Article

Educational and Psychological Measurement
2018, Vol. 78(1) 103–127

DOI: 10.1177/0013164416670211
journals.sagepub.com/home/epm
SAGE

A Study of Reverse-Worded Matched Item Pairs Using the Generalized Partial Credit and Nominal Response Models

Ki Lynn Matlock Cole[1], Ronna C. Turner[2], and W. Dent Gitchel[3]

Abstract

The generalized partial credit model (GPCM) is often used for polytomous data; however, the nominal response model (NRM) allows for the investigation of how adjacent categories may discriminate differently when items are positively or negatively worded. Ten items from three different self-reported scales were used (anxiety, depression, and perceived stress), and authors wrote an additional item worded in the opposite direction to pair with each original item. Sets of the original and reverse-worded items were administered, and responses were analyzed using the two models. The NRM fit significantly better than the GPCM, and it was able to detect category responses that may not function well. Positively worded items tended to be more discriminating than negatively worded items. For the depression scale, category boundary locations tended to have a larger range for the positively worded items than for the negatively worded items from both models. Some pairs of items functioned comparably when reverse-worded, but others did not. If an examinee responds in an extreme category to an item, the same examinee is not necessarily likely to respond in an extreme category at the opposite end of the rating scale to a similar item worded in the opposite direction. Results of this study may support the use of scales composed of items worded in the same direction, and particularly in the positive direction.

[1]Oklahoma State University, Stillwater, OK, USA
[2]University of Arkansas, Fayetteville, AR, USA
[3]University of Arkansas at Little Rock, Little Rock, AR, USA

Corresponding Author:
Ki Lynn Matlock Cole, Research, Evaluation, Measurement, and Statistics, College of Education, Oklahoma State University, 313 Willard Hall, Stillwater, OK 74078, USA.
Email: ki.matlock@okstate.edu

Keywords

generalized partial credit model, nominal response model, category boundary discrimination, category boundary location, item wording direction

Psychological scales, for example, anxiety, depression, and stress inventories, tend to be a combination of positively worded (PW) and negatively worded (NW) items with ordered item responses using a Likert-type format. An ordinal item response theory (IRT) model, such as generalized partial credit model (GPCM), is often applied to response data from a psychological scale (e.g., Baker, Rounds, & Zevon, 2000; Cook et al., 2007; Ebesutani et al., 2012; Fraley, Waller, Brennan, 2000; Karim, 2010; Rodebaugh et al., 2004; Taylor, 2015; Wang, Chen, & Jin, 2015; Yurekli, 2010), but little research uses the nominal response model (NRM) with these types of instruments (DeMars, 2004; Preston, Reise, Cai, & Hays, 2011; Tokuda et al., 2009).

In the first explorations, the NRM was found favorable over the GPCM in terms of model fit when data were known to have unequal discriminations across categories (DeMars, 2004) or when the analyses were used as an exploratory tool to evaluate significant differences in category discriminations. Application of the NRM to commonly used rating scale data indicates that the additional information provided from the NRM, in comparison with the GPCM, is useful to better understand distinctions between response categories on self-reported psychological scales (Preston et al., 2011). In fact, the assumption of a common discrimination across categories by the GPCM may be violated for many items on a scale. Hence, the NRM may be useful to investigate not only the usefulness of rating scale categories but also the impacts of item wording. The purpose of this study was to evaluate the effects of item wording directionality on pairs of reverse-worded items in order to compare the model-fit and estimated item parameters of the GPCM and NRM and to compare the category boundary discriminations for positively and negatively worded pairs of items that are reverse-worded.

Item Directionality

Through much debate, psychological scales continue to be mixed-worded, that is, composed of PW and NW items. Mixed-worded scales are intended to reduce response biases such as extreme responses or straight-line responses and to increase the range of responses (Baumgartner & Steenkamp, 2001; Wong, Rindfleisch, & Burroughs, 2003). Theoretically, if a respondent answers an extreme response to a PW item, an extreme response at the opposite end of the rating scale would be given to the same item worded in the negative direction. However, previous studies of item directionality (e.g., Ebesutani et al., 2012; Preston et al., 2011; Rodebaugh et al., 2004) have not included pairs of reverse-worded items, for instance, ''I am happy''

and ''I am not happy.'' Rather, responses to subsets of PW or NW items were used that were not necessarily measuring matched components of the construct definition. Having a balance of PW and NW items is encouraged to mitigate the contamination of non–construct-related factors on responses, such as item wording directionality, and the effects of acquiescence can be magnified on scales that are disproportionate or unbalanced (Baumgartner & Steenkamp, 2001; Couch & Keniston, 1960). Not all are in favor of using mixed-wording on questionnaires due to potential cultural effects on responses (Wong et al., 2003).

Polytomous Item Response Theory Models

To study the effects of item wording directionality, classical measures, such as observed scores and correlation, have been used. Yet, modern methods of IRT are more recently being used to evaluate psychological and personality assessment scales (Preston et al., 2011; Reise & Henson, 2010). The benefits of IRT include item-level parameter estimation, estimation of item and scale information, and scores of the underlying construct being measured based on specific item responses. Various IRT models may be applied to datasets when the responses are polytomous. Both the GPCM and the NRM are in the same family of models and are considered ''divide-by-total models.'' Each may be applied to responses on the nominal or ordinal scale, such as multiple-choice or rating scale item responses, and both are used in this study for comparing reverse-worded item pairs. The equations of the GPCM and NRM IRT models using the IRT parameterizations of the item characteristics are discussed, followed by the models in the slope–intercept form. A brief discussion of the prior applications of the models is then presented.

Generalized Partial Credit Model

The GPCM models the probability of responding in adjacent categories: 1 versus 2, 2 versus 3, 3 versus 4, etc. Muraki (1992) defines the probability of providing a response to item j's k th category ($X_j = k$) by

$$P(X_j = k|\theta, a_j, \underline{\boldsymbol{b}}_j) = \frac{\exp\left[\sum_{v=1}^{k} a_j(\theta - b_{jv})\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} a_j(\theta - b_{jv})\right]} \quad (1)$$

where θ is the latent trait, a_j is the item discrimination, b_{jv} is the category boundary location (CBL) parameter between the k th and $k-1$ category ($\theta - b_{j1} = 0$), and m_j is the number of category responses. For an item with four category responses, $m_i = 4$, and there are three CBLs: b_{j2}, b_{j3}, and b_{j4}. The CBL is also equal to the location of the intersection of adjacent category curves.

Nominal Response Model

The NRM models the probability of responding in the reference category as compared with each of the other $m - 1$ categories: 1 versus 2, 1 versus 3, 1 versus 4, and so on. Bock (1972) defines probability of responding to item j's k th category as

$$P(X_j = k|\theta, \underline{\boldsymbol{a}}_j, \underline{\boldsymbol{c}}_j) = \frac{\exp\left[c_{jk} + a_{jk}\theta\right]}{\sum_{v=1}^{m_j} \exp\left[c_{jv} + a_{jv}\theta\right]} \tag{2}$$

where a_{jk} is the discrimination of category k for item j, c_{jk} is a location parameter of category k to item j, and a_{jv} and c_{jv} are the discrimination and location parameters, respectively, for each of the $v = 1, 2, \ldots m$ categories.

