pressing needs of this industry. At a minimum, unlike typical performance measures, educational measurement specialists seldom are concerned with response styles, self-deception, acquiescence, and so on. For this and many other reasons, the transition of IRT models and methods to other construct domains is not always that simple. Below we review a few of these conceptual and empirical challenges.

Approximately twenty-five years ago Bollen & Lennox (1991) demonstrated that the type of psychometric interpretation implicit in IRT-type models (i.e. latent variable measurement) critically depends on the latent variable being a 'source' of item responses. When constructs are products of the indictors (items), such as 'social economic standing' or 'quality of family functioning', the latent variable changes depending on which particular items are included on a measure. For instance, Fayers & Hand (2002) note that quality of life scales often include items that function as causal indicators, that is, the concept being measured is defined by the items. With this latter type of construct, where items are 'causal', the entire logic and mathematics of IRT falls apart. Unfortunately, sometimes the IRT hammer appears to be applied to every possible nail without any consideration of the nature of the latent variable.

More recently, Reise & Revicki (2015), for example, have argued that educational constructs and health outcomes differ, typically, in a variety of ways including: (*a*) the interpretation of individual differences on the construct (i.e. unipolar *v.* bipolar), (*b*) the expected distribution on the latent variable (e.g. normal *v.* positively skewed), and (*c*) the conceptual bandwidth of constructs or, stated differently, the diversity of trait manifestations. In what follows, we briefly review these concerns and their implications for the application of IRT to the measurement of psychiatric constructs.

Consider the case of developing a measure of a broad bandwidth construct such as verbal ability in adults. In particular, consider three reasonable expectations: (*a*) there is a nearly infinite number of reading comprehensive paragraphs, vocabulary, analogies, grammar items, and so on from which to build a pool of items, (*b*) a normal distribution is likely adequate to describe individual differences, in the population, and (*c*) both ends of this distribution are meaningful, and psychologically interpretable, that is, the construct is bipolar. This is an ideal situation to apply IRT models because we can expect to find items that vary in location parameter across the entire trait range and thus have good measurement precision across that trait range, which in turn makes CAT a feasible option.

Moreover, although IRT makes no formal normality assumption, item parameters are typically estimated assuming a normal prior for the latent variable. Violating this assumption (i.e. a misspecification) can only lead to biased parameter estimates. Estimation methods that potentially remedy the parameter bias that occurs when the latent variable is misspecified (Woods, 2006) are only now emerging in commercial software (Monroe & Cai, 2014). Alternatively, new IRT modeling approaches that are specifically designed to account for skewed latent trait distributions are currently being researched, for example, see Molenaar (2014). It will be interesting to see how these new psychometric developments ultimately change the way IRT models are applied to psychiatric data in future years.

Now consider the measurement of 'impulsivity' as exemplified by the EY19 or measures of psychopathology more generally. Although the trait arguably manifests across a wide range of behavioral domains, once one has asked a few questions about thoughtless and reckless behavior (e.g. spending), lack of cognitive mediation in decision making (e.g. deliberates carefully before taking action, makes up mind quickly), and labile attention and interests (frequently changes interests or hobbies), the pool of content quickly runs dry. This is certainly not an ideal situation for the development of an item pool for the administration of a CAT; if even feasible, it appears to us that such a pool would contain many overly content redundant items.

It also, to us, remains debatable whether impulsivity, as least as measured by the EY19, is truly a bipolar construct with scores that are interpretable across the entire latent variable continuum, from low (reflecting 'self-control' or 'conscientiousness') to high (reflecting 'impulsiveness'). We believe that impulsivity may be more of a quasi-trait (see Reise & Waller, 2009), or what Lucke (2015) has termed a unipolar trait – definable and meaningful only at one end of the continuum (e.g. pathological gambling). In other words, we speculate that only variation among high scores reflect anything substantively meaningful, while low scores reflect the mere absence of impulsivity, not necessarily high degrees of self-control.

