Title

Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures.

Permalink

https://escholarship.org/uc/item/2jj51001

Journal

Journal of personality assessment, 100(4)

ISSN

0022-3891

Authors

Reise, Steven P
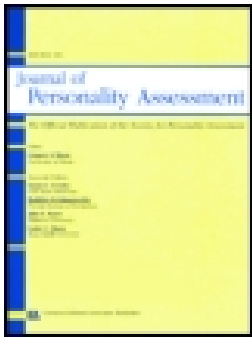Rodriguez, Anthony
Spritzer, Karen L
et al.

Publication Date

2018-07-01

DOI

10.1080/00223891.2017.1381969

Peer reviewed

Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures

Steven P. Reise, Anthony Rodriguez, Karen L. Spritzer & Ron D. Hays

⊞ View supplementary material ↗

🗓 Published online: 31 Oct 2017.

✎ Submit your article to this journal ↗

🔍 View related articles ↗

View Crossmark data ↗

Routledge
Taylor & Francis Group

STATISTICAL DEVELOPMENT AND APPLICATIONS

Check for updates

Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures

Steven P. Reise,[1] Anthony Rodriguez,[1] Karen L. Spritzer,[2] and Ron D. Hays[2]

[1]Department of Psychology, University of California, Los Angeles; [2]Division of General Internal Medicine & Health Services Research, University of California, Los Angeles

ABSTRACT

It is generally assumed that the latent trait is normally distributed in the population when estimating logistic item response theory (IRT) model parameters. This assumption requires that the latent trait be fully continuous and the population homogenous (i.e., not a mixture). When this normality assumption is violated, models are misspecified, and item and person parameter estimates are inaccurate. When normality cannot be assumed, it might be appropriate to consider alternative modeling approaches: (a) a zero-inflated mixture, (b) a log-logistic, (c) a Ramsay curve, or (d) a heteroskedastic-skew model. The first 2 models were developed to address modeling problems associated with so-called quasi-continuous or unipolar constructs, which apply only to a subset of the population, or are meaningful at one end of the continuum only. The second 2 models were developed to address non-normal latent trait distributions and violations of homogeneity of error variance, respectively. To introduce these alternative IRT models and illustrate their strengths and weaknesses, we performed real data application comparing results to those from a graded response model. We review both statistical and theoretical challenges in applying these models and choosing among them. Future applications of these and other alternative models (e.g., unfolding, diffusion) are needed to advance understanding about model choice in particular situations.

Well-known assumptions of unidimensional item response theory (IRT) models are unidimensionality, local independence, and monotonicity. When estimating item parameters using full-information maximum likelihood, it is commonly assumed that the underlying latent trait is normally distributed in the population. In specifying a normal distribution, it is implicitly assumed that the latent variable scale and the estimated item parameters apply to everyone in the calibration population (i.e., there is no mixture). Further, it is assumed that the latent variable is a continuous "bipolar trait" that has substantively meaningful variation across the range of the latent variable (Lucke, 2015, p. 273).

When the normality assumption, or its subsidiary assumptions, are violated, parameter estimates can be highly inaccurate (Azevedo, Bolfarine, & Andrade, 2011; DeMars, 2012; Kirischi, Hsu, & Yu, 2001; Sass, Schmitt, & Walker, 2008; Seong, 1990; Wall, Park, & Moustaki, 2015). Although reviewing the extensive psychometric literature on the effects of normality violations on IRT item parameter and person estimates is beyond the scope of this article, Woods and Thissen (2006) nicely summarized the consequences of non-normality: "There is fairly consistent evidence that, when normality of $g(\theta)$ is assumed, MML estimates of more extreme item parameters (e.g., thresholds around ±2) are nontrivially biased when the true population distribution is

platykurtic or skewed, and if $g(\theta)$ is skewed, the bias increases as the skewness increases" (p. 283).

Monroe and Cai (2014, p. 365) provided a compelling example of the negative consequences of a misspecified normal distribution in the context of a drug abuse treatment outcome studies measure of mental health and emotional distress.

Non-normal latent trait distributions present particular challenges in the application of standard logistic IRT models to personality and psychopathology measures (Reise & Rodriguez, 2016) because it is arguable that for many traits in these domains (e.g., self-esteem [Gray-Little, Williams, & Hancock, 1997], borderline personality disorder [Michonski, Sharp, Steinberg, & Zanarini, 2013], or dark triad traits [Webster & Jonason, 2013]), an assumed normal distribution in a general population might be untenable. In such cases, researchers need to consider alternative IRT models designed to estimate non-normal distributions.

Herein, we describe the strengths and limitations of two such approaches, one nonparametric and the other parametric (i.e., assumes a particular distributional form). Specifically, we review a Ramsay curve model (Woods & Thissen, 2006) that estimates the shape of the latent trait distribution simultaneously with the estimation of the item parameters. We also review a heteroskedastic-skew model (Molenaar, Dolan, & de Boeck, 2012) that both estimates the skewness of the latent trait

CONTACT Steven P. Reise ✉ reise@psych.ucla.edu 📍 Department of Psychology, University of California, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095.

and allows for error variances that increase or decrease as a function of the latent trait. Using a real data set, the results of these models will be compared with the results under a normality assumption.

A skewed latent trait distribution is one way that personality and psychopathology data deviate from the normality assumption. An additional complexity is that some personality and psychopathology constructs are not fully continuous with meaningful individual difference variation across the full range of the latent trait continuum. Although some constructs such as extraversion (vs. introversion), conscientiousness (vs. irresponsibility), and subjective well-being (vs. subjective distress) are, arguably, continuous and bipolar, other important constructs such as substance use or abuse, agoraphobia, and somatic complaints, are not bipolar or fully continuous. We argue that for such constructs, one would not expect a normal distribution; more likely would be a highly skewed or a half-normal distribution.

The challenges that certain personality and psychopathology constructs present for IRT modeling have been long noted. Almost 30 years ago, Reise and Waller (1990) stated that some "personality traits may have an inherently quasi-categorical rather than a full range continuum structure" (p. 57). Observing that clinical assessment instruments have highly peaked information functions in the high (pathological) trait range and a notable lack of items that provide discriminations among individuals in low trait ranges, Reise and Waller (2007) stated, "we believe that the peaked information function for many clinical scales reflects the quasi-trait status of many psychopathology constructs. By the term 'quasi-trait,' we mean that the trait is unipolar (relevant only in one direction) and that variation at the low end of the scale is less informative in both a substantive as well as a psychometric sense" (p. 31).

Reise and Waller (1990, 2007) merged the concepts of unipolar trait and quasi-continuous trait (as opposed to fully continuous) to reference certain types of constructs that are potentially problematic when fitting IRT models using a normality assumption. Here we use the term *quasi-trait* to refer to constructs such as positive psychotic symptoms, where low scores on symptom ratings reflect the absence or irrelevance of the disorder for the individual. We use the term *unipolar trait* to refer to constructs that are most substantively meaningful at one end of the continuum (e.g., alienation, aggression).

Admittedly, the distinction between a quasi (apply only to a subset of the population) and unipolar (only meaningful at one end of the continuum) construct is murky in practice, but needs to be drawn here because recently introduced IRT models were designed explicitly to handle these two types of measurement situations.

Specifically, we review two alternative IRT modeling approaches potentially applicable for quasi- and unipolar traits. The first is a zero-inflated mixture model (Wall et al., 2015) designed to handle the IRT modeling of quasi-traits—when the population is heterogeneous and the continuous trait is only applicable for a subset of the population. This model treats zero and near-zero scores as a distinct latent class and then estimates IRT item parameters with a normality assumption only for a "traited class." We also review a log-logistic model (Lucke, 2015) explicitly designed to handle unipolar traits—traits that

are not fully continuous and are only substantively meaningful at one end of the trait continuum.

In what follows, we review emerging IRT models that might be viable alternatives when the normality assumption for the latent trait in IRT is implausible, either because the latent trait distribution is suspected to be skewed, or the construct is not fully continuous or a unipolar trait. For comparison purposes, we fit a logistic graded response model (GRM; Samejima, 1969) to 29 items from the Patient Reported Outcomes Measurement Information Systems (PROMIS) Anger scale (Pilkonis et al., 2011). We then compare this "business-as-usual" analysis with the four alternative models cited earlier.

