## Title

Response Category Functioning on the Health Care Engagement Measure Using the Nominal Response Model.



## Journal

Assessment, 30(2)

## Authors

Reise, Steven
Hubbard, Anne
Wong, Emily
et al.

## Publication Date

2023-03-01

## DOI

10.1177/10731911211052682

Peer reviewed

Published in final edited form as:
*Assessment*. 2023 March ; 30(2): 375–389. doi:10.1177/10731911211052682.

# Response Category Functioning on the Healthcare Engagement Measure Using the Nominal Response Model

**Steven P. Reise**[1], **Anne S. Hubbard**[1], **Emily F. Wong**[1], **Benjamin D. Schalet**[2], **Mark G. Haviland**[3], **Rachel Kimerling**[4]

[1]Department of Psychology, University of California, Los Angeles

[2]Northwestern University, Feinberg School of Medicine

[3]Department of Psychiatry, Loma Linda University School of Medicine

[4]National Center for PTSD and Center for Innovation to Implementation, VA Palo Alto Health Care System

## Abstract

As part of a scale development project, we fit a nominal response item response theory model to responses to the Healthcare Engagement Measure (HEM). When using the original 5point response format, categories were not ordered as intended for 6 of the 23 items. For the remaining, the category boundary discrimination between category 0 (*Not at all true*) and 1 (*A little bit true*) was only weakly discriminating, suggesting uninformative categories. When the lowest two categories were collapsed, psychometric properties improved greatly. Category boundary discriminations within items, however, varied significantly. Specifically, higher response category distinctions, such as responding 3 (*Very true*) vs. 2 (*Mostly true*) were considerably more discriminating than lower response category distinctions. Implications for HEM scoring and for improving measurement precision at lower levels of the construct are presented as is the unique role of the nominal response model in category analysis.

### Keywords

Nominal Response Model; Item Response Theory; Patient Engagement; Item Discrimination; Category Boundary Discrimination

The overarching goal of this research is to apply the item response theory (IRT) nominal response model (*NRM*; Thissen, Cai, & Bock, 2010; Bock, 1997; Bock, 1972) to evaluate response category functioning on the Healthcare Engagement Measure (HEM). The HEM has 23 items designed to assess the propensity to engage with care (Kimerling, Lewis, Javier, & Zulman, 2020; Schalet, Reise, Zulman, Lewis, & Kimerling, 2021). Our specific goals are to: (a) evaluate whether the response categories are ordered as intended, (b) evaluate the relative quality of each response category by testing whether category boundaries within items vary in discrimination, and (c) establish a set of category scoring

weights that are monotonically related to IRT latent trait estimates. We first describe the engagement construct.

## Patient Engagement

The patient engagement construct (Carman et al., 2013; Gruman et al., 2010) addresses participatory behaviors that can optimize the benefit from healthcare services, such as communication, shared decision-making, and health promoting behaviors (e.g., diet, exercise, and medication adherence). This type of participation in health care allows providers and systems to better align services with patient knowledge, skills, social circumstances, and preferences for their care, which can make care more efficient and effective (Berwick et al., 2008). The measurement of patient engagement has the potential to personalize healthcare, enhance population health management, and serve as a quality measure that is applicable to patient populations with a wide range of conditions and comorbidities.

The National Academy of Medicine's *Vital Directions* initiative has identified patient engagement as one of 15 core measures for healthcare systems (Dazu et al., 2017) and the absence of a candidate measure a critical research gap (Blumenthal & McGinnis, 2015). The HEM was developed to address this deficiency (Kimerling et al., 2020, Schalet et al., 2021). For candidate engagement measures such as the HEM, and other patient-reported performance measures, implementation for healthcare quality improvement means that providers and systems will be accountable for maintaining or achieving certain patient outcome benchmarks. These are high stakes; thus, comprehensive psychometric evaluations are required (Squitieri et al., 2017).

## Present Study: Exploring Category Functioning with The Nominal Response Model

The HEM includes 23 items written at a 6$^{th}$ grade reading level that yield a unidimensional score (Schalet et al., 2021). Proceeding from the operational definition of engagement as behaviors that optimize benefit from healthcare services (Gruman et al., 2010) that are influenced by patients' healthcare systems and social contexts (Carman et al., 2013), the HEM elicits self-efficacy judgments for engagement behaviors (see Table 1). Because these judgments are context-sensitive (Bandura, 2004), better self-efficacy should identify individuals more likely to engage with care. For reviews of construct validity, fitting unidimensional and multidimensional factor and IRT models, evaluation of statistical fit, and differential item functioning, see Kimerling et al. (2020) and Schalet et al. (2021).

As noted, the goal of this study is to address three category functioning issues. The first is to test whether the response categories are ordered as intended; that is, do higher category responses reflect higher standing on the latent trait? In the overwhelming majority of psychometric reports, category ordering is simply assumed; rarely is it tested. In our intensive cognitive testing of potential items and response formats (Schalet et al., 2021), however, based on five categories, we found that participants had difficulties using certain commonly-used response formats, such as frequency, agreement, or confidence

anchors. Frequency anchors, for example, conflated engagement with utilization, and when agreement anchors were substituted, think-aloud responses suggested acquiescence bias. Ultimately, we settled on a set of "novel" (to us) anchors for each prompt (Table 1): 0 = *Not at all*, 1 = *A little bit true*, 2 = *Somewhat true*, 3 = *Mostly true*, and 4 = *Very true*. Clearly, these anchors assume people can reliably assess their engagement behaviors as being on a continuum where terms such as "*a little bit*" and "*somewhat*" are meaningful and can be reliably distinguished.

Given we departed from commonly-used anchors in the development of the HEM (Schalet et al., 2021), more comprehensive analyses of category functioning were required. Not only would we need to test ordering assumption of the categories using the *NRM*, we would need also to explore category functioning in greater detail. Our second research question, thus, addresses the relative quality of each category distinction by testing for the equality of category boundary discrimination within each item. Category boundary discrimination involves judging how discriminating responses are in each of the 4 adjacent categories (i.e., how discriminating is a response of 1 vs. 0, 2 vs. 1, 3 vs. 2, and 4 vs. 3).