Since the GPCM provides a location parameter for adjacent categories, Thissen, Steinberg, and Fitzpatrick (1989) provide a transformation of the discrimination parameters from the NRM to better interpret in relation to adjacent categories. A measure of the category boundary discriminations (CBDs) between categories k and k' is given by

$$CBD_{jk,k'} = a^*_{jk,k'} = a_{jk} - a_{jk'} \tag{3}$$

When $k' = k - 1$, this is the discrimination, or slope, between adjacent categories k and $k - 1$ ($k = 2, 3, \ldots m$). The CBL is equivalent to the location on the θ scale where an examinee is equally likely to respond to categories k and k':

$$CBL_{jk,k'} = c^*_{jk,k'} = \frac{c_{jk'} - c_{jk}}{a_{jk} - a_{jk'}} \tag{4}$$

When $k' = k - 1$, the CBL has a similar interpretation to the b_{jv} parameter from the GPCM.

The benefit of the NRM is that category-specific discrimination parameters may be more informative and contribute more to the estimate of the underlying trait. This may also be helpful in understanding the usefulness of specific response categories as they relate to individual items or identify poorly functioning response categories. However, the additional item parameters increase the likelihood of error variance in the estimations, and these parameters are more difficult to interpret than those from the GPCM. We apply the NRM to investigate the theoretical assumption that examinees may respond in opposite response categories to pairs of items worded in the opposite direction.

Slope–Intercept Form

The parameters of the IRT models are first estimated under the slope–intercept form of the equation. For the purposes of interpreting the item discrimination and location parameters, the slope and intercept parameters are transformed into the conventional IRT parameterizations, as presented above. The slope–intercept form is presented

here for the purposes of comparing the estimates from the two models or pairs of items using confidence intervals of the estimates. The slope–intercept form of the GPCM and NRM is given by Chalmers (2012):

$$P(X_j = k|\theta, a_j, \underline{\boldsymbol{d}}_j) = \frac{\exp[ak_{k-1}(a * \theta) + d_{k-1}]}{\sum_{v=1}^{k} \exp[ak_{v-1}(a * \theta) + d_{v-1}]} \quad (5)$$

For the GPCM, ak is constrained to $ak_0 = 0, ak_1 = 1,$ and so on, and the common slope for each category is given by a. For the NRM, ak is free to vary across categories. If items are truly ordered, then estimates from the NRM follow $ak_0 < ak_1 < \cdots < ak_{m-1}$.

The relationship between the slope and intercept values with the IRT parameterized values from the GPCM is somewhat straightforward, where $a = a$ and $d = -a\sum_{v=1}^{k} b_k$. For the relationship between the slope and intercept values with the IRT values from the NRM, readers are encouraged to review Bock (1972) and Samejima (1979).

Prior Application of the Generalized Partial Credit Model and Nominal Response Model

To our knowledge, only DeMars (2004) and Preston et al. (2011) have compared the GPCM and NRM. In a simulation study of data designed with items to fit either model by DeMars (2004), the GPCM estimated item parameters with substantially smaller root mean square errors (RMSEs) than the NRM when data were constrained to have equal category discriminations for all sample sizes and test length. When data were unconstrained, the NRM estimated item parameters with significantly smaller RMSEs than those from the GPCM when datasets were large (N = 2,000), and only slightly better when the sample size was small (N = 250).

Because of the difficulties of interpreting the multiple category discrimination parameters from the NRM as compared with the GPCM, Preston et al. (2011) estimated a category boundary discrimination and applied the two models to a patient-reported outcomes emotional distress scale measuring feelings of depression, anxiety, and anger in the past 7 days, with a 5-category response (*never, rarely, sometimes, often, always*). Prior to analyses, authors combined the fourth and fifth categories due to low responses in the fifth category. The estimated item discrimination from the GPCM tended to be the average of the CBDs across all categories. For the anxiety and anger scales, the NRM fit the data better than the GPCM, based on the AIC index and the likelihood ratio test; for the depression scale, the two models fit equally well.

Out of the 86 items used by Preston et al. (2011), 25 (29%) displayed significantly different CBDs across categories, but no trend in item content was observed. The first CBD, between responses of never and rarely, tended to be the highest; the next highest CBD was between responses in categories sometimes and often/always; the lowest

CBD, and the least discriminating boundary, was between categories of rarely and sometimes. As a result, estimated trait scores tended to differ more for those responding between the first and second categories as compared with the second and third. This distinction is not allowed when the GPCM is applied. In closing, authors point out the usefulness of evaluating the effects of direction of the item wording using the CBDs from the NRM.

The purpose of this study was to extend the work of Preston et al. (2011) with the following objectives:

1. Compare the model fit and test information from the GPCM and NRM
2. Compare the item information and item parameter estimates from the GPCM and NRM
3. Compare the item parameter estimates for pairs of reverse-worded items.

These techniques were applied to empirical data gathered from three self-reported psychological scales measuring anxiety, depression, and perceived stress, with additional items worded in the reverse direction of the original items in order to make direct comparisons based on item directionality, controlling for item content differences.

Method

Participants

The instrument containing three scales was administered to a volunteer sample of 1,711. Of these, 305 (17.8%) were male, 978 (57.2%) were female, and 428 (25.0%) did not provide their gender. A majority of the sample had some amount of college education: 129 (7.5%) had no college-level education, 491 (28.7) had some college education, 89 (5.2%) had an associate's degree, 301 (17.6%) had a bachelor's degree, and 398 (23.3%) had a graduate or professional degree. In all, 303 participants (17.7%) chose not to respond. A majority of the sample was White (N = 1,192, 69.7%), 102 (6.0%) were African American, 67 (3.9%) were Hispanic, 111 (6.5%) were Native American, Asian, Native Hawaiian, or other minority group.

Data were analyzed for potentially biased response patterns due to acquiescence or inattention to item content observed when respondents answered at the same extreme for both PW and NW items on the same scale. After excluding problematic response patterns and those with incomplete data, the sample sizes were 1,046, 1,118, and 1,388 for the anxiety, depression, and stress scales, respectively. When applying the NRM, DeMars (2003) suggested a sample size of 2,400 over 600; however, de Ayala (2009) provides a guideline of a minimum sample size of 600 for the NRM.

Instruments

Ten items were chosen from each of the following scales: the Zung Self-Rating Anxiety Scale (SAS; Zung, 1971), the Center for Epidemiologic Studies–Depression Scale (CES-D, Radloff, 1977), and the Perceived Stress Scale (PSS; Cohen, 1994). Each contained a combination of positively and negatively worded items; these are referred to as the originally worded items (OW). A corresponding item was written in the opposite direction for each of the selected items, called reverse-worded items (RW). For example, Item 3 from the original PSS reads, ''. . . felt nervous and stress;'' the reverse-worded item reads, ''. . . not felt nervous and stress.'' In total, the instrument was composed of 60 items containing 30 pairs of PW and NW items. Additional demographics were also gathered.

The 10 original items from the SAS (Zung, 1971) contained 5 PW and 5 NW items and had a 4-point response scale: 1 = *A little of the time*, 2 = *Some of the time*, 3 = *A good part of the time*, 4 = *Most of the time*. The CES-D (Radloff, 1977) had two PW and eight NW original items and used a 4-point scale: 1 = *Rarely or none of the time (less than 1 day)*, 2 = *Some or a little of the time (1-2 days)*, 3 = *Occasionally or a moderate amount of time (3-4 days)*, 4 = *Most or all of the time (5-7 days)*. The original 10 items of the PSS scale (Cohen, 1994) had 4 PW and 6 NW items and used a 5-point scale: 1 = *Never*, 2 = *Almost never*, 3 = *Sometimes*, 4 = *Fairly often*, 5 = *Very often*.