Each of the four alternative models makes different assumptions about the origin of the normality violation. Our specific goal in each comparison is not only to highlight the strengths and limitations of alternative modeling approaches, but also to demonstrate how the models might yield different substantive results, or not, in this particular data set. Our overarching goals are to raise awareness of alternative IRT models, as well as to motivate researchers to think more critically about latent distributions.

Example data and psychometric characteristics

The calibration sample consisted of 1,498 nonclinical adults who responded to 29 items administered in the development of the PROMIS Anger measure (content available in Supplemental Materials, Table 1). Anger is a historically important construct in normal range personality, health outcomes, and psychopathology research, and is one of three constructs relevant to negative affect available through the PROMIS (Pilkonis et al., 2011) initiative. As such it has been extensively evaluated[1] (e.g., Shalet et al., 2016). This measure was selected because it is well suited for illustrating the strengths and limitations of the alternative models presented here. We also believe the data are representative of typical IRT applications to clinical measures as reviewed in Reise and Waller (2007): unipolar, high skew for summed scores, and few if any items providing discrimination in the low trait range (see analyses that follow).

The first set of columns in Table 1 displays the percentage of responses in each category (0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often*, and 4 = *always*).[2] Relatively few individuals respond in the two extreme categories, and for many items, approximately 50% are responding 0 (*never*). The last set of columns displays the item–scale (minus the item) correlations, item means, and standard deviations. When summed scores are calculated, coefficient alpha = .97, $M = 25.55$, $SD = 21.22$, skewness = 0.96, and kurtosis = .50.

As seen in Figure 1, the distribution of summed scores appears to be more of a truncated-normal or half-normal distribution with zero inflation; best fitting skewed normal (solid line) and normal distributions (long dashed line) are superimposed for illustrative purposes. Although the distribution of

---

[1]See healthmeasures.net.
[2]Item 29 used a different response format: *not at all, a little bit, somewhat, quite a bit*, and *very much*.

Table 1. Category response percentages and descriptive statistics for the PROMIS Anger scale.

	Scale item	Response percentages					Descriptives		
		0	1	2	3	4	r_{it}	M	S
Item 1	Ang01	0.28	0.32	0.29	0.09	0.02	0.62	1.25	1.03
Item 2	Ang03	0.36	0.27	0.26	0.09	0.03	0.74	1.14	1.08
Item 3	Ang04	0.57	0.24	0.14	0.04	0.01	0.64	0.68	0.94
Item 4	Ang05	0.20	0.30	0.41	0.08	0.01	0.61	1.40	0.93
Item 5	Ang06	0.45	0.27	0.20	0.06	0.02	0.78	0.93	1.02
Item 6	Ang07	0.66	0.19	0.10	0.04	0.01	0.75	0.55	0.91
Item 7	Ang09	0.28	0.36	0.28	0.07	0.01	0.79	1.17	0.95
Item 8	Ang10	0.44	0.25	0.20	0.08	0.02	0.68	0.98	1.07
Item 9	Ang11	0.72	0.13	0.10	0.04	0.01	0.75	0.50	0.91
Item 10	Ang15	0.56	0.23	0.14	0.06	0.01	0.82	0.74	0.99
Item 11	Ang16	0.45	0.27	0.20	0.06	0.02	0.78	0.92	1.01
Item 12	Ang17	0.52	0.27	0.15	0.05	0.01	0.76	0.76	0.95
Item 13	Ang18	0.49	0.23	0.20	0.06	0.02	0.71	0.89	1.06
Item 14	Ang21	0.47	0.27	0.18	0.07	0.01	0.79	0.88	1.00
Item 15	Ang22	0.60	0.19	0.16	0.04	0.01	0.75	0.66	0.93
Item 16	Ang25	0.57	0.24	0.13	0.05	0.01	0.81	0.68	0.93
Item 17	Ang26	0.53	0.24	0.16	0.05	0.02	0.76	0.77	0.99
Item 18	Ang28	0.49	0.26	0.18	0.06	0.02	0.81	0.84	1.01
Item 19	Ang30	0.26	0.37	0.29	0.07	0.01	0.75	1.21	0.95
Item 20	Ang31	0.40	0.33	0.21	0.06	0.01	0.72	0.94	0.95
Item 21	Ang35	0.20	0.36	0.34	0.09	0.01	0.72	1.35	0.93
Item 22	Ang37	0.46	0.29	0.18	0.06	0.01	0.82	0.86	0.97
Item 23	Ang42	0.57	0.25	0.13	0.04	0.01	0.80	0.68	0.93
Item 24	Ang45	0.41	0.28	0.22	0.07	0.02	0.75	0.99	1.02
Item 25	Ang47	0.47	0.27	0.19	0.06	0.02	0.80	0.90	1.02
Item 26	Ang48	0.70	0.14	0.11	0.04	0.01	0.77	0.51	0.91
Item 27	Ang54	0.44	0.30	0.20	0.06	0.01	0.78	0.92	1.00
Item 28	Ang55	0.43	0.28	0.21	0.06	0.02	0.79	0.95	1.02
Item 29	Ang56	0.72	0.15	0.09	0.03	0.01	0.70	0.48	0.88

Note. Response options include 0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often*, 4 = *always*. Item–test correlation denoted (r_{it}). For raw scores, M = 25.55, SD = 21.22, skewness = 0.96, and kurtosis = 0.5.

summed scores is not necessarily a good indicator of the distribution of the latent trait,[3] it is still critically important to inspect it for observed non-normality in the data. Is the latent distribution really normal but observed skew is due to "faulty" item construction or is the latent distribution skewed, possibly with zero inflation? Are excess zeros caused by poor sampling, or is this attributable to a unipolar or quasi-trait? How a researcher answers these questions affects model choice.

The graded response model

The GRM (Samejima, 1969) is commonly employed in the IRT modeling of personality, psychopathology, and health outcomes data. This model has well-known relations with the parameters resulting from item-level ordinal factor analysis (i.e., factor loadings and intercepts). In fact, some IRT software packages report both the factor analytic and IRT parameterization alongside each other in standard output. Herein, we consider the GRM as the "default" or "business-as-usual" model. In estimating the GRM, it is commonly assumed that there exists an underlying normally distributed latent variable (trait) with meaningful variation across the full range of the trait continuum.
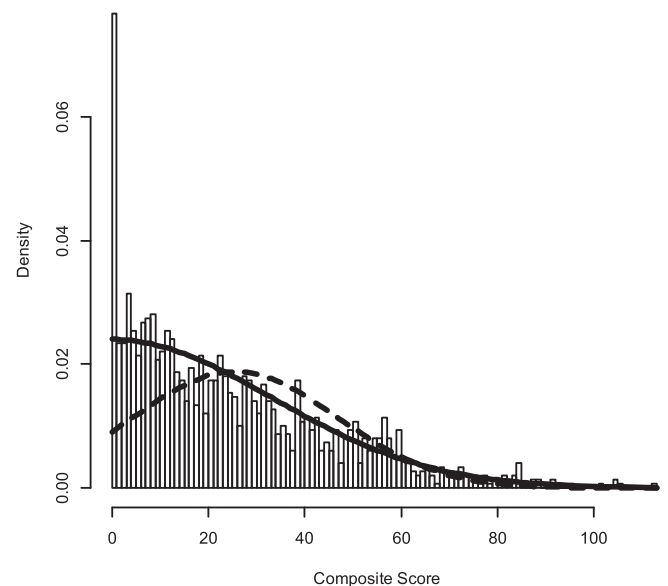

Figure 1. Histogram of composite scores for Anger scale with best fitting skewed normal (solid line) and normal distribution (dashed line).