When category boundary discriminations are equal within items, this suggests that all categories are functioning with equal effectiveness in differentiating individual differences across the trait continuum. In turn, such evidence supports both the decision to use five response categories and the wording of the selected anchors. On the other hand, when category boundary discriminations vary within items, this suggests uninformative, non-discriminating categories (too many response options) or a poor anchor wording. Note, that as with category ordering, equality of category boundary discrimination within items must be assumed in commonly-applied IRT models, such as in the graded response model (GRM; Samejima, 1969).[1] Our investigation of category boundary discrimination on the HEM is partially based on our belief that such analyses should be standard practice in new instrument development. It is also partially based on our theory of the engagement construct and previous research findings. Specifically, we conceive patient engagement to be a unipolar construct[2] as opposed to a bipolar construct (i.e., a construct that is more meaningful at one end of the continuum and ambiguous or absent at the other end, such as gambling addiction; Lucke, 2014). In turn, we believe that this "unipolar" aspect may lead to the categories being differentially discriminating, as they have been in similar unipolar patient-reported outcome measures. Examples include Preston, Reise, Cai and Hays (2011); they examined the category boundary discrimination within items that are part of the Patient-Reported Outcomes Measurement Information Systems (PROMIS) Depression, Anxiety, and Anger item pool[3] (Pilkonis et al., 2010). With a primary goal of analyzing category boundary discrimination variation, they found that 25 of the 86 items examined had significant variation – higher categories tended to be more discriminating (see also, Preston & Reise, 2015). In turn, they argued that such findings could be attributable to the unipolar

[1]In fact, lack of ordering of threshold parameters and variation in category boundaries are meaningless concepts in the *GRM* and cannot be tested because it is a cumulative boundary homogeneous model.
[2]Prior work suggests that patient engagement was conceptually orthogonal to patient disengagement (Kimerling et al., 2020).
[3]These are arguably unipolar because the low end is absence of symptoms, not happiness, calmness, or pleasantness.

nature of psychopathology constructs. We will elaborate on unipolar traits, response formats, and category boundary discrimination in more detail in the discussion.

Finally, as part of investigating category ordering and category boundary discrimination using the *NRM*, the analyses naturally yield a set of "scoring weights" for each category that reflect the effect of each category response on the estimation of the latent trait. As we show below, these "optimal" weights can be used to create a set of weighted summed scores that are, in turn, perfectly monotonically related to the IRT latent trait estimate. We argue that these scoring weights are not only complementary to IRT category, item, and test information analyses, they are more useful in terms of judging item quality and utility than the typically-reported table of factor loadings or item-test correlations.

## The Nominal Response Model

To understand how the *NRM* is used to address these issues, we now provide a description of the technical details of the model and a brief review of previous applications. We borrow heavily from the notation in Thissen, Cai, and Bock (2010). To illustrate, we describe the *NRM* applied to Item #1 (*I know I can always follow my doctor's instructions*) when scored using four categories (0 = *Not at all/A little bit true*, 1 = *Somewhat true*, 2 = *Mostly true*, and 3 = *Very true*).

Like all IRT models, the chief objective of the *NRM* is to develop a set of functions, called category response curves (CRCs) that relate individual differences on a latent trait (in the present case, Patient Engagement), symbolized by $\theta$, to the probability of responding to an item in a specific way (e.g., responding in the 3rd category). For a $K = 4$ response category item, in the original *NRM* parameterization (Bock, 1972), the relation between trait level and the log-odds of responding in a particular response category $k(k = 0 \ldots K - 1)$ is written as a linear function:

$$z_k = \exp(a_k\theta + c_k) \quad (1)$$

Where, $a_k$ is the slope and $c_k$ is the intercept of the line relating trait level ($\theta$) to the log-odds of responding in a category $k(k = 0 \ldots K - 1)$, symbolized by $z_k$. For reasons soon to be clear, the spread or variance of the $a_k$ parameters within an item reflects the item's ability to discriminate among individuals along the trait continuum, and the intercept reflects relative category popularity (i.e., proportion responding in a given category). Typically, the scale for the trait is specified to have a mean of zero and standard deviation of 1.0 in the population, and, thus, it can be interpreted like a *z*-score. One $a_k$ and $c_k$ parameter needs to be estimated for each response category. To accomplish this, a constraint is needed such as the sum of $a_k$ values equal 0 and the sum of $c_k$ values equal 0. For Item #1, these are $a_k = -1.13$, $-0.61$, 0.24, and 1.49, and $c_k = -1.81$, $-0.38$, 1.28, and 0.92. Once these parameters are estimated using a program such as ***mirt*** (Chalmers, 2012), CRCs reflecting the probability of responding in a given category conditional on trait level can be derived as:

$$P(k=0\ldots K-1\,|\,\theta)=\frac{\exp(z_k)}{\sum_{k=0}^{K-1}\exp(z_k)}=\frac{\exp(a_k\theta+c_k)}{\sum_{k=0}^{K-1}\exp(a_k\theta+c_k)} \quad (2)$$

The category response curves for Item #1 are shown in the top panel of Figure 1. For this "divide-by-total" model (Thissen & Steinberg, 1986), the $a_k$ and $c_k$ parameter values are difficult to interpret directly in terms of describing how well the items and category boundaries are performing in their relative discrimination and in testing whether the response categories are ordered as expected (i.e., whether responses in higher categories reflect higher trait levels as assumed). Therefore, these parameters are often transformed to more readily interpreted values (Thissen, Steinberg, & Fitzpatrick, 1989). Specifically, for an item with $K$ categories, $K-1$ category boundary (e.g., 0 vs. 1, 1 vs. 2, and 2 vs. 3) discriminations (*CBD*s; Preston, Reise, Cai, & Hays, 2011) and intersections are derived as:

$$a_{k*}=\alpha_k-\alpha_{k-1} \quad (3)$$

$$c_{k*}=(c_{k-1}-c_k)/a_{k*} \quad (4)$$

Where $a_{k*}$ is a *CBD* (0.52, 0.85, and 1.25 for Item #1) reflecting how discriminating the distinction between two adjacent response categories is (i.e., categories 0 vs. 1, 1 vs. 2, and 2 vs. 3). The $c_{k*}$ parameter is an "intersection" reflecting where the CRCs for two options intersect (−2.72, −1.94, and 0.28 for Item #1; see vertical lines in Figure 1) or the trait level where the higher category response ($k$) becomes more likely than the lower ($k-1$). Finally, a 2-parameter function for the dichotomous distinction between category $k$ and $k-1$ can be written as:

$$P(k|k,k-1)=\frac{1}{1+\exp\left(-a_k^*(\theta-c_k^*)\right)} \quad (5)$$

Which equals a 2-parameter item response curve showing the probability of response $k$ (the higher category) given that responses are either in $k$ or $k-1$. Thus, adjacent categories are only ordered if $a_k$ are ordered, because only then will the *CBD* ($\alpha_k-\alpha_{k-1}$) be positive. For the present example, the *CBDs* are: 0.52, 0.85, and 1.25, respectively, indicating that the categories, indeed, are ordered from 0 to 3 and as trait levels increase, the probability of responding in the higher adjacent category increases (we show these *CBD* curves based on Equation 5 in the bottom of Figure 1). Moreover, inspection of the *CBD* reveals that a response in category 3 (vs. 2) is more ($CBD=1.25$) is two and a half times more discriminating than a response in category 1 (vs. 0) where $CBD=0.52$. Implications of this follow.