Analysis

PW items were reverse-coded prior to analysis. Data from each scale were analyzed separately. The Cronbach's alpha reliabilities for the sets of 20 items for each scale of anxiety, depression, and stress were .874, .900, and .938, respectively. To validate the subsets of RW items, the reliability of the OW and RW were compared. The measure of Cronbach's alpha for the subsets of OW and RW for the anxiety, depression, and stress data were .737 and .796, .819 and .822, and .890 and .874, respectively.

The only prior study of the category boundary discriminations, to our knowledge, was done by Preston et al. (2011). Authors reported that a five-response scale was used, but that the fourth and fifth category responses were combined due to low response-rate in the fifth category. We also observed low response rates of the highest category (for no items was the last category never selected), but chose not to combine the upper categories in order to have a clear comparison of category responses for PW and NW items.

The distributions of observed sum scores were evaluated for normality by measures of skewness and kurtosis. For each of the three scales, measures of skewness were 0.773, 0.750, and 0.098 for the anxiety, depression, and perceived stress scale scores, respectively. Measures of kurtosis were 2.989, 2.818, and 2.894 for the three scales. Because the data from the perceived stress scale more closely follows a normal distribution, it is presented in more detail in the results.

Principal axis factoring with a Promax rotation was applied to each set of 20 items using the ''psych'' package in R (Revelle, 2015). For the anxiety items, all but one pair of items loaded on the same factor; Items 10 and 20 loaded on a secondary factor. For the depression items, all but Items 7 and 17 loaded on a single factor, with this pair loading on a secondary factor. Interestingly, these four items regarded the respondent's sleep patterns. The two factors were only moderately correlated, with correlations of 0.38 and 0.42 for the anxiety and depression scales, respectively. With only two items loading on the second factor, relatively strong internal consistency reliability, and the popular use of the aggregate set of items in practice, it was decided to maintain a unidimensional model in the IRT analysis. The 20 perceived stress items resulted in a 2-factor solution with PW items loading on one factor and NW items loading on the other; the factors had a correlation of 0.77. Cohen and Williamson (1988) supported the use of unidimensional analyses when items loaded in this way.

The ''mirt'' package in R (Chalmers, 2012) was used to conduct the GPCM and NRM analyses. Significance was evaluated using a 95% confidence interval. Comparisons of item parameter estimates from the same item were made across the two models; if the estimated parameter from the GPCM was contained in the 95% confidence interval of the NRM, the estimates were not significantly different. Comparisons of the item parameter estimates on the PW and NW item were also made. The IRT parameterizations of the item discrimination and locations (Equations 1 and 2) are presented in the tables, figures, and interpretations of the results, but the significance is based on the estimates and standard errors of the slope and intercept of the models (Equation 5). Transformations of the estimates under the slope–intercept form and IRT parameterizations are presented above.

Results

Objective 1: Comparing Model Fit and Test Information From the Generalized Partial Credit Model and Nominal Response Model

For all sets of 20 items, the NRM fit significantly better than the GPCM according to the Akaike information criterion (AIC) and the log likelihood χ^2 test for nested models, presented in Table 1. Since two of the three criteria favored the NRM, it was chosen as the better fitting model.

Test information curves of the GPCM and NRM (shown in Figure 1) tended to be similar for the anxiety scale data for the entire range of the θ scale, with slightly more information provided from the GPCM for $\theta > 1.2$. For the depression scale data, information was slightly higher from the NRM for $\theta < 1.1$, and the GPCM provided more information for $\theta \geq 1.1$. For the perceived stress scale data, the two models provided similar information for those at $\theta < -2.6$; the NRM provided more information between $-2.6 \leq \theta < 0.7$, and the GPCM provided more information for $\theta \geq 0.7$. Inconsistencies in the trends of test information curves for the three datasets indicate

Table 1. Model Fit Statistics.

Scale	Model	AIC	BIC	−2LL	χ^2_{df}
Anxiety	GPCM	43574.25	43970.47	43414.26	
	NRM	43440.70	44035.03	43200.70	$\chi^2_{40} = 213.55$*
Depression	GPCM	47937.02	48338.56	47777.02	
	NRM	47783.19	48385.50	47543.18	$\chi^2_{40} = 233.83$*
Perceived Stress	GPCM	62156.27	62679.83	61956.26	
	NRM	61001.08	61838.78	60681.08	$\chi^2_{60} = 1275.19$*

Note. GPCM = generalized partial credit model; NRM = nominal response model; AIC = Akaike information criterion; BIC = Bayesian information criterion; −2LL = −2 log likelohood.
*$p < .01$.

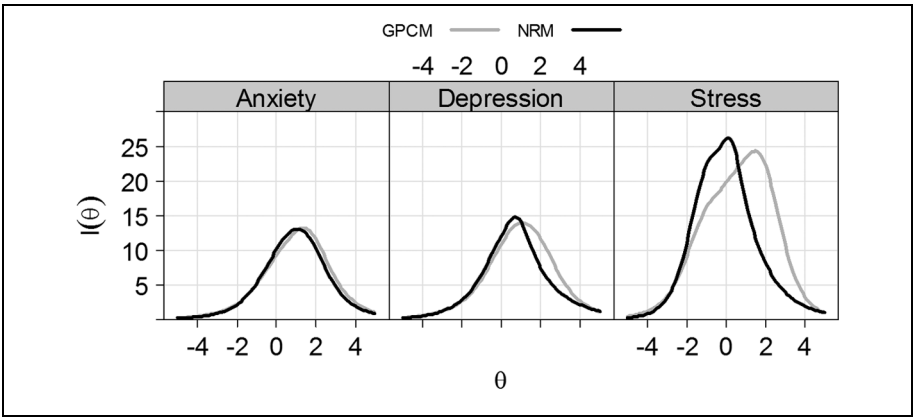

Figure 1. Test information curves from the generalized partial credit model (GPCM) and nominal response model (NRM) for the 3 scales of 20 items.

that the choice of model may depend on the choice of instrument and the decision may not be favorable for the entire θ scale.

Objective 2: Comparing Item Information and Parameter Estimates From the Generalized Partial Credit Model and Nominal Response Model

The estimated item parameters from the GPCM and that the transformed item parameters (see Equations 3 and 4) from the NRM for the anxiety, depression, and perceived stress scales are presented in Tables 2, 3, and 4, respectively. Items are ordered in the table by pairs, with the NW item listed first, followed by the PW item within each pair.

Table 2. Estimated Item/Category Discrimination and Boundary from the GPCM and NRM for the Anxiety Scale.