In the GRM for each item (i), one slope parameter (α_i) is estimated. Items with larger slope parameters are considered more discriminating or informative. For each item, $K - 1$ intercept parameters $\gamma_{i(j=1...K-1)}$ are also estimated where K is the number of response options. These intercepts are then transformed into $K - 1$ location parameters $\beta_{i(j=1...K-1)}$, where $\beta_{ij} = -\frac{\gamma_{ij}}{\alpha_{ij}}$. Thus, the location parameter is not independent of

---

[3]The observed summed score distribution is a function of both the true latent trait distribution and the properties of the items (Lord, 1953). Thus, even if the true distribution is normal, if a test was either too easy or difficult, or if the location parameters are not symmetric around zero, the observed scores will be skewed.

Table 2. Item slope, location, and intercept parameter estimates under the graded response model, assuming a normally distributed latent trait.

	Slope	Location				Intercept			
	α	β_1	β_2	β_3	β_4	γ_1	γ_2	γ_3	γ_4
Item 1	1.71	−0.77	0.40	1.72	2.98	1.31	−0.69	−2.94	−5.09
Item 2	2.25	−0.40	0.46	1.52	2.48	0.90	−1.03	−3.42	−5.58
Item 3	1.80	0.26	1.18	2.21	3.26	−0.47	−2.12	−3.98	−5.86
Item 4	1.57	−1.21	0.04	1.96	3.60	1.90	−0.06	−3.08	−5.65
Item 5	2.65	−0.13	0.70	1.69	2.55	0.34	−1.86	−4.48	−6.75
Item 6	2.92	0.47	1.16	1.81	2.70	−1.38	−3.39	−5.28	−7.87
Item 7	2.85	−0.66	0.42	1.59	2.75	1.87	−1.20	−4.53	−7.86
Item 8	1.97	−0.17	0.69	1.68	2.77	0.34	−1.36	−3.32	−5.46
Item 9	3.14	0.65	1.13	1.77	2.65	−2.03	−3.55	−5.56	−8.32
Item 10	3.35	0.18	0.89	1.62	2.52	−0.61	−2.97	−5.42	−8.46
Item 11	2.76	−0.12	0.71	1.67	2.53	0.33	−1.95	−4.61	−6.98
Item 12	2.60	0.08	0.94	1.85	2.79	−0.20	−2.45	−4.81	−7.25
Item 13	2.25	0.01	0.76	1.76	2.52	−0.02	−1.72	−3.96	−5.67
Item 14	2.75	−0.05	0.75	1.63	2.84	0.14	−2.06	−4.48	−7.80
Item 15	2.62	0.32	0.97	1.97	2.94	−0.83	−2.54	−5.16	−7.69
Item 16	3.30	0.21	0.98	1.75	2.69	−0.70	−3.23	−5.78	−8.88
Item 17	2.62	0.13	0.90	1.82	2.55	−0.33	−2.34	−4.76	−6.67
Item 18	3.23	0.01	0.76	1.64	2.42	−0.03	−2.47	−5.30	−7.83
Item 19	2.49	−0.76	0.39	1.62	2.75	1.88	−0.98	−4.04	−6.86
Item 20	2.20	−0.28	0.76	1.90	3.11	0.62	−1.67	−4.17	−6.82
Item 21	2.19	−1.05	0.20	1.60	2.96	2.30	−0.45	−3.51	−6.47
Item 22	3.21	−0.08	0.78	1.68	2.62	0.25	−2.51	−5.38	−8.42
Item 23	3.27	0.21	1.01	1.84	2.54	−0.68	−3.31	−6.03	−8.31
Item 24	2.38	−0.24	0.64	1.69	2.72	0.57	−1.52	−4.01	−6.45
Item 25	3.00	−0.07	0.72	1.62	2.47	0.21	−2.17	−4.87	−7.41
Item 26	3.33	0.59	1.11	1.84	2.54	−1.98	−3.69	−6.12	−8.46
Item 27	2.68	−0.16	0.73	1.71	2.61	0.43	−1.96	−4.57	−7.00
Item 28	2.79	−0.17	0.65	1.65	2.53	0.47	−1.82	−4.60	−7.05
Item 29	2.42	0.69	1.34	2.06	2.79	−1.66	−3.23	−4.97	−6.75
M	2.63	−0.09	0.76	1.75	2.73	0.10	−2.08	−4.59	−7.09
SD	0.50	0.47	0.30	0.15	0.26	1.12	0.95	0.86	1.04
M^a	2.17	−0.19	0.81	2.04	3.15				
SD^a	0.43	0.65	0.55	0.43	0.47				

Note. α = slope; β = location; γ = intercept.
[a]Pilkonis et al. (2011) results.

the slope parameter, and, consequently, highly discriminating items tend to have locations that are clustered around zero, and less discriminating items have locations more spread out.

Item parameter estimates for the Anger scale under the GRM are shown in Table 2. These parameters were estimated using marginal maximum likelihood as implemented in Multidimensional Item Response Theory (MIRT; see Chalmers, 2012) assuming a normally distributed latent trait with a mean of 0 and variance of 1 in the population (for identification). The default number of quadrature nodes used in MIRT is 61 (range specified to be −4 to 4). For informational purposes, the log-likelihood of the model is −38135.64, Akaike's information criterion (AIC) = 76561.28, Bayesian information criterion (BIC) = 77331.5, root mean square error of approximation (RMSEA) = 0.04 (95% CI [0.038, 0.049]), standardized root mean square residual (SRMSR) = .040, and comparative fit index (CFI) = .989. The M_2 fit index (Maydeu-Olivares & Joe, 2006) is 1027.31 with 290 df, $p < .01$. Thus, judging by the practical fit indexes, the estimated parameters recover the data well, but the M_2 statistical index suggests that a closer examination of fit at the item level is needed.

To understand the parameters of the GRM, in Figure 2 we display three plots based on the results for Item 1. In the top panel is the log-odds of responding in and above Categories 1, 2, 3, and 4, respectively, as a function of the latent variable. This plot makes clear the interpretation of the $K − 1 = 4$ intercepts; they are the log-odds of responding in or above a category $j = 1, 2, 3,$ and 4, for individuals with trait levels of 0 (the mean). Moreover, the item slope parameter reflects the steepness of these functions; for Item 1 with a slope = 1.7, the log-odds of responding in the next highest category or above increases by a factor of 1.7 for a 1 SD unit change on the latent variable.

In the middle panel of Figure 2 is shown the more familiar threshold response curves (TRCs) for Item 1. These represent the probability of responding in and above categories $j = 1, 2, 3,$ and 4, respectively, as a function of the latent variable. The vertical dashed lines show that the location parameters represent the point on the latent variable continuum where the individual has a 50% chance of responding in and above a given category $j = 1 \ldots 4$. Finally, for the bottom panel, if we label the $K − 1$ TRCs as $TRC_1, TRC_2, TRC_3,$ and TRC_4, respectively, then the probability of responding in a particular category is 1 − TRC_1, $TRC_1 − TRC_2$, $TRC_2 − TRC_3$, $TRC_3 − TRC_4$, and $TRC_4 − 0$. These are called category response curves (CRCs). For any point on the latent variable, the sum of the CRCs equals 1.