More recently, there is a new and more useful parameterization of the *NRM* presented in Thissen, Cai and Bock (2010). This parameterization not only allows the nominal model to

be extended to the multidimensional case (e.g., Falk & Ju, 2020), it makes clear the relative psychometric strength both between items and between categories within items. Specifically, the new parameterization

$$z_k = a_{i*}a_k S\theta + c_k \quad (6)$$

Where, $a_{i*}$ is the item slope reflecting the overall discrimination capacity of the item, $a_k S$ is the "scoring coefficient" (Muraki, 1992) for response category $k$, and $c_k$ is the intercept parameter equal to the original parameterization. To identify the model for estimation, $a_0 S$ and $c_0$ are set to 0, and $a_{K^S-1} = K - 1$. For the example item, the slope, $a_{i*} = 0.87$. In turn, this value can be compared to other items to judge relative discrimination, such that higher values of $a_{i*}$ indicate more discriminating items. The "scoring coefficients" are $ak_0 = 0$, $ak_1 = 0.60$, $ak_2 = 1.57$, and $ak_3 = 3$. They are called "scoring coefficients" because when multiplied by the item slope, they indicate the "scoring weight" one should give each category response to yield an "optimal" (most reliable) weighted composite that is perfectly monotonically related to an IRT trait level estimate, as demonstrated shortly. For example, for Item #1, a response in category

$$0 \text{ is scored } 0 * 0.87 = 0,$$

$$1 \text{ is scored } 0.60 * 0.87 = 0.52,$$

$$2 \text{ is scored } 1.57 * 0.87 = 1.37, \text{ and}$$

$$3 \text{ is scored } 3 * 0.87 = 2.61.$$

Notice that the distances between each scoring weight within an item are not necessarily equal; the larger the distance, the more discriminating the higher category (see Anderson, 1977 and Andrich, 1978, for a more technical discussion). For the example item, a transition from 0 to 1 has a scoring weight of .52 (exactly equal to the $CBD_1 = 0.52$), a transition from 1 to 2 increases the scoring weight to 1.37 (a difference equal to $CBD_2 = 0.85$), and a transition from 2 to 3 has a scoring weight of 2.61 (a difference equal to $CBD_3 = 1.25$). That is, the scoring weights simply reflect the slope of the *CBD*s which, in turn, reflect the distance between category slopes ($a_k$) in the original parameterization. Thus, the more spread the $a_k$ in the original parameterization, the larger the slope of the *CBD*s, and more importantly, in a psychometric sense, a response in a particular category, such as category 1, 2, or 3, has a larger effect on the trait level estimate, as we will show in the results section.

Finally, we note that important submodels can be easily derived from this modern parameterization by first fixing the $a_k S$ scoring coefficient parameters to be equal distance (e.g., 0, 1, 2, 3) between items. This yields a generalized partial credit model (*GPCM*;

Muraki, 1992) where items may differ in discrimination ($a_i^*$) but categories are assumed equally discriminating. An even more constrained model can then be specified by forcing the $_i^*$ to be equal across items (and maintaining the identification constraint that the variance of the latent trait is 1.0). This partial credit model (*PCM*; Masters, 1982) assumes that all items are equally discriminating as well as all categories within items. Under the *PCM*, unit-weighted composite scores are a sufficient statistic for trait level estimates.

## Method

### Measure

Table 1 shows the item content for the 23-item HEM (Schalet et al., 2021). There are five response options: "*Not at all true*", "*A little bit true*," "*Somewhat true*," "*Mostly true*," and "*Very true*." Our provisional scoring rule assigns numbers from 0 to 4 the respective categories.

### Participant dataset

The participant data ($N = 7,122$) were based on a national mail survey of adult (aged 1881) users of Veteran Administration (VA) health care facilities in the contiguous United States. The data obtained represented users from 136 medical centers, with care for one or more of the following four conditions in the past year: hypertension, diabetes, depression, or post-traumatic stress disorder. Survey respondents were 79% male, 15% Hispanic or Latinx, 26% Black or African-American, and 59% White. The highest education level was high school or GED for 24% of the sample, with 32% college graduates and beyond.

### Analyses

Previous published analyses demonstrated that the 23-item set was sufficiently unidimensional for IRT parameter estimation (Schalet, et al., 2021).[4] Our analyses proceeded by examining the *NRM* parameter estimates for the 5-category data to evaluate whether the items were ordered as intended. We also fit more constrained models (*PCM* and *GPCM*) and tested these relative to the *NRM*. We scored the data (i.e., estimated latent trait levels) with expected a posteriori (EAP; Bock & Mislevy, 1982) estimation and plotted a variety of information and characteristic curves to clarify category functioning. All analyses were conducted in *R*, with basic psychometrics in the *psych* package (Revelle, 2019) and IRT analyses in *mirt* (Chalmers, 2012).

## Results

### Analysis of 5 Response Categories:

Our first objective was to evaluate whether all five original response options were ordered as intended (i.e., higher options reflect higher levels of Patient Engagement) and contribute meaningfully to measurement precision. We, thus, estimated the *NRM* based on all five response options. Prior to reviewing these results, we note that responses in category 0 (*Not at all true*) were exceptionally rare with less than 5% responding in this lowest category on

[4]Note that all analyses reported here were conducted prior to Schalet et al. (2021), and they informed all the analyses in that report.

most items. With that noted, results of the *NRM* are shown in Table 2. The first column shows the overall item slopes ($a^*$); $ak_0$ to $ak_4$ show the scoring coefficients that are to be multiplied by the item slope to obtain the scoring weights. The $c_0$ to $c_4$ are category intercepts, which can easily be transformed into intersections, as described previously.

The distinction between the first two categories was very small, average scoring coefficient = .01 (but note, 6 values were negative making the mean less meaningful). Although not shown in Table 2, the average *CBD* was 0.04, 0.63, 1.02, and 1.69, respectively. In other words, the distinction between category 0 and 1 is not at all informative or discriminating in terms of trait levels. Perhaps more importantly in Table 2 is the observation that six items (#1, 3, 5, 6, 14, and 18) had negative $ak_i$ parameter estimates indicating a lack of ordering among the first two categories. For these items, it is not clear that responding in category 1 (*A little bit true*) refers to higher trait levels than responding in category 0 (*Not at all*).

We proceed after collapsing categories 0 and 1, thus, creating items with four categories, which we use in all subsequent analyses. The collapsed responses produced higher factor loadings, item-test correlations, item intercorrelations (.41 vs .40), and higher item slopes ($\bar{a}^* = 1.12\ vs.\ 0.85$), and as we show below, higher *CBD* for the first category distinction. In short, the psychometric properties of the measure improved; five categories may have caused nuisance variance.