Direction	Item	Discrimination					Boundary					
What column best describes how often you have felt or behave this way during the past several days.		GPCM	NRM				GPCM			NRM		
		a	a^*_{12}	a^*_{23}	a^*_{34}	$\bar{a}^*$	b_1	b_2	b_3	c^*_{12}	c^*_{23}	c^*_{34}
−	I feel more nervous and anxious than usual. (1)	0.880	0.819	0.906	1.030	0.918	−0.577	0.830	1.936	−0.604	0.781	1.791
+	I feel calm and secure. (11)	2.123	2.003	2.146	2.031	2.060	−0.580	0.489	1.709	−0.591	0.465	1.715
−	I feel afraid for no reason at all. (2)	1.153	1.172	1.240	0.731	1.048	1.417	1.667	2.488	1.381	1.630	3.011
+	I feel safe. (12)	1.563	1.596	1.424	1.753	1.591	0.225	1.274	2.240	0.211	1.298	2.144
−	I feel like I am falling apart and going to pieces. (3)	1.219	1.251	1.230	1.008	1.163	0.547	1.387	1.681	0.520	1.379	1.745
+	I am able to hold myself together well. (13)	1.500	1.500	1.582	1.523	1.535	−0.009	1.415	2.687	−0.024	1.368	2.676
−	I feel that everything is not right and that bad things will happen. (14)	1.281	1.547	0.918	0.776	1.080	0.800	1.610	2.056	0.698	1.944	2.332
+	I feel that everything is all right and nothing bad will happen. (4)	0.925	0.870	1.187	0.587	0.881	−1.437	0.123	1.350	−1.520	0.113	1.748
−	I feel restless and fidgety. (15)	0.966	0.915	1.235	0.653	0.934	−0.459	1.708	1.505	−0.491	1.456	1.757
+	I feel calm and can sit still easily. (5)	1.148	1.332	1.243	0.830	1.135	−1.184	0.242	1.304	−1.131	0.252	1.488
−	I can feel my heart beating fast. (6)	0.676	0.517	1.134	0.385	0.679	0.364	2.122	2.769	0.456	1.489	4.203
+	My heart beat remains normal. (16)	0.971	1.014	1.339	0.024	0.792	0.398	1.141	2.465	0.347	1.036	55.583
−	Breathing in and out is difficult. (17)	0.648	1.417	0.606	−1.532	0.164	2.759	2.771	1.876	1.525	3.299	0.440
+	I can breathe in and out easily. (7)	0.636	1.133	0.940	−0.718	0.452	1.120	1.512	1.622	0.626	1.424	0.078
−	I am bothered by stomach aches or indigestion. (8)	0.371	0.291	0.382	0.555	0.409	0.700	2.980	1.893	0.897	2.840	1.481
+	My stomach feels healthy and normal. (18)	0.532	0.490	0.627	0.424	0.514	−0.148	0.652	2.143	−0.159	0.573	2.521

(continued)

Table 2. (continued)

Direction	Item	Discrimination					Boundary					
What column best describes how often you have felt or behave this way during the past several days.		GPCM	NRM				GPCM			NRM		
		a	a^*_{12}	a^*_{23}	a^*_{34}	$\bar{a}^*$	b_1	b_2	b_3	c^*_{12}	c^*_{23}	c^*_{34}
−	My hands are usually cold and clammy. (19)	0.293	0.287	0.127	−0.483	−0.023	3.010	3.193	2.646	3.125	7.024	−1.236
+	My hands are usually dry and warm. (9)	0.237	0.328	0.278	0.068	0.225	0.114	0.209	0.219	0.018	0.227	0.176
−	I have difficulty falling asleep and getting a good night's rest. (20)	0.366	0.236	0.347	0.538	0.374	−0.799	1.555	0.691	−1.106	1.565	0.589
+	I fall asleep easily and got a good night's rest. (10)	0.436	0.058	0.520	0.594	0.391	−1.637	−0.615	0.844	−9.397	−0.625	0.695
	Mean all items[a]	0.896	0.939	0.971	0.539	0.816	0.231	1.313	1.806	0.220	1.185	1.545
	Mean NW items[a]	0.785	0.845	0.813	0.366	0.675	0.776	1.982	1.954	0.640	1.820	1.611
	Mean PW items[a]	1.007	1.032	1.129	0.712	0.958	−0.314	0.644	1.658	−0.247	0.613	1.471

Note. Four-point response scale was used, where: 1 = *A little of the time*, 2 = *Some of the time*, 3 = *A good part of the time*, 4 = Most of the time. PW items were reverse-coded. GPCM = generalized partial credit model; NRM = nominal response model; NW, negatively worded; PW, positively worded.
[a]Items with a boundary estimates $|c^*| > 5$ were excluded from the calculation of the mean.

Table 3. Estimated Item/Category Discrimination and Boundary From the GPCM and NRM for the Depression Scale.

Direction	Item	Discrimination					Boundary					
Please tell me how often you have felt this week during the past week…		GPCM	NRM				GPCM			NRM		
		a	a^*_{12}	a^*_{23}	a^*_{34}	$\bar{a}^*$	b_1	b_2	b_3	c^*_{12}	c^*_{23}	c^*_{34}
−	I was bothered by things that usually don't bother me. (1)	0.825	0.870	0.823	0.640	0.778	0.001	1.590	2.978	−0.016	1.601	3.497
+	I did not get bothered by things easily. (11)	0.529	0.902	0.666	0.062	0.543	−1.553	0.123	1.901	−1.147	0.198	12.613
−	I had trouble keeping my mind on what I was doing. (2)	0.708	0.589	0.815	0.643	0.682	−1.331	0.631	1.927	−1.514	0.547	2.065
+	I was able to keep my mind on what I was doing. (12)	0.981	1.096	1.027	0.707	0.943	−1.086	0.535	2.290	−1.037	0.533	2.796
−	I felt depressed. (3)	1.688	1.478	2.052	1.298	1.609	0.318	0.958	1.845	0.345	0.892	2.024
+	I felt hopeful. (13)	2.057	2.176	2.626	0.771	1.858	−0.340	0.865	2.084	−0.360	0.827	3.195
−	I felt that everything I did was an effort. (4)	0.562	0.905	0.578	0.101	0.528	−0.923	0.837	2.162	−0.720	0.913	9.149
+	I was able to complete tasks without too much effort. (14)	1.157	1.171	1.258	0.894	1.108	−1.095	0.660	2.189	−1.097	0.628	2.522
−	I felt pessimistic about the future. (15)	0.697	1.194	0.506	0.066	0.589	0.371	1.124	2.182	0.172	1.551	14.470
+	I felt hopeful about the future. (5)	1.001	1.057	1.463	0.008	0.843	−0.145	0.829	2.040	−0.187	0.764	131.250
−	I felt fearful. (6)	0.809	0.855	0.710	0.795	0.787	1.227	1.499	2.923	1.164	1.615	2.950
+	I felt safe. (16)	1.063	1.195	0.881	0.633	0.903	1.082	1.612	3.052	0.982	1.797	4.038
−	My sleep was restless. (7)	0.460	0.314	0.515	0.521	0.450	−0.650	0.662	1.769	−0.815	0.555	1.633
+	My sleep was sound. (17)	0.469	0.391	0.601	0.309	0.434	−1.167	0.614	1.819	−1.345	0.498	2.515
−	I was unhappy. (18)	1.469	1.678	1.314	0.767	1.253	0.181	1.190	2.327	0.145	1.272	3.089

(continued)

Table 3. (continued)

Direction	Item	Discrimination					Boundary					
Please tell me how often you have felt this week during the past week…		GPCM	NRM				GPCM			NRM		
		a	a^*_{12}	a^*_{23}	a^*_{34}	$\bar{a}^*$	b_1	b_2	b_3	c^*_{12}	c^*_{23}	c^*_{34}
+	I was happy. (8)	1.889	1.942	2.149	0.689	1.593	−0.032	1.205	2.288	−0.052	1.177	3.581
−	I felt lonely. (9)	0.848	0.867	0.827	0.901	0.865	0.179	1.153	2.169	0.164	1.157	2.118
+	I did not feel lonely. (19)	0.756	1.061	0.901	0.139	0.700	0.261	0.647	1.359	0.101	0.707	3.892
−	I could not get "going." (10)	0.669	0.526	0.665	1.040	0.744	−0.602	1.166	2.303	−0.686	1.107	1.838
+	I was able to get "going." (20)	0.922	0.858	1.105	0.647	0.870	−0.725	0.857	2.227	−0.767	0.767	2.777
	Mean all items[a]	0.978	1.056	1.074	0.582	0.904	−0.301	0.938	2.192	−0.334	0.955	2.783
	Mean NW items[a]	0.874	0.928	0.881	0.677	0.828	−0.123	1.081	2.259	−0.176	1.121	2.402
	Mean PW items[a]	1.082	1.185	1.268	0.486	0.980	−0.480	0.795	2.125	−0.491	0.789	3.164

Note. Four-point response scale was used, where: 1 = *A little of the time*, 2 = *Some of the time*, 3 = *A good part of the time*, 4 = *Most of the time*. PW items were reverse-coded. GPCM = generalized partial credit model; NRM = nominal response model; NW, negatively worded; PW, positively worded.