Returning to Table 2, all PROMIS Anger items are highly discriminating, but there is a large range. Item 4 ("I disagreed with people") has a slope of 1.57, and Item 26 ("I felt like I needed help for my anger") has a slope of 3.33. Item 26 is about 4.5 times more informative or discriminating than Item 4 ($3.33^2/1.57^2$), or, it would take about 4.5 items like Item 4 to achieve the same precision as one Item 26.
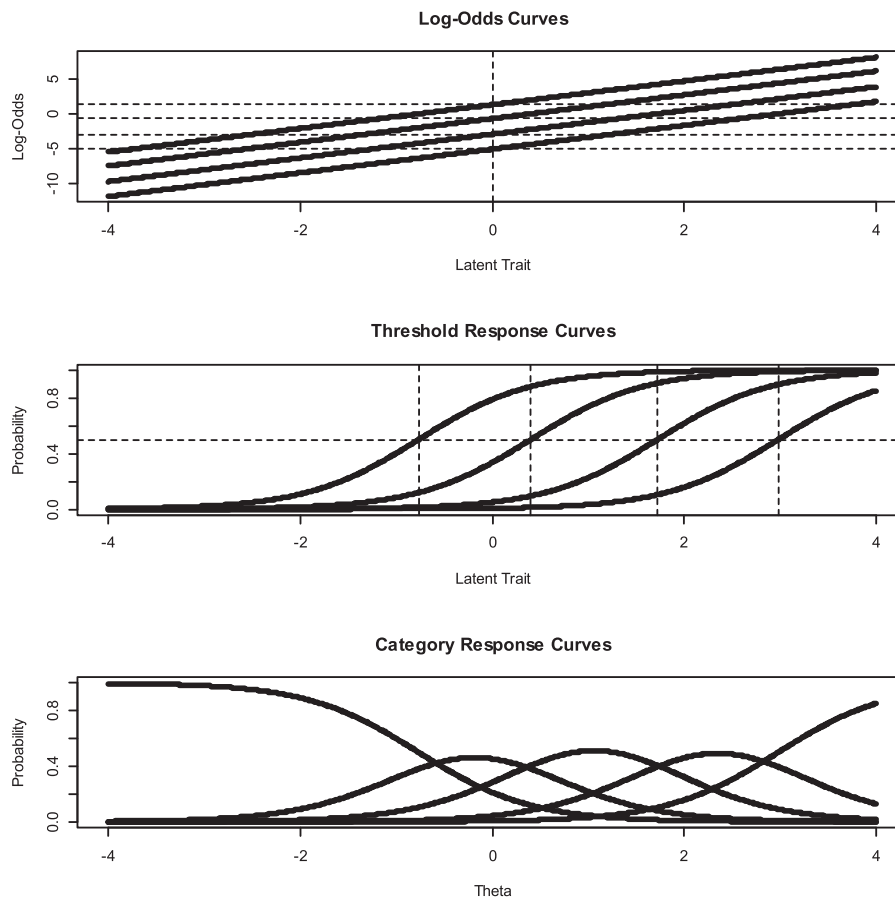
Log-Odds Curves

Threshold Response Curves

Category Response Curves

Figure 2. Log-odds, threshold response curves, and category response curves for PROMIS Anger Item 1 under the graded response model.

Despite there being five response options designed to spread measurement precision across the trait range, there are no items with location parameters below −1.21 (Item 4, the least discriminating item). With few exceptions, even the first location parameter tends to be around zero, or only slightly negative, suggesting that an individual has to be around the mean on the latent variable to have a 50% chance of responding in the second category or above. On the positive side of the theta continuum, the fourth location parameters are extreme, suggesting that to respond in the highest category, individuals must be around 2.5 to 3.6 *SD*s above the mean on the latent variable. Clearly, the Anger scale is a "peaked test," with location parameters clustered in the positive end of the continuum, and few, if any, items that discriminate best in the low trait range. This is the type of psychometric situation that Reise and Waller (1990, 2007) referred to in their discussion of quasi-continuous traits.

Models for quasi- and unipolar traits

The previously described GRM assumed a continuous, normally distributed latent variable. Nevertheless, the observed item and summed scores are highly skewed, and when a model was fit, threshold parameters were highly concentrated at the high end of the trait continuum. To justify this incongruity, one might attribute this to problems with the measure or the data, that is, faulty item construction (e.g., not enough response options, or, the anchors somehow are too extreme and thus cannot distinguish between low trait individuals), or

oversampling individuals from low trait ranges. Alternatively, we can reconceptualize the construct with a corresponding change in model and assumptions. In the following two sections, we consider two alternative modeling strategies.

A zero-inflated mixture model

Wall et al. (2015) developed a zero-inflated mixture (ZIM) IRT model in the context of psychiatric symptom measurement where it is common to find many individuals responding with zero or few symptoms. In other words, the model was developed explicitly to handle measurement situations where the construct is a so-called quasi-trait (a trait applicable to only a subset of the population) and the population is heterogeneous; for one pathological class of individuals, the trait is meaningful and symptoms can be used to scale individuals along a continuum, whereas for another untraited or nonpathological class, the construct is inapplicable. Wall et al. (2015) cited studies that documented the severe bias in IRT item parameter estimates, especially the discrimination parameter, when a normal distribution is assumed, but the data are zero inflated. In turn, the authors demonstrated that their ZIM model yields more accurate calibration.

The basic idea underlying the ZIM model is to estimate the percentages in the population that belong to the traited and untraited classes and then estimate the item parameters for the GRM in the traited class only. The latent trait is assumed to be normal in both classes, but one class is

degenerate; item parameters are not estimated for the degenerate class. This is similar to, but not exactly the same, as discarding cases scoring zero or near zero, and estimating item parameters based on the remaining cases. The authors argue that the mixture approach has a superior statistical justification because due to measurement error, some people with zero raw scores are likely in the traited class, whereas some people with nonzero raw scores likely belong to the untraited class.

We estimated a ZIM model using M*plus* (Muthén & Muthén, 2016) code supplied by the authors, with modification for these data. Results showed that the percentage of individuals estimated to be in the untraited class was 5.5% (a subset of the 6.0% of individuals with 0 raw total scores) and, thus, 94.5% was in the traited class. In comparison to GRM parameter estimates shown in Table 2, the slope parameters in the ZIM model for the traited class are much lower ($M = 1.94$ vs 2.63); controlling for the zero inflation leads to smaller slope parameters (see Supplemental Materials Table 2).

When test information curves are drawn that reflect the precision of measurement across the trait range, the test information functions for the GRM and mixture model are peaked at relatively high trait levels, but the mixture model, which is presumably more accurate, provides about half the information of the GRM (see Figure 3), and thus larger standard errors (which equal approximately 1 divided by the square root of test information). Finally, the Pearson correlation between latent trait estimates for the 1,406 individuals in the traited class from the mixture and GRM was .994. Thus, the models provide essentially the same relative ordering of individuals albeit with much larger standard errors in the ZIM model.

There are three important limitations of the mixture model applied in this context. First, the model assumes that the non-normality arises from a degenerate class, and once this degenerate class is removed from the calibration sample (down-weighted during estimation), the distribution is normal. The model, as presently implemented, does not allow for a skewed distribution to be applied after removing the untraited class. Thus, the parameters might still be biased due to an incorrect latent distribution.

Second, and resulting from the first limitation, one loses sample size because no meaningful latent trait estimates can be derived for the degenerate class. If this approach were to be applied in multivariate research where many constructs are measured, it is not at all clear how researchers are to proceed with such missing data when the data are missing as a consequence of construct irrelevance. Our third concern is purely substantive. As noted, the model was derived in the context of psychiatric constructs (alcohol use or abuse in particular) where the mixture formulation of nonpathological and pathological groups might make relatively more substantive sense. It is not clear to us what the interpretation of a class of "no anger" group would mean substantively, unless we view anger as measured in the PROMIS items as a pathological condition.

A unipolar log-logistic model

In the model presented in this section, a skewed distribution is treated as an inherent result of the measurement of unipolar constructs, especially for disorders such as alcohol, nicotine, and substance abuse, where it makes little sense to create a norm-referenced score. Lucke (2015) stated, "it makes little or no sense to assert that a person has a below-average level of addiction to alcohol or an above-average level of addiction to gambling." He then further argued that "The anchor for the scale should therefore be 'no disorder'" (p. 272).

To put these views into practice, Lucke (2015) proposed a log-logistic (LL) model for dichotomous item response data. This model was proposed in the context of unipolar traits and applied to a measure of gambling addiction. In the LL model, the latent trait begins at zero and continues to positive infinity. Response patterns of all zeros are assigned $\theta = 0$. A polytomous LL TRC for responding in or above category j ($j = 1 \ldots 4$) is then defined:

$$TRC_{ij} = P(x \geq j \,|\, \theta) = \frac{\lambda_{ij}\theta^{\eta_i}}{1 + \lambda_{ij}\theta^{\eta_i}}$$

CRCs are defined in the same way as the GRM: $\mathrm{CRC}_0 = 1 - \mathrm{TRC}_1$, $\mathrm{CRC}_1 = \mathrm{TRC}_1 - \mathrm{TRC}_2$; $\mathrm{CRC}_2 = \mathrm{TRC}_2 - \mathrm{TRC}_3$; $\mathrm{CRC}_3 = \mathrm{TRC}_3 - \mathrm{TRC}_4$; $\mathrm{CRC}_4 = \mathrm{TRC}_4 - 0$. Model parameters are defined as follows. The λ_{ij} parameters, $K - 1$ per item, are analogous to the intercept parameters in the GRM and have been referred to as easiness parameters. The λ parameters are always positive, and higher values signify that a large proportion of people respond in or below a given category. The η_i parameter, one per item, is a discrimination parameter and analogous to the slope in the GRM; in fact, it is exactly the same so items are no more or less relatively discriminating in either model. Finally, for each item, $K - 1$ location parameters, the point on the latent trait where the probability of responding in or above category $j = 1 \ldots K - 1$ is .50, is $\delta_{ij} = \left(\frac{1}{\lambda_{ij}}\right)^{\frac{1}{\eta_{ij}}}$.