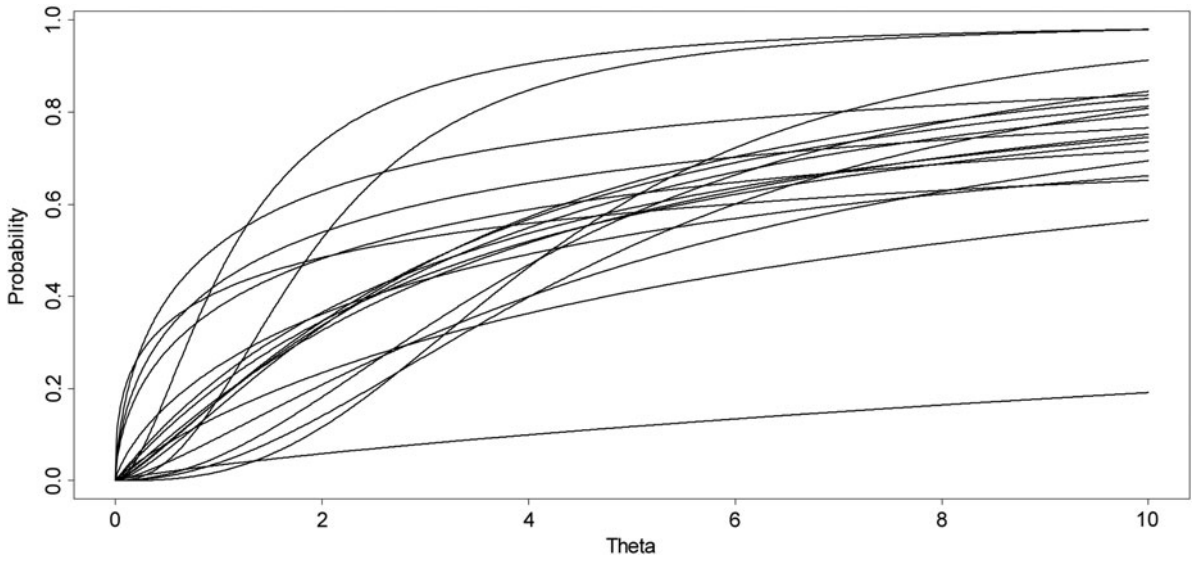


**Fig. 7.** Item response curves for the EY19 items in the log-logistic model.

Observe that some evidence of the quasi-trait nature of impulsivity is to be found in the present results. First, in the EY19, and in other self-report measures we have examined, one cannot find a good range of item location parameters. We note that this concern has been raised by Reise & Waller (2009) as being generally true in psychopathology measurement. Second, and related to the first, is that scores on the EY19 are not normally distributed, either raw or latent trait estimates, and are, in fact, highly positively skewed. If one takes the next step and entertains the idea of a unipolar trait of impulsivity, then one analytic approach that may be taken is to fit a zero-inflated mixture IRT model (Wall *et al.* 2015).

In this type of model, a latent class is defined to be a 'no trait'/'no symptoms' or 'normal' group and the percentage of such cases in the population is estimated. Then, assuming normality, one or more latent classes are identified to represent 'traited', 'symptomatic', or 'clinical' groups, and IRT item parameters are estimated based only on this latter latent class. This model may be ideal for the assessment of constructs where a researcher believes that low and zero scores are essentially meaningless, while variation in higher scores is meaningful and valid. This model, however, was developed for data where there are many more zero scores than observed in the EY19 data, and thus we will not illustrate an application.

Alternatively, in the presence of high skew, instead of assuming normality, a more reasonable latent trait distribution may be the log-normal (where zero is the lowest latent trait score possible, low scores reflect an absence of the trait and high scores reflect severity of pathology relative to absence of the disorder). Moreover, a more appropriate IRT model may be the log-logistic instead of the logistic (see Lucke, 2014, 2015, for technical details). The log-logistic model is shown in equation (5).

$$P(x = 1|\theta) = \frac{\lambda\theta^{\eta}}{1 + \lambda\theta^{\eta}}, \tag{5}$$

where $\lambda$ is a multiplicative parameter with higher values shifting the IRC to the right, and $\eta$ is an item discrimination parameter.

This model can be estimated using Bayesian methods as described in Lucke's research, but for simplicity, observe that 'naive' parameters can be found as simple transformations of the parameters from the 2PL model. Specifically, $\eta = \alpha$ and $\lambda = \exp(-\alpha\beta)$; theta estimates in the log-logistic can be found as $\exp(\theta_{2PL}) = \theta_{LL}$. The item parameter values are shown in Table 3 and the corresponding IRCs for the 19 items are shown in Fig. 7. Compared to a logistic model [equation (1)], in the log-logistic [equation (5)] individual differences at the high end of the trait are expanded and individual differences at the low end are contracted. This means that for individuals scoring low on the measure, theta estimates are rather homogeneous and near zero. For individual's scoring relatively high on the measure, theta estimates are more spread out.

At this point, we are not prepared to argue one way or another in regards to the correct latent variable measurement model for impulsivity, or other psychiatric constructs. This modeling issue, however, is important to raise here not only to motivate researchers to consider the continuous *v.* quasi-continuous nature of constructs, but also to point out that typically IRT modeling and associated applications may not work well in particular construct domains. This may be especially true in psychiatric domains where the construct

is unipolar, the latent variable highly positive skewed, and the range of possible trait indicators very narrow.

## Acknowledgements

This work was supported by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580 (PI: Robert Bilder), and RL1DA024853 (PI: Edythe London). Additional research support was obtained through the National Institutes of Health the NIH Roadmap for Medical Research Grant (AR052177; PI: David Cella).

## Declaration of Interest

None.

## Notes

[1] There are also multidimensional IRT models (Reckase, 2009), but these are beyond the present scope.

[2] For this sample, the items were administered in random order as part of a much larger battery, thus there is no possibility of item order effects here.

[3] Note that the location parameter in the 3PL, 3PLU, and 3PL models is no longer interpreted as the point on the latent variable where the probability of endorsement is 0.50.