[a]Items with a boundary estimates $|c^*| > 5$ were excluded from the calculation of the mean.

Table 4. Estimated Item/Category Discrimination and Boundary From the GPCM and NRM for the Perceived Stress Scale.

		Discrimination						Boundary							
		GPCM	NRM					GPCM				NRM			
Direction	Item	a	a^*_{12}	a^*_{23}	a^*_{34}	a^*_{45}	$\bar{a}^*$	b_1	b_2	b_3	b_4	c^*_{12}	c^*_{23}	c^*_{34}	c^*_{45}
−	1	0.966	0.578	1.052	1.281	0.535	0.862	−2.570	−0.684	1.865	2.341	−3.580	−0.716	1.560	3.521
+	11	1.275	1.540	1.460	0.719	−0.098	0.905	−1.299	0.276	2.143	2.831	−1.216	0.248	3.161	−17.755
−	2	0.879	1.486	1.234	0.821	−0.303	0.810	−1.970	−0.133	1.158	1.962	−1.575	−0.101	1.300	−2.462
+	12	1.971	2.358	2.585	0.929	−0.125	1.437	−0.895	0.418	1.575	2.357	−0.877	0.360	2.326	−12.016
−	3	1.139	1.201	1.140	1.251	0.998	1.148	−2.270	−1.524	0.518	1.205	−2.174	−1.539	0.440	1.274
+	13	1.019	1.466	1.308	0.810	0.237	0.955	−1.549	−0.551	0.894	2.128	−1.387	−0.459	1.041	6.042
−	14	1.281	1.985	1.840	0.891	−0.723	0.998	−1.345	0.338	1.405	2.260	−1.170	0.284	1.806	−1.051
+	4	1.463	1.922	1.909	0.840	−2.866	0.451	−0.552	0.664	1.780	2.923	−0.554	0.606	2.473	−0.195
−	15	1.793	1.971	1.942	1.705	1.064	1.671	−1.645	−0.081	1.337	2.057	−1.579	−0.124	1.358	2.625
+	5	1.585	1.507	1.939	1.470	0.144	1.265	−1.238	0.156	1.463	2.627	−1.277	0.092	1.552	15.083
−	6	1.546	1.742	1.699	1.206	0.773	1.355	−1.090	0.229	1.451	1.888	−1.059	0.184	1.648	2.497
+	16	2.051	2.561	2.823	0.667	0.179	1.558	−0.901	0.474	1.652	2.651	−0.874	0.412	3.030	12.872
−	17	0.416	1.813	1.215	−0.340	−1.367	0.330	−3.358	−0.376	1.482	4.360	−1.499	0.002	−1.109	−1.309
+	7	1.073	1.765	1.399	0.558	−1.559	0.541	−1.399	0.298	1.994	2.778	−1.156	0.297	3.213	−0.631
−	18	2.008	2.297	2.342	1.734	0.582	1.749	−1.632	−0.078	1.057	1.911	−1.548	−0.109	1.268	3.464
+	8	1.960	2.084	2.422	1.243	0.393	1.536	−1.343	0.295	1.314	2.344	−1.321	0.232	1.608	5.807
−	9	1.042	1.005	1.052	1.130	0.883	1.018	−2.119	−0.570	1.411	2.033	−2.133	−0.613	1.311	2.292
+	19	0.564	1.149	0.961	0.014	−0.762	0.341	−1.979	−0.464	1.608	4.438	−1.392	−0.195	50.214	−2.600
−	10	1.985	2.140	2.023	1.695	1.496	1.839	−0.954	0.276	1.206	1.740	−0.941	0.220	1.248	1.949
+	20	2.003	2.441	2.825	1.129	−2.184	1.053	−0.905	0.446	1.836	2.529	−0.885	0.381	2.543	-0.007
Mean all items[a]		1.401	1.751	1.759	0.988	−0.135	1.090	−1.550	−0.029	1.458	2.469	−1.410	−0.027	1.665	0.669
Mean NW[a]		1.305	1.622	1.554	1.137	0.394	1.177	−1.895	−0.260	1.290	2.176	−1.726	−0.251	1.069	1.280
Mean PW[a]		1.496	1.879	1.963	0.838	−0.664	1.004	−1.206	0.201	1.626	2.761	−1.094	0.197	2.328	−0.858

Note. Five-point scale was used, where: 1 = *Never*, 2 = *Almost never*, 3 = *Sometimes*, 4 = *Fairly often*, 5 = *Very often*. PW items were reverse-coded. GPCM = generalized partial credit model; NRM = nominal response model; NW, negatively worded; PW, positively worded.

[a]Items with a boundary estimates $|c^*| > 5$ were excluded from the calculation of the mean.

Table 5. Item Stems From the Perceived Stress Scale Items.

In the past month, how often have you . . .
been upset because of something that happened unexpectedly. (1)
remained calm when something happened unexpectedly. (11)
felt you were unable to control the important things in your life. (2)
felt you were able to control the important things in your life. (12)
felt nervous and "stressed." (3)
not felt nervous and stressed. (13)
felt unsure about your ability to handle your personal problems. (14)
felt confident about your ability to handle your personal problems. (4)
felt that things were not going your way. (15)
felt that things were going your way. (5)
found that you could not cope with all the things you had to do. (6)
found that you were able to cope with all the things you had to do. (16)
been unable to control irritations in your life. (17)
been able to control irritations in your life. (7)
felt that you were not on top of things. (18)
felt that you were on top of things. (8)
been angered because of things that were outside of your control. (9)
have not gotten angry at things that were outside your control. (19)
felt difficulties were piling up so high that you could not overcome them. (10)
felt that you could successfully confront difficulties as they occurred. (20)

Item Information. Item information curves (IICs) from the GPCM and NRM were somewhat similar for the items on the anxiety scale. Items providing little to no information when estimated from the GPCM had a similar information curve from the NRM; likewise, those providing more information at some θ tended to provide information with a similar distribution. IICs from items on the depression scale tended to follow the same shapes when estimated with either model, but curves from the GPCM were shifted to the right, with peaks at a higher θ, than those from the NRM, which was also displayed on the test information curve from the depression scale.

IICs from items on the perceived stress scale (Table 5) when fit to the GPCM and NRM were more different. Figure 2 displays information curves for pairs of items from the perceived stress scale; in order to compare the effect of the model, compare lines of the same color and different styles (i.e., compare dotted black lines to solid black lines [for NW items] or compare dotted gray lines to solid gray lines [for PW items]). For some items, the two models yielded similar IICs when estimated, particularly the negatively worded Items 3 and 9. For others, the distribution of information from the GPCM (dotted lines) was shifted to the right of the NRM (solid line), providing more information at higher θs, whereas the NRM provided more information around the middle of the θ scale.

Items 17 and 19 provided little to no information across the entire θ scale when estimated with the GPCM, but some amount of information was provided when estimated with the NRM.