## Analysis of 4 Response Categories:

We now turn to the psychometric analysis of the 4 response category data. In Table 3 are shown item-test correlations, item means, and standard deviations. Also shown are factor loadings and communalities (1 minus loading squared). Observe that Items #10 (*I know I can get a provider to deal with my main health concerns*), #11 (*I can make sure my concerns are fully addressed before I leave appointments*), #20 (*I can get the care I needed without getting discouraged*), and #21 (*I know I can get the information I need about the pros and cons of treatment*) had the highest loadings and item-test correlations, and Item #5 (*It is easy for me to refill medications on time*) had the lowest.

We next fit the *PCM*, *GPCM*, and *NRM* to the 4-category data. Results are shown in Tables 4, 5, and 6, respectively. In Table 4, the *PCM* model results show an average item slope of 1.19 (note: variance was fixed to one and a constant slope estimated), and scoring coefficients are simply the integers 0, 1, 2, and 3 (i.e., $ak_0 \ldots ak_3$ would be multiplied by 1.19 in weighted scoring). In Table 5, the *GPCM* results are similar to factor loadings and item test correlations in Table 3 in that Items #10 ($a^* = 1.67$), #11 ($a^* = 1.88$), #20 ($a^* = 1.89$), and #21 ($a^* = 1.99$) had the highest slopes, whereas item #5 ($a^* = 0.80$) had the lowest. To clarify what those results imply about scoring under the *GPCM*, consider Items #5 and #20. For Item #5, scoring weights would be:

$$0.80 * 0 = 0, 0.80 * 1 = .80, 0.80 * 2, = 1.60 \text{ and } 0.80 * 3 = 2.40.$$

On the other hand for Item #20, scoring weights would be:

$1.89 * 0 = 0, 1.89 * 1 = 1.89, 1.89 * 2 = 3.78$ and $1.89 * 3 = 5.67$.

Thus, equal item scores do not mean equal things between items under the *GPCM*. Finally, chi-square model comparison of the *PCM* vs. *GPCM* was 2,451 on 22 *df*, $p = .00$, and AIC and BIC also favored the *GPCM*. Thus, in a statistical sense, the *GPCM* is favored over the more restricted *PCM* model suggesting that the items vary in overall discrimination.

The top of Table 6 displays the new parameterization for the *NRM*. The chi-square test comparing the *NRM* with *GPCM* yielded 2,056 on 46 *df*, p = .00 and AIC and BIC also favored the *NRM*.[5] We, thus, conclude that at least in a statistical sense, the *NRM* is superior. As with the *GPCM*, items vary significantly in slope with an average of 1.12. Noting that item information is (very) roughly a function of the square of the item slope, some items (e.g., #21) provide almost eight times the psychometric information as other items (e.g., #5); $.64^2 = .41$ vs. $1.70 = 3.20$.

Moreover, comparison of *GPCM* and *NRM* fit suggest that category boundary discriminations within items also varied. As shown in the bottom of Table 6, scoring weights averaged 0.00, 0.64, 1.66, and 3.35 for categories 0 to 3, respectively, and *CBD*s averaged 0.64, 1.02, and 1.70, for distinctions, 0 (*Not at all true or A Little Bit True*) vs 1 (*Somewhat true*), 1 (*Somewhat true*) vs. 2 (*Mostly true*), and 2 (*Mostly true*) vs. 3 (*Very true*), respectively. The scoring weight is a direct index of the responses effect on an individual's trait level estimate. Thus, whether judged by the distance between scoring coefficients (top Table 6), scoring weights, or when the results are converted to *CBD*s, responses in higher categories are more discriminating and with greater effects (i.e., spread people out more) than responses in lower categories, especially a response in category 3 (*Very true*) vs. 2 (*Mostly true*), which had an average *CBD* = 1.70. Note, the degree to which higher categories result in higher scoring weights depends on, or is moderated by, the overall item discrimination.

To make the concept of scoring weight clearer, in Figure 2 is shown the distribution of EAP estimated trait levels (top) and the perfect monotonic relation between weighted composite scores derived from the scoring weights and EAP trait levels (bottom). Clearly, the EAP level estimates have an upper ceiling effect, and individuals cannot score above 2.2 even if they score in the highest category on every item, as many do. This is due to not having enough information in the very highest trait ranges to distinguish the very high from the high (if such a distinction is even meaningful). Shown on the bottom of Figure 2 is the perfect monotonic relation between weighted composite based on the *NRM* scoring weights and EAP trait level estimates. Not only does such a curve demystify the IRT trait scale (i.e., it is just a weighted composite rescaled to a zero one metric), it also makes clear why we can refer to scoring weights as response "*impacts*" – the larger the scoring weight for a given category, the more the trait level estimate changes. Finally, this figure makes clear why the weighted raw score is a sufficient statistic for the latent trait under the *NRM*.

[5]It is also appropriate to note that, although the NRM may fit better and be more valid, this does not mean that all model applications are materially different; for example, trait level estimates based on the PCM, *GPCM*, and *NRM* are correlated above .98.

To probe these *NRM* results in more detail, in Figure 3 we show both the category response curves (top) for all items and well as the item information curves (bottom). Notice that for some items, such as #1, the curves for categories 2 and 3 dominate the trait continuum, suggesting that, perhaps, these items should be dichotomously scored. The category response curves for categories 0 and 1 indicate that responding in these categories is relatively unlikely at all trait ranges, even low trait ranges. In contrast, for items such as #17 to #22, all of the categories have at least some range of the trait for which the response is most likely. The consequence of this is very clear in the lower graph in that items like #20 and #21 provide considerably more psychometric information (i.e., reduction in error) than items such as #4 and #5.

To understand the broader effects of these results on measurement precision, in Figure 4 (top) is displayed the overall test information curve for the *NRM*. The overall curve is peaked, implying that the items measure best in the middle of the trait continuum. Note that one divided by the square root of information is roughly equal to the conditional standard error of measurement. Where does that information or measurement precision come from? The lower curves in Figure 4 show the test information from responses in categories 0, 1, 2, and 3. Clearly, 3 is dominating, suggesting that this category is most meaningful, which we know to be true because the *CBD* for the distinction between 2 and 3 is 1.70.

To amplify, in Figure 5 are shown the category response curves for Item #5 (one of the worst items with *CBD* = 0.30, 0.45, and 1.19, respectively) and Item #21 (one of the best items with *CBD* = 1.03, 1.69, and 2.65, respectively). Clearly, only options 2 and 3 are most likely for Item #5 suggesting this should be a dichotomy – there is no point on the latent trait where option 0 and 1 are most likely. Under these conditions it is impossible for those categories to provide much discrimination among people. On the other hand, for Item #21, all options are most likely somewhere along the continuum. Accordingly, each of the categories contributes meaningfully to measurement precision. The exact amounts are shown in Figure 6 with the category and item information for Item #5 on top and Item #21 on bottom. For Item #5, only categories 2 and 3 provide meaningful information. For Item #21, all categories are informative, especially 3.