## References

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

**Cai L** (2013). *flexMIRT: a Numerical Engine for Flexible Multilevel Multidimensional Item Analysis and Test Scoring* (*Version 2.0*) (*Computer software*). Vector Psychometric Group: Chapel Hill, NC.

**Cai L, Thissen D, du Toit SHC** (2011). *IRTPRO for Windows* (*Computer software*). Scientific Software International: Lincolnwood, IL.

**Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R** (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology* **63**, 1179–1194.

**Chalmers RP, Pritikin J, Robitzsch A, Zoltak M** (2015). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software* **48**, 1–29.

**Chen WH, Thissen D** (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics* **22**, 265–289.

**Choi SW** (2009). Firestar: Computerized Adaptive Testing simulation program for polytomous item response theory models. *Applied Psychological Measurement* **33**, 644–645.

**Choi SW, Schalet B, Cook KF, Cella D** (2014). Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment* **26**, 513–527.

**Embretson SE, Reise SP** (2000). *Item Response Theory for Psychologists*, Mahwah, NJ: Erlbaum.

**Eysenck SB, Eysenck HJ** (1978). Impulsiveness and Venturesomeness: their position in a dimensional system of personality description. *Psychological Reports* **43**, 1247–1255.

**Fayers PM, Hand DJ** (2002). Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*) **165**, 233–253.

**Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ** (2013). NIH Toolbox for assessment of neurological and behavioral function. *Neurology* **80**, S2–S6.

**Han KCT, Paek I** (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement* **13**, 1–13.

**Horn NR, Dolan M, Elliott R, Deakin JFW, Woodruff PWR** (2003). Response inhibition and impulsivity: an fMRI study. *Neuropsychologia* **41**, 1959–1966.

**Lucke JF** (2014). Positive trait item response models. In *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting* (ed. R. E. Millsap L. A. van der Ark, D. M. Bolt and C. M. Woods), pp. 199–213. New York: Springer.

**Lucke JF** (2015). Unipolar item response models. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (ed. S. P. Reise and D. Revicki), pp. 272–284. Routledge: New York.

**Maydeu-Olivares A, Cai L, Hernández A** (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: a Multidisciplinary Journal* **18**, 333–356.

**Meijer RR, Tendeiro JN, Wanders RBK** (2015). The use of nonparametric item response theory to explore data quality. In *Handbook of Item response Theory Modeling: Applications to Typical Performance Assessment* (ed. S. P. Reise and D. A. Revicki), pp. 85–110. Routledge: London, England.

**Molenaar D** (2014). Heteroscedastic latent trait models for dichotomous data. *Psychometrika* 625–644.

**Monroe S, Cai L** (2014). Estimation of a Ramsay-curve item response model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement* **74**, 343–369.

**Mungas D, Reed BR, Marshall SC, González HM** (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology* **14**, 209–223.

**Muthén LK, Muthén BO** (2012). Mplus. *The Comprehensive Modelling Program for Applied Researchers: User's Guide, 5*.

**Orlando M, Thissen D** (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement* **24**, 50–64.

**Orlando M, Thissen D** (2003). Further investigation of the performance of S-X2: an item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* **27**, 289–298.

**Partchev I** (2015). Package 'irtoys': simple interface to the estimation and plotting of IRT models (https://cran.rproject.org/web/packages/irtoys/irtoys.pdf).

**R Development Core Team** (2015). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria (http://www.R-project.org).

**Reckase M** (2009). *Multidimensional Item Response Theory*. Springer: New York.

**Reise SP, Revicki D** (2015). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Routledge: New York.

**Reise SP, Waller NG** (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods* **8**, 164–184.

**Reise SP, Waller NG** (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology* **5**, 27–48.

**Rizopoulos D** (2015) ltm: latent trait models under IRT (https://cran.r-project.org/web/packages/ltm/ltm.pdf).

**Schalet BD, Cook KF, Choi SW, Cella D** (2014). Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders* **28**, 88–96.

**Steinberg L, Thissen D** (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods* **1**, 181–197.

**Tavares HR, Andrade DFD, Pereira CADB** (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology* **27**, 679–685.

**Thomas ML** (2011). The value of item response theory in clinical assessment: a review. *Assessment* **18**, 291–307.

**Van der Ark LA** (2014). New developments in Mokken scale analysis in R. *Journal of Statistical Software* **48**, 1–27.

**Wall MM, Park JY, Moustaki I** (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement* **39**, 583–597.

**Woods CM** (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods* **11**, 253–270.

**Wu EJ, Bentler PM** (2011). *EQSIRT: a User-friendly IRT Program* (*computer software*). Multivariate Software: Encino, CA.

**Xu MK, Gaysina D, Barnett JH, Scoriels L, van de Lagemaat LN, Wong A, Jones PB** (2015). Psychometric precision in phenotype definition is a useful step in molecular genetic investigation of psychiatric disorders. *Translational Psychiatry* **5**, e593.

**Yang FM, Kao ST** (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry* **26**, 171–177.