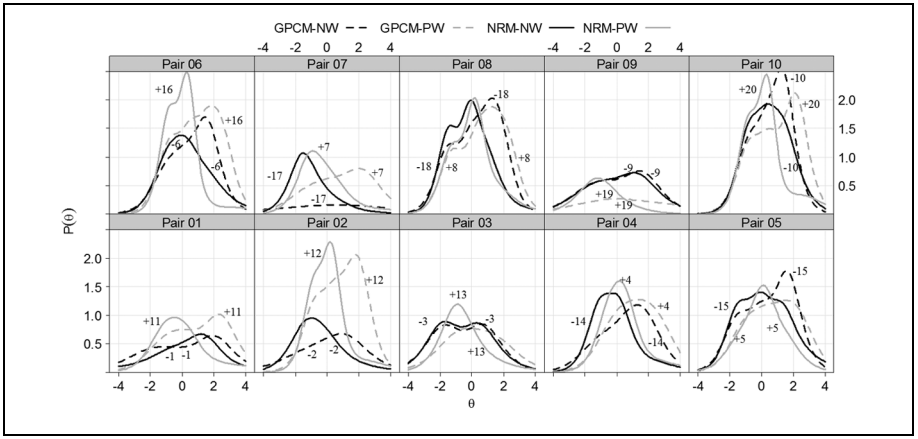

Figure 2. Item information curves from the generalized partial credit model (GPCM; dotted lines) and nominal response model (NRM; solid lines) for pairs of positively worded (PW; gray lines) and negatively worded (NW; black lines) items from the Perceived Stress Scale.

Discrimination. The single estimated discrimination value from the GPCM tended to be more correlated with the average of the CBD estimates from the NRM for the anxiety and depression scales (these were $r = 0.97$ and 0.99, respectively). Within specific CBDs, the GPCM estimated discrimination was most correlated with the CBD between the second and third of the four categories ($r = 0.93$ and 0.96 for the anxiety and depression scales, respectively) and least correlated with the CBD between the last third and fourth categories ($r = 0.70$ and 0.49 for the anxiety and depression scales, respectively). For the perceived stress scale, the correlation between the GPCM estimated discrimination and the average of the CBDs was $r = 0.83$; the GPCM estimate was most correlated with the discrimination between the second and third of five categories ($r = 0.90$) and least correlated with the CBDs between the last fourth and fifth categories ($r = 0.22$).

The single estimated discrimination from the GPCM must be generalized for all categories, but the NRM provides an estimate for the discrimination among adjacent categories. A CBD near zero indicates little to no distinction between adjacent categories (Preston et al., 2011). On all scales, the CBD between the last categories (third and fourth on the anxiety and depression scales or fourth and fifth on the perceived stress scale, corresponding to higher levels of the trait) tended to have the lowest estimates from the NRM, indicating little discrimination between examinees responding in the last two categories. Though ordinal Likert-type response scales were used, negative CBDs were present for some items, indicating nonordered responses. Three items on the anxiety scale (Items 7, 17, and 19) and nine items on the perceived stress scale (Items 2, 4, 7, 11, 12, 14, 17, 19, and 20) had negative CBDs between the highest categories indicating a higher levels of anxiety and perceived stress, respectively.

This indicated that the selection of the lower of the two adjacent categories was more likely a response for someone with a high level of the trait than was the highest category.

The significance of the differences in the estimated slope (a) across the GPCM and NRM was evaluated by comparing the overlap of the 95% confidence intervals of the estimated slope and intercept parameters. The specific category slopes from the NRM were free to vary, but those from the GPCM were fixed ($ak_0 = 0,\ ak_1 = 1,\ ak_2 = 2,\ ak_3 = 3$), and there was no standard error of estimate. Only significant differences between estimated a are discussed. The differences for these measures were considered significant if the fixed value of the GPCM was not contained in the 95% confidence interval from the NRM estimate.

Overall, the estimated slope for all categories (a) had larger standard errors, on average, from the GPCM than from the NRM. The $\bar{a}$ for all 20 items on the anxiety scale was 0.077 from the GPCM and 0.095 from the NRM. On the depression scale, the average standard errors of the slope were 0.073 and 0.083 for the GPCM and NRM, respectively. And on the perceived stress scale, the average standard errors were 0.071 and 0.081, respectively.

The estimated slope (a, constant for all categories) was significantly different for two items from the anxiety scale (Items 17 and 19), no items on the depression scale, and nine items from the perceived stress scale (Items 4, 5, 7, 11, 12, 14, 16, 19, and 20). The two items from the anxiety scale were NW items, and all but one of the nine items on the perceived stress scale that were significantly different for the two models were PW. Most items from the perceived stress scale that displayed significantly different slope estimates were those that also had a negative estimated CBD between the last two categories. Significant differences were also reported for category slopes (ak_{k-1}) from the two models for many items on all three scales. Again, there was no clear trend of the effect of the wording direction.

Category Boundary Locations. CBLs are equivalent to estimates of the θ where the probability of responding in adjacent categories is equal. The lack of equal discrimination across categories when using the NRM as compared with the GPCM had strong effects on the category boundary locations and category characteristic curves. For some items, CBLs and the category curves tended to be similar when estimated with each model, but for others the curves were different. Item information curves are directly related to the estimated item parameters. As a result, items that had very similar IICs when estimated with the GPCM and NRM had very similar item category curves (ICCs) from the two model fits. Figure 3 displays the ICCs of eight items selected items (four pairs of PW and NW items) estimated with the GPCM and NRM. Items 3 and 9 had similar IICs from the GPCM and NRM fits; likewise, Items 3 and 9 have similar ICCs.

The other six items in Figure 3 (Items 7, 8, 13, 17, 18, and 19) had very different category curves from the GPCM and NRM. In most cases, the last category, Category 5, had no functionality across any portion of the θ scale. Consider Item 7, ''. . . been able to control irritations in your life,'' which is positively worded and

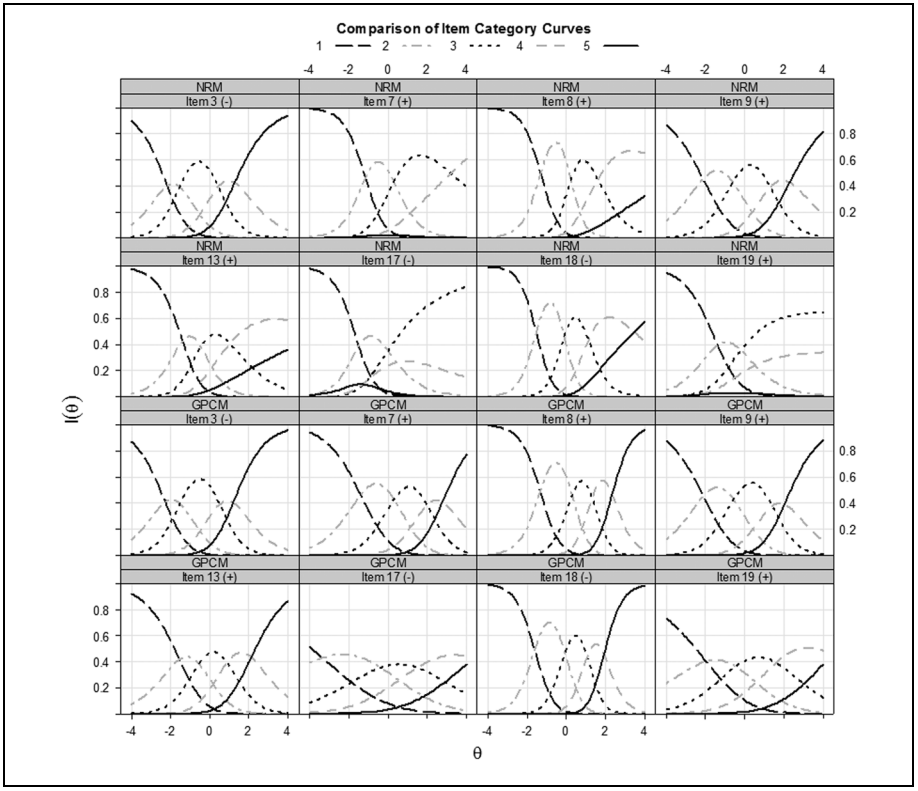

Figure 3. Category characteristic curves from the generalized partial credit model (GPCM) and nominal response model (NRM) for eight selected items (four positively worded [PW] and four negatively worded [NW] item pairs) from the Perceived Stress Scale.