## Discussion

We fit the *NRM*, as well as two more constrained models, the *GPCM* and *PCM*, to a sample of 7,122 responses to the 23-item HEM. The goal of these analyses was to provide a detailed psychometric analysis of response category functioning to: (a) evaluate category ordering, (b) evaluate relative category functioning by testing whether categories boundaries discriminations vary within items, and (c) establish a set of category scoring weights. Phrased slightly differently, we asked, are categories ordered, how discriminating are they, and are they equally discriminating within items? Below we address our findings for each of these questions in turn and comment on their implications for the HEM.

In the development of patient-reported outcomes, a critically important consideration is how many response categories to have and how to verbally anchor them (Krosnick & Fabrigar, 1999). Evaluation should address whether the categories are ordered as intended

and whether each is providing a useful discrimination of individual differences. This is especially important when applying novel response formats, as is the case with the HEM.

Our results show that the HEM items were better scaled on 4 rather than 5 response categories. With 5 categories, fitting the *NRM* did not support the hypothesis that the response categories were ordered for all items such that higher item scores imply higher trait levels. For six items, the category responses of 1 (*A Little Bit*) and 0 (*Not at all true*) were reversed with a score of zero suggesting a higher trait level than a score of 1. For the remaining items, the CBD = .04 for the distinction between category 1 (vs 0), suggesting that a response in category 1 (vs. 0) does not discriminate well among individuals, or, equivalently, provide much psychometric information (i.e., error reduction). As a consequence, scoring coefficients (and weights) for category 1 (vs. 0) were very small. Thus, 5 options as we anchored them are too many.

In future applications of the HEM, we, thus, recommend scoring the instrument as 4 categories. When we collapsed the lower two categories, all categories were then logically ordered 0…3 for all items, and, thus, had positive *CBD* parameters (*CBD* = 0.64, 1.02 and 1.70). Moreover, scoring in four categories is supported by item slopes that show the items are more discriminating and, thus, informative compared to when scored using five categories. Although it is arguable that the results suggest eliminating the lowest response category, note that when response categories were evaluated in cognitive testing prior to this study (Schalet et al., 2021), some individuals found the items difficult to answer without the absolute lower bound response anchor of "*Not at all*." This extreme anchor may be helpful for some respondents with certain response styles or function as a context for other response categories, and, thus, should be retained at scale administration.

Beyond informing on category ordering, application of the NRM contributed important information regarding between and within item functioning. Model comparison tests of the *GPCM* and *PCM* suggested that items varied significantly in discrimination. Specifically, items varied widely in discrimination with four items providing much higher information (i.e., contribution to error reduction) than others (i.e., Items #10, 11, 20, and 21). These four items concern confidence in obtaining a provider and acquiring needed information. A few items had exceptionally low discriminations such as Items #5 and 6. These items address confidence in refilling medications and obtaining health care services, neither of which is directly related to a patient's engagement with their provider, *per se*.

For highly discriminating items, each category had at least some range of the latent trait where the category was most likely. In fact, the top 4 items are so superior to the remaining items, it appears that once one measures those aspects of the construct, other aspects may contribute relatively little information. This is a critical recognition for future short form construction. For some less discriminating items, such as #1 (*I know I can always follow my doctor's instructions*), graphs of category response curves (Figure 3) suggested that only two of the categories (2 and 3) were highly likely regardless of trait level (0 and 1 are likely only for the most extremely low respondents). In turn, such a finding suggests that the item functions like a dichotomy – if low respond 2, if high respond 3.

The *NRM* nicely captures this phenomenon in the category scoring weights, which are 0.00, 0.52, 1.38, and 2.62 for Item #1 (*CBD* = 0.52, 0.85, and 1.25) – responding in 1 increases weighted scores 0.52, responding 2 increases weighted scores more than double to 1.38, and responding 3 almost doubles that to 2.62. We believe that these occurrences raise the issue of whether the relatively lower overall discrimination for Item #1 is due to the content validity of the item (efficacy following a doctor's instructions) or that our particular response anchors are faulty in some way, not allowing people to discriminate at the lower end, or that in reality, there is no such behavior as "a little" or "somewhat" efficacy in following a doctor's orders – you "mostly" or "always" do. These are excellent topics for follow on studies.

Perhaps more interesting, model comparison tests revealed that the *NRM* provided a superior fit relative to the more constrained *GPCM* suggesting that *CBD* within items varied significantly, and, thus, the categories are not providing equal discriminations of individual differences across the trait range as intended. In fact, higher categories tended to have greater effects on trait level estimates. Specifically, the *CBD* parameters were 0.64 (0 vs. 1), 1.02 (1 vs. 2) and 1.70 (2 vs. 3). For highly discriminating items, such as #21, *CBD* were 1.03, 1.69, and 2.65. For poorly discriminating items, such as #5, *CBD* were 0.30, 0.45, and 1.19. In either case, judging by the fact that the first *CBD* (*Not At All True* or *A Little Bit True* vs. *Somewhat true*) is below 1.0 for 22 of 23 items, it appears that discrimination at the lower end of the engagement continuum may be a greater measurement challenge than differentiating between highly engaged patients.

An applied implication of finding varying *CBD* is that models, such as the *GPCM* or *GRM*, which assume equal *CBD*, may not be appropriate.[6] In terms of the design of the measure, such results may imply that the anchors used for the lower categories are sub-optimal, relative to the anchors used for the higher categories. Alternatively, and more substantively, we believe that the phenomenon of *CBD* increasing for higher categories may be due to the unipolar nature of the construct (Lucke, 2014) – at the higher end it is easier to make more reliable distinctions, whereas at the lower end, it is more ambiguous what behaviors or cognitions may differentiate relatively low from lower. This is a concern because the ability to discriminate between patients at the lower end of the engagement continuum and to track their change after intervention, is a primary concern; this range of scores identifies patients who are more likely to need adjunctive services or a greater intensity of care, where more highly engaged patients are confident expressing preferences, following treatment plans, and alerting providers to their needs regarding treatment options or self-management support.

Finally, our third application of the *NRM* are the scoring weights derived by multiplying the overall item discrimination by the scoring coefficient. We characterize scoring weights as "impacts" or "importance" weights in that they clearly show each response contributes to a weighted raw score. Forming a composite of item responses weighted by the scoring weights is a sufficient statistic for the IRT trait level estimate. From an applied perspective,

[6]Nevertheless, despite our findings here, in Schalet et al. (2021), the *GRM* was used to model HEM responses. Schalet et al. ultimately selected the (somewhat inappropriate) *GRM* for the HEM because it is the most-commonly applied and best understood model. In terms of ordering individuals on the trait scale, model choice made little if any difference, however; trait scores when estimated under the *NRM* and GRM correlated .99.

this is advantageous because it is much easier to compute a weighted raw score using a spread sheet than it is to obtain an EAP trait level estimate. Not only does this demystify the latent trait scale in IRT under the *NRM* as a monotonic function of weighted item scores, reporting category scoring weights, such as those in the bottom of Table 6, may often be more illuminating in terms of evaluating item value or quality and can nicely complement the analyses of IRT category and item information.