The parameters of the dichotomous or polytomous log-logistic model can be estimated using Bayesian methods.
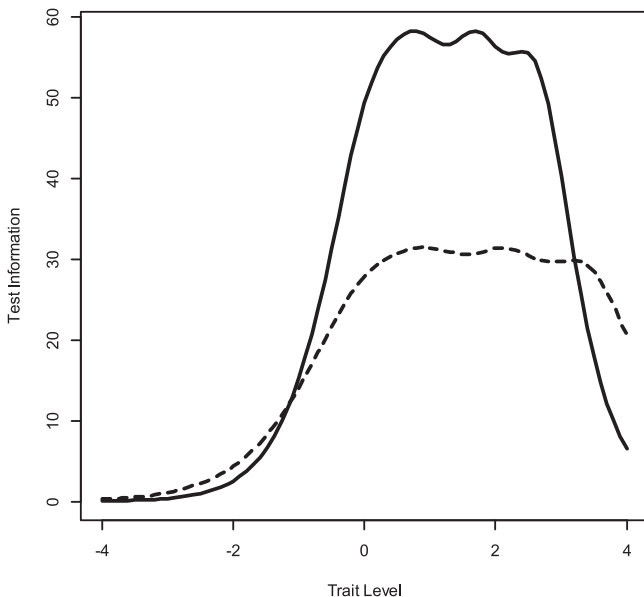

Figure 3. A comparison of test information for the graded response model (solid line) and the zero-inflated mixture model (dashed line).

However, for present purposes, we take advantage of the fact that logistic models and LL models are transformations of each other (e.g., if the latent variable in the GRM is normal with a mean of zero and standard deviation of 1, then in the LL model, the distribution is log-normal with the same mean and standard deviation).[4] Thus, using the parameters for the GRM shown in Table 2, we can transform to an LL model as follows: $\lambda = \exp(\gamma)$; $\eta = \alpha$; and $\delta = \exp(\beta)$. Latent trait scores estimated in the GRM can also be transformed to the LL metric as $\hat{\theta}_{LL} = \exp(\hat{\theta}_{GRM})$.

Item parameter estimates for the LL model are available in the Supplemental Materials Table 3. Understanding the difference between the GRM and LL rests on understanding the effects of the transformation of the latent scale. To clarify, what the LL model does is massively compress negative theta estimates and estimates that are around 0 in the GRM metric. For example consider that $\exp(-3.0) = 0.049$; $\exp(-2.0) = 0.135$; $\exp(-1.0) = 0.367$; $\exp(0) = 1.0$; $\exp(1.0) = 2.71$; $\exp(2.0) = 7.38$; and $\exp(3.0) = 20.08$. For this reason, theta estimates are very highly skewed in the LL model. In turn, this new metric has profound implications for the item and test information functions. Due to the metric "squeezing" at the low end and expansion at the high end, test information is very peaked and extremely high at the low end, as shown in Figure 4. The corresponding standard errors, also shown, indicate that the standard errors for a trait-level estimate change remarkably as a function of theta. The correlation between trait-level estimates in the GRM and LL is $r = .72$. Despite the fact that the two estimates are simple nonlinear tranforms of each other where the rank ordering remains the same, the correlation is far from perfect due to scale compression and expansion noted earlier.

In our view, the LL model has many virtues to recommend it in terms of the present Anger measure. Most important, it does not assume normality, it allows for the scoring of all individuals, and it appears consistent with theory if the researcher believes the construct to be a unipolar trait. On the other hand, this is a relatively recently proposed model, and much remains unknown (e.g., how to evaluate fit, test for differential item functioning, its robustness to zero inflation, etc.). Some of our concerns are practical (e.g., accuracy of parameter estimation under varying distribution conditions) and some are technical (e.g., the information function in this model has some implausible properties; it gets very high at low trait levels due to the "squeezing" effect of the transformation).

Models for non-normal latent traits

The preceding models treat the non-normality in the data as arising from very different mechanisms, zero inflation and the unipolar nature of the construct, respectively. In this section, we review two models that assume a continuous underlying
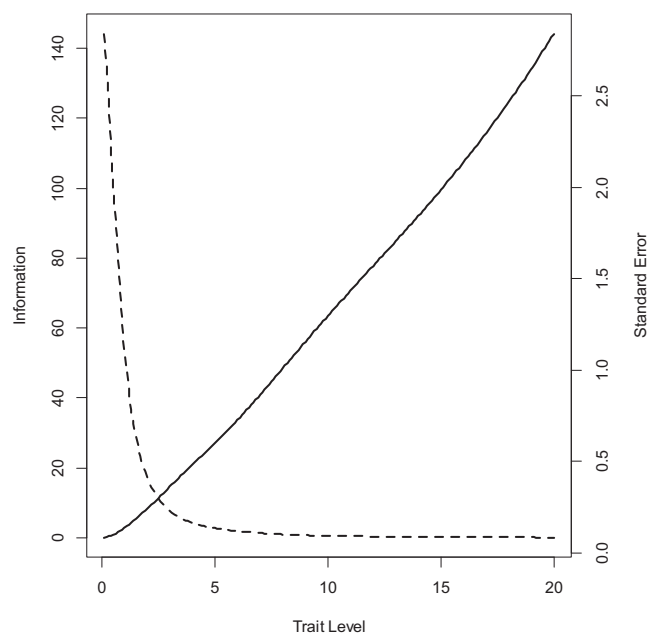

Figure 4. Test information (solid line) and standard errors (dashed line) in the log-logistic model.

latent distribution, but allow that distribution to be non-normal.[5]

Ramsay-curve IRT

When a normal latent trait distribution is (incorrectly) assumed during parameter estimation, this misspecification can lead to distorted item parameter estimates. One possible remedy to this problem is to estimate the shape of the latent trait distribution and then estimate parameters based on a correctly specified latent trait distribution. In Ramsay-curve (RC) IRT (Woods, 2006, 2015; Woods & Thissen, 2006), the latent trait distribution is estimated at the same time as the item parameters.

At its most basic level, the latent trait distribution in RC IRT is estimated using a smooth function to describe the density for the latent trait; this is a nonparametric technique, but there are limits on the type of distribution that can be reasonably approximated. To date, all research on parameter recovery under RC IRT relies on creating non-normal distributions through the mixture of normal distributions. We know of no studies of parameter recovery in the presence of zero inflation. In this analysis, we used RCLOG V2 (Woods, 2006) to estimate a latent density underlying the Anger items. There are several important user options in running RCLOG. For the sake of brevity, for technical details and suggested user defaults, we refer readers to the original research and user manual.

In this analysis, we allowed RC IRT to evaluate solutions running from 2 degrees and 2 knots (a normal distribution) to 6 degrees and 6 knots. Knots and degrees are technical jargon.

---

[4]With the caveat that if the parameters in the GRM are biased, their translation must be in error as well. This is why future work should consider the Bayesian estimation of the LL GRM where a researcher can have better control over prior distributions.