was recoded. The boundary between categories indicating lower levels of perceived stress (Categories 1 and 2 and between Categories 2 and 3) were similar for the two models (see Table 4 for values); however, the boundaries between categories likely from respondents with high levels of perceived stress (Categories 3 and 4 or 4 and 5) were very different. From the GPCM, someone with a perceived stress level $\theta > 2.778$ is likely to respond in the *Never* category; from the NRM, no one is likely to respond *Never*, and someone with $\theta \geq 2.778$ is likely to respond in the *Sometimes* or *Almost Never* categories. Therefore, the use of the NRM indicates that the fifth category is irrelevant to estimating the measure of perceived stress for this item.

ICCs from the anxiety scale data were very similar from the GPCM and NRM for corresponding items. From the depression scale, five of the 20 items had occurrences where the fourth category indicating the highest level of depression was problematic when the NRM was fit.

The correlations of the CBLs for corresponding boundaries from the two models were higher for the lower and middle CBLs than for the last categories. On the anxiety scale, the correlation between boundary locations between the first and second categories from the GPCM and NRM was 0.72, between the second and third was 0.87, and between the third and fourth was 0.47. On the depression scale, the correlation between adjacent categories, that is, the first and second, second and third, and third and fourth categories, from the GPCM and NRM was 0.98, 0.96, and 0.12, respectively. On the perceived stress scale, the correlation between adjacent categories, that is, the first and second, second and third, third and fourth, and fourth and fifth categories, from the GPCM and NRM was 0.72, 0.98, 0.15, and −0.50, respectively. This indicated that the boundaries between the categories measuring low to moderate amounts of the trait tended to increase or decrease at similar magnitudes when estimated with the two models, and boundaries from the two IRT models at the upper end of the trait distribution tended to vary substantially. For almost all items, the NRM produced a wider range of CBLs than the GPCM. This indicates that the NRM provides a more informative estimate of θ for a wider range of the trait scale than the GPCM.

Previous results discussed the IRT parameter values. The significance of the differences in category boundary estimates are discussed in terms of the intercept estimates from the slope-intercept form of the model. The category intercept estimates (d_{k-1}) were estimated with smaller standard errors, from the GPCM than from the NRM at all category levels within each scale. The average standard errors of the estimated category intercepts from the GPCM were 0.012, 0.157, and 0.242 and 0.112, 0.196, and 0.336 from the NRM for $d_1, d_2,$ and d_3, respectively, on the anxiety scale. On the depression scale, the standard errors also increased on successive categories; on average these were 0.100, 0.142, and 0.232 for the GPCM and 0.110, 0.171, and 0.229 from the NRM. This increase in errors on successive categories is most likely due to the lower response rate of the highest category. Standard errors of the category intercept parameters were largest, on average, on the perceived stress scale items. When the GPCM was used, average errors were 0.141, 0.177, 0.222, and 0.338 for $d_1, d_2, d_3,$ and d_4, respectively; when the NRM was used, these were 0.200, 0.236, 0.297, and 0.415.

The category intercept estimates were significantly different for the GPCM and NRM estimates for d_3 on Items 7 and 17 from the anxiety scale, d_3 on Item 8 from the depression scale, and for various categories of 14 items on the perceived stress scale. (Note, d_3 is not the CBL that was previously discussed; see the section ''Slope–Intercept Form'' for clarification.) Items 7 and 17 on the perceived stress scale (''I have been able to control irritations'' or ''unable to control irritations in your life,'' respectively) from the NRM had estimated slope and discrimination values that were not ordinal, and the fourth category was not estimated to be a response from any respondent along the θ scale. Items that had significantly different category intercepts were from those that also had significantly different slope estimates; there were no items for which the estimated intercept was significantly different and the

slope was not. This is most likely due to the dependency of the estimated category intercept on the category slope.

Objective 3: Comparing Item Parameter Estimates for Pairs of Reverse-Worded Items

Overall, the estimated discrimination values for the PW items tended to be larger than the discrimination values for the NW items for both the GPCM, when a single a-value was estimated, and the NRM, when discrimination was estimated at each category change. Additionally, the category boundary locations (CBLs) tended to range from the middle to upper end of the trait scale for the NW items, whereas the PW items were more likely to have a lower boundary on the negative side of the scale.

Information. IICs were different for some pairs of PW and NW reverse-worded items when the same model was applied, while curves were similar for other pairs of items. Unfortunately, there was no clear pattern of the effect of the direction of the item wording. Shown in Figure 2, the IICs of PW items (gray) tended to be more leptokurtic, and information from NW items (black) tended to be more platykurtic. For some pairs of items (1 and 11, 2 and 12, 4 and 14, 6 and 16), IICs from the NW items (black lines) tended to be more platykurtic, providing less information at the peak, but more information across a wider range of the θ scale. For other pairs, the NW items tended to provide as much (e.g., Items 8 and 18) or more information (e.g., Items 9 and 19) as the paired PW item. This trend was not consistent for a specific model. IICs of matched items from the anxiety and depression scales tended to be more similar than those from the perceived stress scale.

Discrimination. Overall, the PW items tended to have a higher estimated discrimination value than the NW items when the GPCM and NRM were applied. Within pairs of items, the categories did not discriminate similarly for many pairs when estimated with either IRT model, and for some pairs the estimated discriminations were similar across pairs for one model and not the other. On the anxiety scale, Item 1 was NW and had an estimated discrimination from the GPCM model of $\hat{a} = 0.88$; the corresponding PW item 11 had an estimated discrimination of $\hat{a} = 2.12$. This indicated that the original NW item was less informative than the PW item, and that the PW item provided a more reliable estimate of the level of anxiety. The NRM allowed the discrimination between adjacent categories to vary. For Item 1 the CBDs were 0.82, 0.91, and 1.03 for adjacent categories 1 and 2, 2 and 3, and 3 and 4, respectively. The distinction between Categories 3 and 4 was more informative to the estimate of anxiety than was the distinction between other adjacent categories. For Item 11, the CBDs were 2.00, 2.15, and 2.03; for this item, the distinction between categories was highly informative between all adjacent categories.

Significantly different slope parameters were estimated for some pairs of PW and NW items for each model and scale. On the anxiety scale, Items 1 and 11 (‘‘I feel

more nervous and anxious than usual'' and ''I feel calm and secure'') had significantly different slopes when estimated with either model; other pairs of items on the anxiety scale were only significantly different when estimated with the GPCM (Items 4 and 14 and Items 6 and 16) or only from the NRM (Items 7 and 17 and Items 9 and 19).

Six of the 10 pairs of items from the depression scale had significantly different estimated slopes when estimated with the GPCM and only two pairs were significantly different when the NRM was applied. Item slopes were also significantly different for eight item pairs from the perceived stress scale for one or both models. There was no trend of the estimated slope of the PW item consistently estimated with a significantly higher or lower slope than the NW items.