In terms of patient reported outcomes (PRO) implementation in healthcare settings, deriving an IRT-based trait level estimate from *NRM*-based scoring weights may have the potential to facilitate PRO integration into electronic medical records. To date, integration of PRO systems with electronic health records (EHR) are challenging, but PRO collections systems outside the EHR can be burdensome to providers and staff (Gensheimer et al., 2018). Trait level estimates more precise than traditional summed score to scale score conversion tables could potentially be calculated outside of a PRO system using typical EHR informatics, such as mapping scoring weights to standardized terminologies for measure items and responses.

## Conclusion

In developing and analyzing the HEM, the present analyses led us to three clear conclusions. First, five response categories, as we anchored them, appears to be too many – the lower two categories provide no discrimination of individual differences. Second, when collapsing the first two categories, and then scoring as four categories, psychometric properties improve; that is, all categories provide meaningful discriminations. There were, however, several items where the CRCs suggested that only two categories are used by respondents across a wide range of the trait. Third, higher categories provided substantively more discrimination than lower.

In terms of understanding these findings we offer two suggestions. First, we suggested that engagement may function more like a unipolar dimension than a bipolar dimension. One explanation of the lower discrimination of lower response categories is that the construct is not as well defined at the low end, and as a consequence, regardless of number of categories or the anchors, people can only distinguish among behaviors reflecting higher ranges of the construct.[7] An alternative explanation is that there are simply too many response options, or the anchors for the lower categories are not optimal in some way. These explanations are not mutually exclusive. Regardless, our findings should inform future HEM revisions or other engagement measures or similar health-related constructs.

Finally, we argue that the *NRM* is of great value in addressing three critical questions in scale development and analysis: (a) are the response categories ordered as intended, (b) are the categories within items equally discriminating (or provide differential discrimination), and (c) how does category response contribute to an optimally-weighted composite that will be perfectly monotonically related to the IRT latent trait estimate? In future scale

[7]Another explanation is that engagement is a complex construct in the sense that there are many ways to be less engaged, such as communication and self-management difficulties; whereas to be more engaged means that self-efficacy generalizes across the 3 domains (bifactor subfactors) of behaviors, and more highly engaged people are more likely to rate items uniformly in the upper two categories.

development efforts, we strongly suggest always testing the response options through application of the *NRM*, regardless of whether one intends to use the model for applied purposes. Too often, scale developers use 5-, 7-, or even 9-point response formats but do not provide any empirical evidence that each category provides unique, reliable, and valid information about trait standing. Routine *NRM* applications would address this commonly-overlooked issue.

## Research Support:

This work was supported by 1I01HX002317 from the United States (US) Department of Veterans Affairs Health Services Research and Development Service. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Partial support for Reise, Hubbard, and Wong, was also provided by MH118514 ("National Neuropsychology Network, R. Bilder, PI).

## References

Andrich D (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561–573.

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

Chalmers RP (2012). "mirt: A Multidimensional Item Response Theory Package for the R Environment." Journal of Statistical Software, 48, 1–29.

Bandura A (2004). Health promotion by social cognitive means. Health Education & Behavior, 31, 143–164. 10.1177/1090198104263660 [PubMed: 15090118]

Berwick DM, Nolan TW, & Whittington J (2008). The triple aim: Care, health, and cost. Health Affairs, 27, 759–769. 10.1377/hlthaff.27.3.759 [PubMed: 18474969]

Blumenthal D, & McGinnis JM (2015). Measuring vital signs: An IOM Report on Core Metrics for Health and Health Care Progress. JAMA, 313, 1901–1902. 10.1001/jama.2015.4862 [PubMed: 25919301]

Bock RD, & Mislevy RJ (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431–444.

Carman KL, Dardess P, Maurer M, Sofaer S, Adams K, Bechtel C, & Sweeney J (2013). Patient and family engagement: A framework for understanding the elements and developing interventions and policies. Health Affairs, 32, 223–231. 10.1377/hlthaff.2012.1133 [PubMed: 23381514]

Dzau VJ, McClellan MB, McGinnis J, & et al. (2017). Vital directions for health and health care: Priorities from a national academy of medicine initiative. JAMA, 317, 1461–1470. 10.1001/jama.2017.1964 [PubMed: 28324029]

Falk CF, & Ju U (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. Frontiers in Psychology, 11, 72. [PubMed: 32116902]

Gensheimer SG, Wu AW, Snyder CF, PRO-EHR Users' Guide Steering Group, & PRO-EHR Users' Guide Working Group. (2018). Oh, the Places We'll Go: Patient-Reported Outcomes and Electronic Health Records. The Patient, 11, 591–598. 10.1007/s40271-018-0321-9 [PubMed: 29968179]

Glasgow RE, Wagner EH, Schaefer J, Mahoney LD, Reid RJ, & Greene SM (2005). Development and validation of the Patient Assessment of Chronic Illness Care (PACIC). Medical Care, 43, 436–444. 10.1097/01.mlr.0000160375.47920.8c [PubMed: 15838407]

Gruman J, Rovner MH, French ME, Jeffress D, Sofaer S, Shaller D, & Prager DJ (2010). From patient education to patient engagement: Implications for the field of patient education. Patient Education and Counseling, 78, 350–356. 10.1016/j.pec.2010.02.002 [PubMed: 20202780]

Hibbard JH, Mahoney ER, Stockard J, & Tusler M (2005). Development and testing of a short form of the patient activation measure. Health Services Research, 40(6 Pt 1), 1918–1930. 10.1111/j.1475-6773.2005.00438.x [PubMed: 16336556]

Huntink E, Koetsenruijter J, Wensing M, & van Lieshout J (2019). Patient cardiovascular risk self-management: Results from a randomized trial of motivational interviewing delivered by practice nurses. Family Practice, 36, 460–466. 10.1093/fampra/cmy087 [PubMed: 30277507]

Kimerling R, Lewis ET, Javier SJ, & Zulman DM (2020). Opportunity or Burden? A Behavioral Framework for Patient Engagement. Medical Care, 58, 161–168, 10.1097/MLR.0000000000001240 [PubMed: 31688570]

Lucke JF (2014). Unipolar item response models. In Reise S & Revicki D (Eds). Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment New York: Routledge (p. 272–284)

Masters GN (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Muraki E (1992). A generalized partial credit model: Application of an EM algorithm. ETS Research Report Series, 1992(1), i–30.