[5]In the original presentation of the Anger item bank (Pilkonis et al., 2011), the authors assumed a continous latent variable but acknowledged that the normality assumption for the calibration might be questionable. Nevertheless, they concluded that the effects of violations, if any, were minimal, and proceeded with a standard calibration with normality assumption.
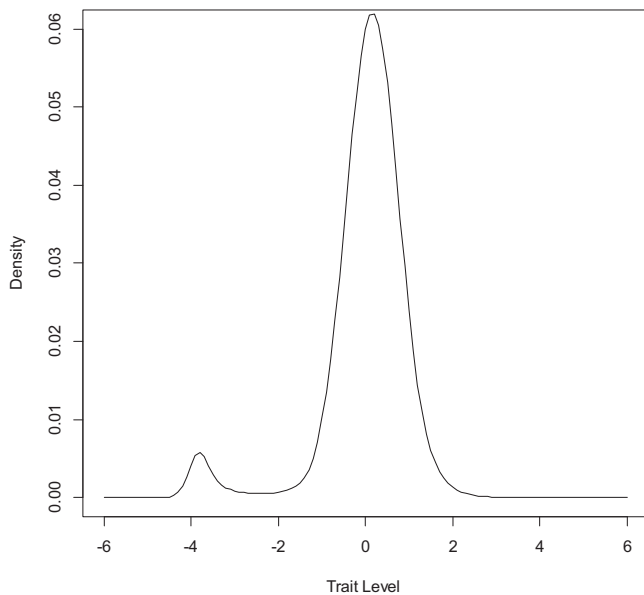
Figure 5. The latent density estimated by Ramsay curve logistic.

Degrees refers to the degrees of the polynomial for the Ramsay curves used to approximate the latent distribution. Knots refers to the number of joinings of the Ramsay curves. Together, degrees and knots refer to the flexibility of the possible distribution; fewer are more restricted with the limiting case of a normal distribution, whereas more allow for greater departures from non-normality (i.e., skew and kurtosis). Examination of fit indexes output from RCLOG suggested that the solution with 6 knots and 2 degrees was best, although alternative solutions were very close. In this solution, the log-likelihood was -38102.47, which differed significantly from a normal distribution, $\chi^2(4) = 67.11$. Most importantly, skewness of the latent distribution was estimated to be -1.96 and kurtosis was estimated to be 8.78. The specific distribution estimated is shown in Figure 5, which displays one small hump at low trait levels, and then an essentially normal distribution.

Item parameter estimates based on the best fitting model are available in Supplemental Materials Table 4. Comparing these values to the GRM, the glaring difference is that the slopes are all much higher in the RC model than GRM ($M = 3.75$ in RC and $M = 2.63$ in GRM), implying that the RC model yields more precise trait-level estimates. However, we caution that these parameter estimates might be misleading. The basis of our concern is the negative skew estimate and unusually high kurtosis. In particular, the negative skew estimate for data that are clearly positively skewed indicates that the estimation might be problematic due to the excess zero distribution or preponderance of people clustered around raw total scores of zero.[6] The correlation between trait level estimates in the GRM and RC models is .99.

---

[6]We note that Woods (2015) successfully implemented Davidian curves to Anger items drawn from the PROMIS project. However, there is currently no available software to implement this method so we could not explore that alterative to Ramsay curves here.
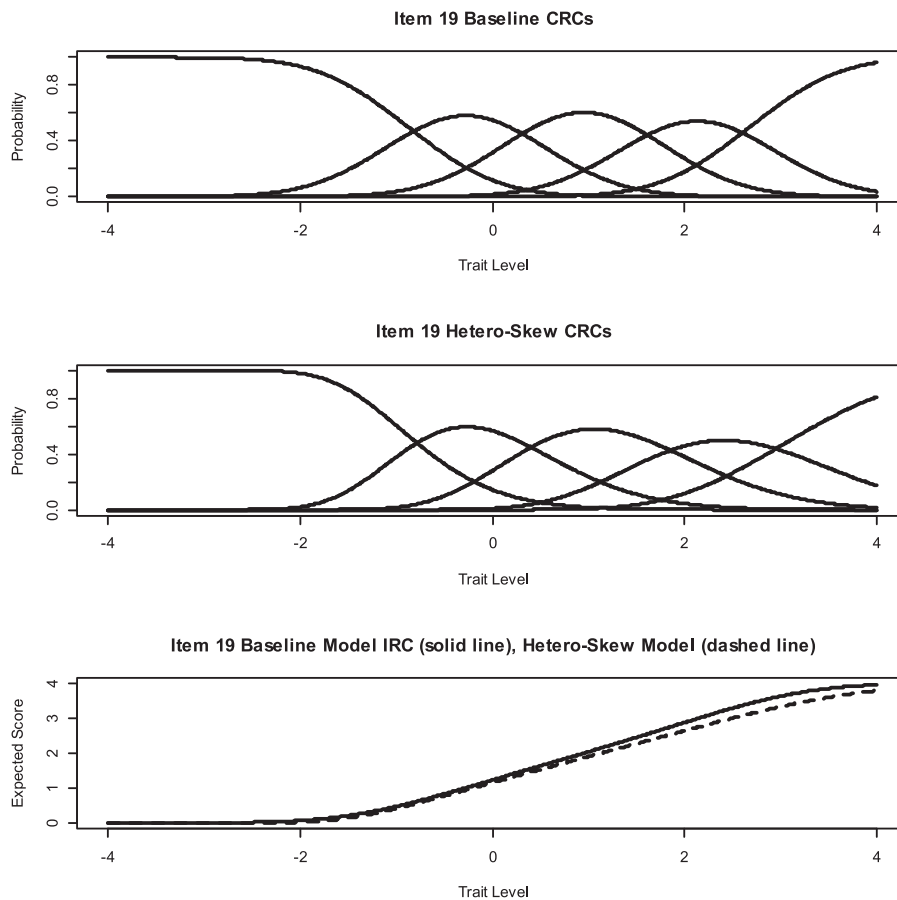
The heteroskedastic-skew graded response model

The preceding model attempts to estimate a nonparametric but smooth density function to represent the latent trait distribution. The model in this section, called the heteroskedastic-skew (HS) model (Molenaar et al., 2012), also attempts to estimate a non-normal density, but with a specific parametric form, namely, a skewed normal distribution (Azzalini, & Capatanio, 1999). In addition, the HS model also attempts to account for violations of homogeneity of variance, which is one possible, but seldom discussed source of observed non-normality in item response.

Three features of the HS model are critical to understand. First, the model is based on the normal-ogive version of the GRM. This makes the model akin to an item-level ordinal factor analytic model—a "factor loading and intercept" parameterization easily transformable into an IRT "slope and threshold" parameterization. For example, we can describe item functioning as:

$$y_i^* = \nu_i + \lambda_i\theta + \varepsilon_i$$

where y_i^* is a continuous normal response propensity that is "polytomized" through the K ordinal item response categories, ν_i is the item intercept (the expected score on y_i^* when $\theta = 0$), λ_i is the (unstandardized) factor loading (regression slope), and ε_i is a residual with variance, σ_ε^2. For each item, a linear regression is estimated with the latent variable as the predictor and a normally distributed latent response propensity as the outcome. In the preceding, error variances for each item are assumed to be homoskedastic, with an expected value of zero.

Second, the logistic GRM described previously imposes symmetric category response curves, which can lead to problems in scoring individuals on the latent variable (Samejima, 2000). By virtue of allowing for heteroskedastic errors, the HS model does not necessarily produce symmetric CRCs (see Molenaar et al., 2012, p. 473). Third, the developers of the HS model noted that observed skewness in the data can be caused by at least two factors: (a) the latent trait distribution could be skewed, and, (b) heteroskedasticity of residuals. In theory, the HS model allows for skewness to be estimated after separating out the effects of heteroskedastic error and, thus, can provide a "cleaner" estimate.

With this foundation in mind, the basic idea of the HS GRM is to simultaneously estimate: (a) the parameters of the normal-ogive GRM, (b) a heterogeneity parameter for each item that allows items to violate homoskedasticity, and (c) the skewness of the latent trait distribution. This latter estimate is based on assuming a parametric skewed-normal distribution. For specific details of the model, and conversion of factor analytic to IRT model parameters, we refer readers to Molenaar et al. (2012).