Category Boundary Location. Trends of the comparisons of CBLs for pairs of PW and NW items were not consistent for the three scales. On the anxiety scale, the CBL of each category of the PW item tended to be at a lower value of the θ scale than the NW item for most pairs of items (excluding Items 3 and 13), when comparing estimates from the same model. The trend was not as apparent on the depression scale, yet the CBL between the middle categories (second and third) were very similar across six of the 10 pairs of items. In almost all cases, the CBL of the first and second categories was lower for the PW items than for the NW items. The CBL between the upper third and fourth categories did not follow a clear trend when comparing PW and NW item pairs. Overall, the PW items tended to be estimated with a wider range of the θ scale than those from the NW items. The perceived stress scale had five response categories. For all pairs, except the last CBL of Items 7 and 17, the estimated CBL for each category tended to be higher for the PW than for the NW items. This was the opposite trend compared to observations for the anxiety and depression scales. On the NRM, the same was true for the lower category CBLs. However, when the NRM was applied, the CBL between Categories 4 and 5 was extremely high or extremely low, as previously discussed. This indicated that the last category response is not likely for any examinee, even when θ is high, and thus not useful in estimating high levels of perceived stress.

The CBLs were calculated using a transformation of the IRT parameterized location estimates. These IRT parameters were transformations of the slope and intercept estimates from the models. Only standard errors of the slope and intercept estimates were obtained, so significantly different corresponding category intercepts for PW and NW items were evaluated. For nine item pairs from the anxiety scale, seven pairs from the depression scale, and all 10 pairs from the perceived stress scale, the estimated category intercept for at least one category was significantly different for pairs of reverse-worded items, and in many cases for all categories. It should be noted that item pairs may have significantly different category intercepts, but that does not necessarily imply they also have significantly different category location parameters due to the relationship between parameters in the different formulations.

Conclusion

In general, the NRM fit the data for the anxiety, depression, and perceived stress scales better than the GPCM. The NRM provided more information for the majority of the participants, including most of the low- and moderate-scoring respondents for the three scales, whereas the GPCM tended to only provide more information for respondents who were more than one standard deviation above the mean on the scales.

The NRM model allows for estimation of the discrimination for each category among adjacent categories, whereas the GPCM model estimates a single discrimination per item. Adjacent category boundary discriminations were calculated from the NRM estimates to better evaluate the effectiveness of the NRM over the GPCM. Item CBDs for the highest category, corresponding to the highest indicator of the trait, may be smaller than the next-to-the-highest CBD for the anxiety and depression scales, and even negative for many items on the perceived stress scale. This indicates that a person with a higher level of the anxiety, depression, or stress trait is more likely to choose the next to the highest response option rather than the highest response (e.g., ''a good part of the time'' vs. ''most of the time''; ''fairly often'' vs. ''very often''). This was also observed in the ICC comparisons of items and the CBLs. In these cases, the highest category response options are not likely to be used by anyone along the theta scale, and are not useful in adding information to their latent trait estimates. Studies are also needed to determine the impact of categories with low response-rates on polytomous IRT models.

The use of the NRM might be preferred over the GPCM in this situation where traits that are being measured tend to vary significantly for only a portion of the sample, and certain ranges of the item response categories may not function accurately for a subgroup within the sample. The NRM allows the user to identify items in which certain categories do not function effectively within the population. The results indicate that the incidence of high levels of anxiety, depression, and perceived stress is relatively low for this sample. This might be expected in the general population. These results might be expected to differ with a sample from a clinical population at-risk for anxiety, depression, and stress.

In comparing PW and NW items, PW items may be more effective than NW items in differentiating between subjects having different levels of a trait, but NW items tend to provide a more reliable estimate of the trait across a wider range of the trait scale. Significance tests indicated that the slope estimates (corresponding to discrimination) were rarely significantly different for the anxiety and depression scales using GPCM and NRM estimates. However, many of the perceived stress items had significantly different slope estimates using the two models. Corresponding intercept estimates (related to category boundaries) of PW and NW pairs of items tended to be significantly different from many items on the anxiety and perceived stress scales, but rarely on the depression scale. Inconsistent results across instruments causes those applying IRT evaluate items and score respondents to cautiously select the model, as estimates may vary greatly for some scales or constructs, yet be very similar for others. Future studies may investigate further why such differences are present

for some datasets, but not others, when the measured constructs are related. For example, effects of nonnormal distributions on polytomous IRT model estimations may be studied further.

Certain pairs of PW and NW items may function comparably when reverse-worded, but others could result in different estimates and potential outcomes. Instrument developers who plan to create scales in which PW and NW items are used are encouraged to conduct trials with matched pairs to identify which ones can be used interchangeably and which PW or NW items may be problematic or introduce sources of error that might diminish the effectiveness of their scale. A great debate remains on the use of items worded in the same direction on a scale or a mixture of PW and NW. Results of this study may support the use of scales comprised of items worded in the same direction for certain constructs, while supporting the use of mixed items for others.

Authors' Note

As of 2017, author Ki Lynn Matlock is publishing works under Ki Lynn Matlock Cole.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A comparison of graded response and Rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics*, *25*, 253-270.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143-156.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

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

Cohen, S. (1994). *Perceived stress scale*. Retrieved from http://www.mindgarden.com/documents/PerceivedStressScale.pdf

Cohen, S., & Williamson, G.M. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The Social Psychology of Health* (pp. 31-67). Newbury Park, CA: Sage.

Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C. H., Crane, P. K., . . . Reeve, B. B. (2007). IRT health outcomes data analysis project: An overview and summary. *Quality of Life Research*, *16*, 121-132.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, *60*, 151-174.

de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*, 275-288. doi:10.1177/0146621603253188

DeMars, C. E. (2004, April). *A comparison of the recovery of parameters using the nominal response and generalized partial credit models*. Poster presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The loneliness questionnaire–short version: An evaluation of reverse-worded and non-reverse-worded items via item response theory. *Journal of Personality Assessment*, *94*, 427-437. doi:10.1080/00223891.2012.662188

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*, 350-365.

Karim, J. (2010). An item response theory analysis of Wong and Law emotional intelligence scale. *Procedia—Social and Behavioral Sciences*, *2*, 4038-4047. doi:10.1016/j.sbspro.2010.03.637

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, *71*, 523-550.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Reise, S. P., & Henson, J. M. (2010). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, 18, 93–103. doi: 10.1207/S15327752JPA8102_01

Revelle, W. (2015). *Package 'psych'*. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation scale. *Psychological Assessment*, *16*, 169-181. doi:10.1037/1040-3590.16.2.169

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville, TN: Department of Psychology, University of Tennessee.

Taylor, J. M. (2015). Psychometric analysis of the ten-item perceived stress scale. *Psychological Assessment*, *27*, 90-101.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *2*, 161-176.

Tokuda, Y., Okubo, T., Ohde, S., Jacobs, J. Takahashi, O. Omata, F., . . . Fukui, T. (2009). Assessing items on the SF-8 Japanese version for health-related quality of life: A

psychometric analysis based on the nominal categories model of item response theory. *Value in Health*, *12*, 568-573.

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effect in mixed-format scales. *Educational and Psychological Measurement*, *75*, 157-178. doi: 10.1177/0013164414528209

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research*, *30*, 72-91.

Yurekli, H. (2010). *The relationship between parameters from some polytomous item response theory models* (Master's thesis). Florida State University, Tallahassee. Retrieved from http://diginole.lib.fsu.edu/islandora/object/fsu%3A253923

Zung, W. W. K. (1971). A rating instrument for anxiety disorders. *Psychosomatics*, *12*, 371-380.