Mutebi A, Slack M, Warholak TL, Hudgens S, & Coons SJ (2016). Interpretation of verbal descriptors for response options commonly used in verbal rating scales in patientreported outcome instruments. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25, 3181–3189. 10.1007/s11136-016-1333-3 [PubMed: 27294436]

Packer TL, Kephart G, Ghahari S, Audulv A, Versnel J, & Warner G (2015). The Patient Activation Measure: A validation study in a neurological population. Quality of Life Research, 24, 1587–1596. 10.1007/s11136-014-0908-0 [PubMed: 25557496]

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment, 18, 263–283. [PubMed: 21697139]

Preston KS, & Reise SP (2014). Detecting faulty within-item category functioning with the nominal response model. In Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (pp. 404–423). New York: Routledge.

Preston K, Reise S, Cai L, & Hays RD (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.

Revelle W psych: Procedures for Psychological, Psychometric, and Personality Research 1.9.12.31 ed. Evanston, IL: The Comprehensive R Archive Network; 2019.

Rick J, Rowe K, Hann M, Sibbald B, Reeves D, Roland M, & Bower P (2012). Psychometric properties of the Patient Assessment Of Chronic Illness Care measure: Acceptability, reliability and validity in United Kingdom patients with long-term conditions. BMC Health Services Research, 12, 293. 10.1186/1472-6963-12-293 [PubMed: 22938193]

Roberts NJ, Kidd L, Dougall N, Patel IS, McNarry S, & Nixon C (2016). Measuring patient activation: The utility of the Patient Activation Measure within a UK context— Results from four exemplar studies and potential future applications. Patient Education and Counseling, 99, 1739–1746. 10.1016/j.pec.2016.05.006 [PubMed: 27217050]

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement

Schalet BD, Reise SP, Zulman DM, Lewis ET, & Kimerling R (2021). Psychometric evaluation of a patient-reported item bank for healthcare engagement. Quality of Life Research, 1–12.

Simmons LA, Wolever RQ, Bechard EM, & Snyderman R (2014). Patient engagement as a risk factor in personalized health care: A systematic review of the literature on chronic disease. Genome Med, 6, 16. 10.1186/gm533 [PubMed: 24571651]

Stepleman L, Rutter MC, Hibbard J, Johns L, Wright D, & Hughes M (2010). Validation of the patient activation measure in a multiple sclerosis clinic sample and implications for care. Disability and Rehabilitation, 32, 1558–1567. 10.3109/09638280903567885 [PubMed: 20590506]

Thissen D, & Steinberg L (1986). A taxonomy of item response models. Psychometrika, 49, 501–519.

Thissen D, Steinberg L, & Fitzpatrick AR (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 26, 161–176.

Thissen D, Cai L, Bock RD, Nering ML, & Ostini R (2010) Handbook of Polytomous Item Response Theory Models: Developments and Applications 43–75.

Turner-Bowker DM, Bayliss MS, Ware JE, & Kosinski M (2003). Usefulness of the SF-8™ Health Survey for comparing the impact of migraine and other conditions. Quality of Life Research, 12, 1003–1012. [PubMed: 14651418]

Ware JEKM, Kosinski M, Dewey JS, & Gandek B (2001). How to score and interpret single-item health status measures: a manual for users of the SF-8 health survey

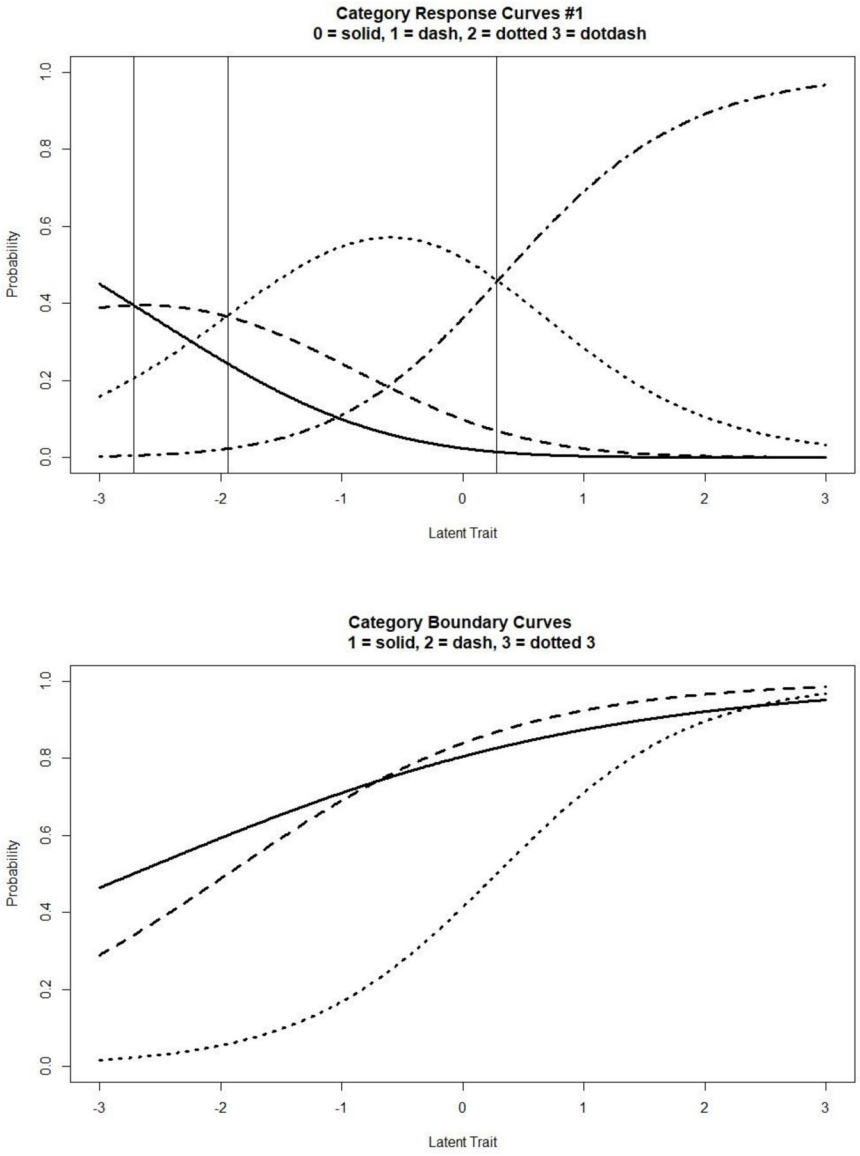
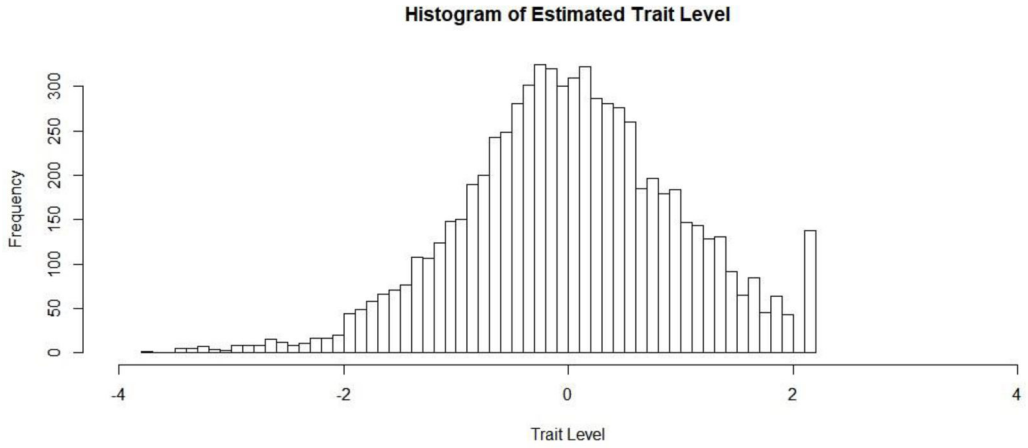