We estimated two nested models using an OpenMx (Boker et al., 2011) program provided by the model developers: (a) baseline model with no skew or heteroskedasticity, and a full model with both skew and heteroskedasticity estimated. To identify the model, we fixed the first two threshold parameters to their values estimated in the GRM. Item parameter estimates (i.e., factor loadings, thresholds, intercepts, residual variance,

Figure 6. Category response curves (CRCs) and item response curves (IRCs) for Item 19 under baseline model and heteroskedastic-skew model.

heteroskedastic residuals, and IRT slope) for the baseline and full models are available in Supplemental Materials Tables 5 and 6.

Most important is the comparison of the baseline model with the full model. For the baseline model (skewness = 0, heteroskedasticity = 0), model fit indexes were −2 the log-likelihood = 76733.50, df = 43297, AIC = −9860.50, and BIC = −119924.62. For the full model, fit indexes were −2 the log-likelihood = 76610.57, df = 43267, AIC = −9923.43, and BIC = −119876.40. The chi-square difference between the baseline and full model was 122.93 on 30 df, which is significant at $p \leq$ .01. The estimated skewness of the latent trait was 0.28, which is essentially indistinguishable from a normal distribution. The factor loadings ($M = 0.96$ vs. $M = 0.92$) and IRT slopes ($M = 2.22$ vs. $M = 2.05$) are slightly lower in the full than the baseline model. All other item parameter estimates in the full model are essentially the same as the baseline, with the exception of the heteroskedasticity parameters (δ_1), which are now estimated.

The critical issue with these parameters is whether they are of sufficient magnitude to impact the CRCs relative to the baseline model. To explore this issue, in the top two panels of Figure 6 are displayed the CRCs for Item 19 under the baseline and full models. This item had a large positive heteroskedasticity parameter (0.48). It appears that the CRCs under the two models are nearly identical. In the bottom panel, we compare the item response curves for Item 19 under the two models; they are nearly overlapping except in the high trait ranges, where the

expected item scores are lower for the full model. In Figure 7, we provide the comparison of test response curves (expected summed score as a function of the latent trait) under the two models. As one would predict from the slightly lower factor loadings in the full model, expected scores are slightly lower in the full model. This is likely a difference that makes no practical
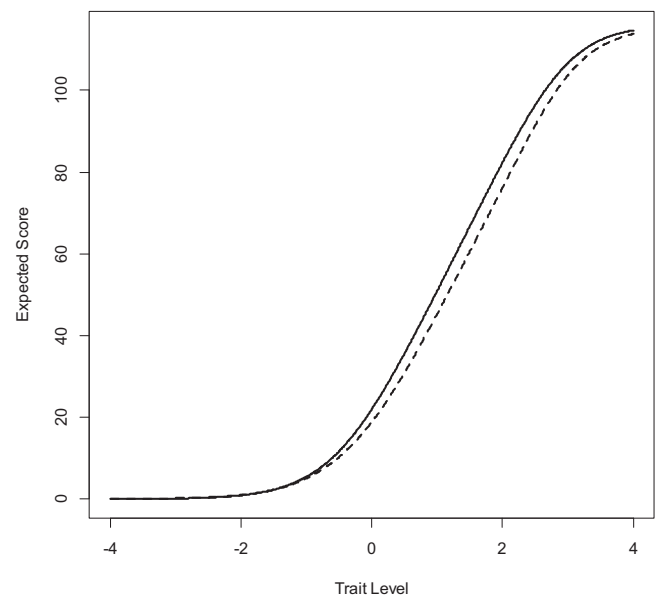

Figure 7. Test response curves for baseline (solid line) and heteroskedastic-skew (dashed line) models.

difference. Note that trial-level estimates from the HS models were not actually estimated due to software limitations, but given the similarity in item parameters with the GRM it is reasonable to expect that their correlation with the GRM would be near 1.0.

In review, although the HS full model with estimated skewed distribution is statistically superior to the baseline GRM with normal distribution, just as in the RC-IRT analysis, we question whether the present data are consistent with the estimated latent distribution. In RC-IRT the distribution is smooth but nonparametric, and thus avoids misspecifications caused by assuming a specific form. In the HS model, what is estimated is the skew of a parametric skewed normal distribution. In this regard, it is critically important to recognize that the limit of this distribution is a skewness of 1.0. As a consequence, no matter how skewed the true latent distribution is, the model can only accommodate that skew up to a certain point. Clearly, more research is needed to clarify parameter estimates under the HS model with unipolar traits, extreme skew, or excess zeros present in the data. In addition, the substantive interpretation and practical implications of heterogeneous residuals requires further research. In this research, the role of estimating heterogeneous residuals was merely to decontaminate the estimate of skewness from one source of possible bias.

Discussion

IRT models are valuable psychometric tools when the model-derived latent variable scale (θ) accurately reflects individual differences on the trait the researcher is trying to measure and the estimated item parameters faithfully reflect the relation between trait levels and the probability of category response. To be used effectively, however, models such as the logistic GRM (Samejima, 1969) make many assumptions about the latent trait (causative, not emergent[7]), the item response data (local independence), the calibration sample (homogeneous, representative), the nature and shape of the latent variable (continuous, normal), the distribution of errors, and the parametric form of the model (linear relation between theta and log-odds of responding). The validity of the conclusions drawn from any IRT model application is threatened to the degree that any of these assumptions are violated.

Similar to Pilkonis et al. (2011), we applied the logistic GRM, with normality assumption, to responses to a 29-item measure of anger. We then considered four alternative models that relaxed one or more of the assumptions listed previously. The RC-IRT (nonparametric) and HS (parametric) models allow the researcher to estimate a fully continuous latent distribution simultaneously with the item parameters, and thus relax the normality assumption. Both approaches also allow a likelihood ratio test to compare the non-normal versus normal distribution model.

The ZIM and LL models relax normality in different ways, and neither model allows for a simple likelihood ratio statistical test of whether it is a significant improvement over the GRM.

The ZIM assumes that the population is heterogeneous and the observed normality violations are caused by an untraited or nonpathological latent class. When this subsample is estimated and removed, the GRM with normality assumption is then applied to the traited or "pathological" group. The LL model replaces the logistic function in the GRM with the LL, and replaces the normality assumption with an assumption of log-normality. It can be used in the same situations as the ZIM model.

Review of practical differences

An important practical concern is determining whether the alternative models yield either different scalings of individual differences, or provide a different view on the psychometric properties of the items and the scale (CRCs, scale response curves, scale information). In this section, we review model differences for the Anger scale in terms of scoring and psychometric evaluation.

Correlations of latent trait estimates generated from the GRM with estimates from the ZIM (for those in the traited class), RC, and HS models are all nearly perfect. These results suggest that these models make little difference in terms of relative standing on the estimated latent trait. The only real difference for the Anger data is in the standard error, which would be larger in the ZIM model (because of lower slopes) and smaller in the RC model (because of the higher slopes).

The one distinctive model in terms of scoring was the LL model, where latent trait estimates correlated $r = .72$ with those from the GRM. Although some might view this as a high linear correlation, implying similar patterns of external relations, it is important to note two differences. First, the anchor for the scale and the interpretation is very different. In the GRM, the anchor is the mean of $\theta = 0$ and scores are interpreted relative to that mean. In the LL, the anchor is $\theta = $ zero—no disorder—and scores reflect severity of the disorder. Second, relative to the GRM, differences between people near the low end of the scale are compressed, whereas differences between people toward the high end of the scale are expanded. Thus, the substantive (heritability coefficients, correlates with neurobiological parameters or life outcomes) and psychometric results (indexes of clinically important differences) based on the Anger data under these two models could differ dramatically.[8]

In terms of psychometric properties, as noted earlier, one major difference was the reduced slope parameters in the ZIM model compared to the GRM, suggesting that they are inflated in the GRM due to excess zeros, roughly, more zero scores than expected under a normal distribution. The implication is that if a sample contains fewer or greater noncases, the slope parameters of the GRM will change accordingly. On the other hand, the RC model suggested higher slopes once a non-normal latent trait distribution was estimated. Nevertheless, we believe that the RC results are

---

[7]It is also worth noting that not all constructs are best conceived as latent variables; some constructs are best represented as emergent variables (Bollen & Lennox, 1991; Fayers, Hand, Bjordal & Groenvold, 1997). IRT or factor-analytic models are inappropriate for this latter type of construct. Full discussion of this issue is beyond the scope of this article. We have assumed for simplicity that the latent variable measure framework is sensible.

[8]We note that IRT models have been criticized (Goldstein, 1980) exactly because if you change the basic model from a logistic model, very different scaling of individual differences might occur. Historically, the choice of a logistic function was not based on any substantive consideration or proven validity, but rather simply on mathematical convenience. This remains true today.

untrustworthy in this particular application due to the extreme skew caused by the excess zeros identified in the ZIM model.

This same concern with the possibility of excess zeros applies to the interpretability of the HS model as well. The HS model that contained item heterogeneity parameters (allowing error variances to increase or decrease as a function of trait level) and estimated a skewed normal distribution fit better than the GRM. Moreover, because the estimated skewness was very small, the better fit could be mostly attributable to the estimated item heterogeneities. Nevertheless, allowing for heterogenous variances did not result in item or TRCs that differed appreciably from the GRM. Because of a lack of research on this model, we do not know the degree to which possible excess zeros lead to biases in heterogeneity or skewness estimates.

Finally, as in the case of scoring, the LL model provided the largest contrast with the normal theory GRM. In relative terms, items are just as discriminating in the GRM and LL model, but where that discrimination is located is vastly different. In the LL model, information is very high in the near-zero trait range, indicating that the item set yields a precise discrimination between people who are low on Anger (i.e., the majority of subjects) and those who are not. In terms of differentiating among individuals at the positive end of the scale, standard errors are relatively larger. These psychometric differences have implications for all types of applications of IRT models including linking, computerized adaptive testing, and the study of differential item functioning. We note in closing that just as GRM slope parameters can be inflated by excess zeros, so can the analogous parameters in the LL model. In short, the LL results presented here could be misleading if one considers some zero scores as excess zeros.

Deciding between approaches in practice

Throughout, we have not considered whether the alternative models provide a statistically better fit than the GRM that assumes normality; with a large enough sample, we assume that any model without a restrictive normality assumption will display a superior statistical fit. In practice, there are no ready fit indexes or rules of thumb for deciding between the models considered here. Rather, what is required is the thoughtful consideration of mostly theoretical questions. For example, if a researcher considers the latent variable to be fully continuous with meaningful variation on both ends of the scale (bipolar), but the latent trait distribution might be skewed, then RC and HS models are viable candidates.

However, we warn that not only do these methods have limitations in the type of distribution that can be estimated, but the item response data must allow for the accurate estimation of a non-normal latent trait. For short scales (e.g., five items), and for scales that do not include items that discriminate well across the trait range,[9] the ability of any algorithm to correctly

estimate a latent distribution is severely compromised. Moreover, if the skew is caused by excess zeros, possibly due to poor sampling, estimating a true latent distribution would be nearly impossible.

If the construct is considered unipolar or a quasi-trait, then the ZIM and LL models can be considered. In deciding between these two models, a researcher must ask questions such as this: What do low scores reflect, low trait standing or absence of the trait? A critical difference between these models is that the ZIM model assumes a normal distribution for the population, but the sample is contaminated by excess zeros. To obtain correct population parameters, one needs to identify and eliminate these cases from the calibration. The LL assumes a highly skewed distribution in the population. If that assumption is justifiable, and it make sense to reference scores and clinical change relative to a zero anchor, the LL model might be the more appropriate choice. It is clear to us that more research is needed on the robustness of LL model parameters to excess zeros. It is also possible, in theory, to develop an LL model with excess zeros analogous to the ZIM model.

Conclusion

Lucke (2015) argued that a chief virtue of IRT modeling is that it allows researchers to develop measurement models that are consistent with the theory of the construct (see also Asparouhov & Muthén, 2015). Indeed, his LL model was selected not merely because it can account for highly skewed response data—dozens of monotonically increasing functions can do that—but rather because the LL model and log-normal trait scale are potentially more consistent with the cognitive neuroscience of addictive behavior. Lucke is by no means alone in proposing that measurement models need to be consistent with the hypothesized underlying response processes and what is known about the nature of specific constructs.

Stark, Chernyshenko, Drasgow, and Williams (2006) and Weekers, Anke, and Meijer (2008) considered the IRT modeling of personality data in terms of an unfolding response process.[10] Van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011) considered a diffusion model for the response process that might be appropriate for bipolar traits, but not for unipolar traits that are anchored at no ability or no trait at the low end. Although neither of these models was detailed here, they are examples of new psychometric developments of potential value as we move toward the next generation of IRT applications in the personality, psychopathology, and health outcomes domains. We hope that this article provides motivation for researchers to more carefully consider the nature of the latent trait, and to explore the application of alternative models. Only then can

---

[9]We note that the sole psychometric justification for having multiple, ostensibly, ordered response options, rather than yes–no, true–false, is to allow for better differentiation among individuals across the assumed latent trait continuum. When location parameters in the GRM are bunched at one end of the continuum (e.g., all in the positive trait range), the items are not differentiating among individuals across the continuum. One possible reason is that the low end of the trait does not exist—it is a unipolar or quasi-trait.

[10]By response process, we mean the theory of how trait levels are linked to item responses. Traditional test theory models are dominance models: The probability of item endorsement depends on the degree to which an individual's trait level is higher than an item's location (e.g., easy vs. hard). In an "unfolding" response process, the probability of endorsement is determined by the absolute distance between trait level and item location. The closer the trait level is to the item location, the more likely an endorsement.

we obtain additional substantive insights and findings from these models.

Funding

References

Asparouhov, T., & Muthén, B. (2015). Structural equation models and mixture models with continuous non-normal skewed distributions. *Structural Equation Modeling*, *34*, 1041–1058. doi:10.1080/10705511.2014.947375

Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centered parameterization. *Computational Statistics & Data Analysis*, *55*, 353–365. doi:10.1016/j.csda.2010.05.003

Azzalini, A., & Capatanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B*, *61*, 579–602. doi:10.1111/1467-9868.00194

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., … & Mehta, P. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314. doi:10.1037/0033-2909.110.2.305

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in M*plus* for estimating item response theory parameters for non-normal populations. *Structural Equation Modeling*, *19*, 610–632.

Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research*, *6*, 393–406.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234–246.

Gray-Little, B., Williams, V. S., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*, 443–451.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162. doi:10.1177/0146621012203197

Lucke, J. F. (2015). Unipolar item response models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 272–284). New York, NY: Routledge/Taylor & Francis Group.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732. doi:10.1007/s11336-005-1295-9

Michonski, J. D., Sharp, C., Steinberg, L., & Zanarini, M. C. (2013). An item response theory analysis of the DSM–IV borderline personality disorder criteria in a population-based sample of 11- to 12-year-old children. *Personality Disorders: Theory, Research, and Treatment*, *4*(1), 15–22.

Molenaar, D., Dolan, C. V., & de Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, *77*, 455–478. doi:10.1007/s11336-012-9273-5

Muthén, L. K., & Muthén, B. O. (2016). M*plus: Statistical analysis with latent variables: User's guide* (pp. 1998–2012). Los Angeles, CA: Muthén & Muthén.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*, 263–283. doi:10.1177/1073191111411667

Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychological Medicine*, *46*, 2025–2039. doi:10.1017/S0033291716000520

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*, 319–335. doi:10.1007/BF02296149

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*, 65–88. doi:10.1080/08957340701796415

Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., … Cella, D. (2016). Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. *Journal of Clinical Epidemiology*, *73*, 119–127. doi:10.1016/j.jclinepi.2015.08.036

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299–311. doi:10.1177/014662169001400307

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, *39*, 583–597. doi:10.1177/0146621615588184

Webster, G. D., & Jonason, P. K. (2013). Putting the "IRT" in "dirty": Item response theory analyses of the Dark Triad Dirty Dozen—An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences*, *54*, 302–306.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models. *European Journal of Psychological Assessment*, *24*(1), 65–77.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for non-normal latent variables. *Psychological Methods*, *11*, 253–270. doi:10.1037/1082-989X.11.3.253

Woods, C. M. (2015). Estimating the latent density in unidimensional IRT to permit non-normality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 60–84). New York, NY: Routledge/Taylor & Francis Group.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301. doi:10.1007/s11336-004-1175-8