**Figure 1.**
Category Response Curves for Item #1 under the Nominal Response Model.



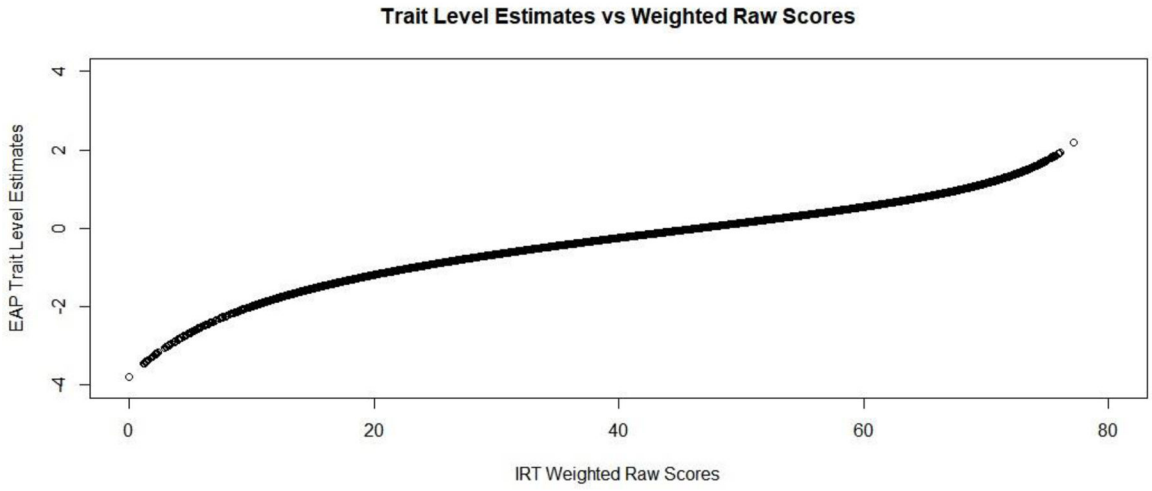




**Figure 2.**

Distribution of EAP Estimated Trait Levels and Relation Between Weighted Composite Scores and EAP Trait Level Estimates.

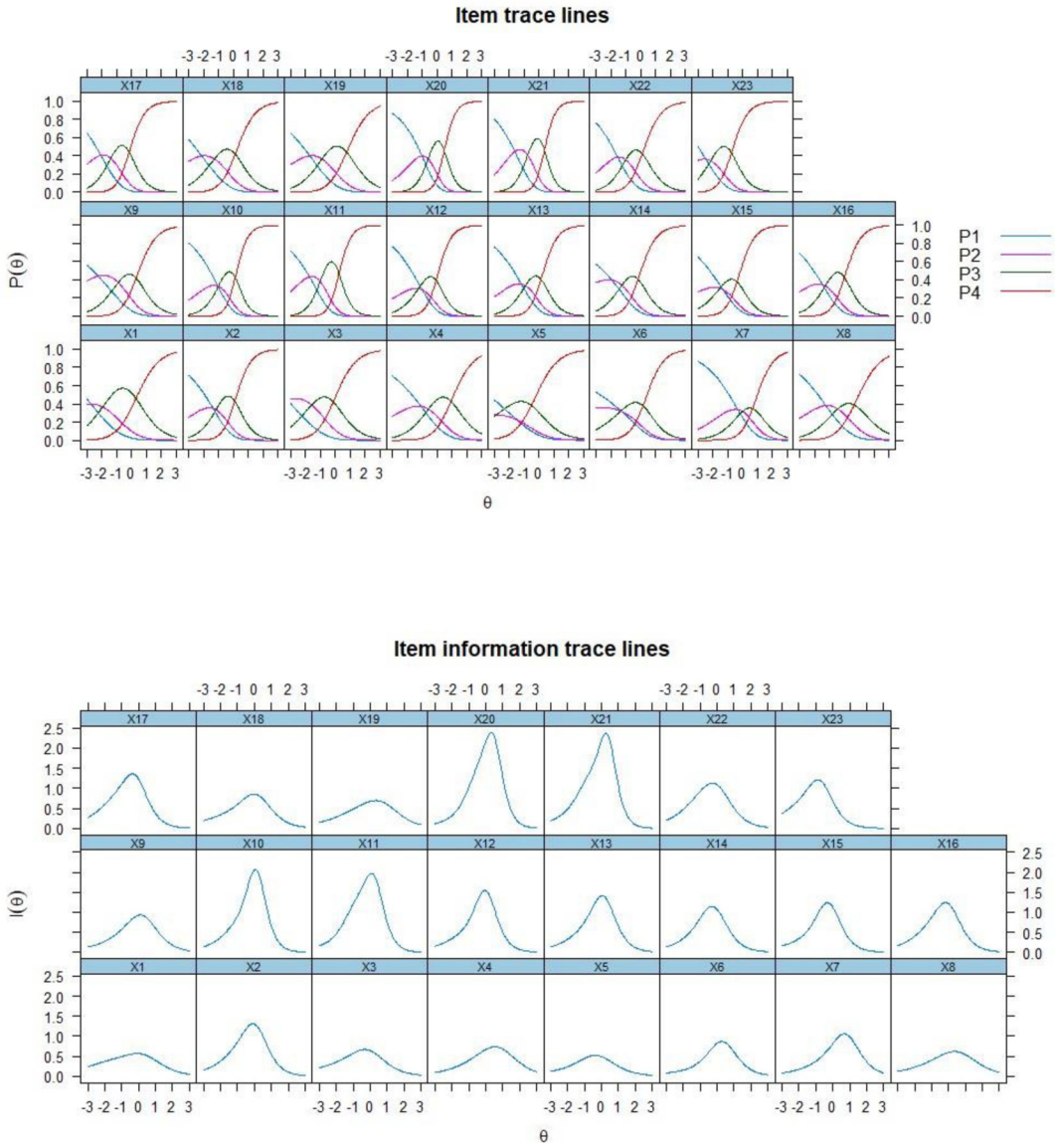



**Figure 3.**
Category Response Curves and Item Information Curves Under Nominal Response Model.

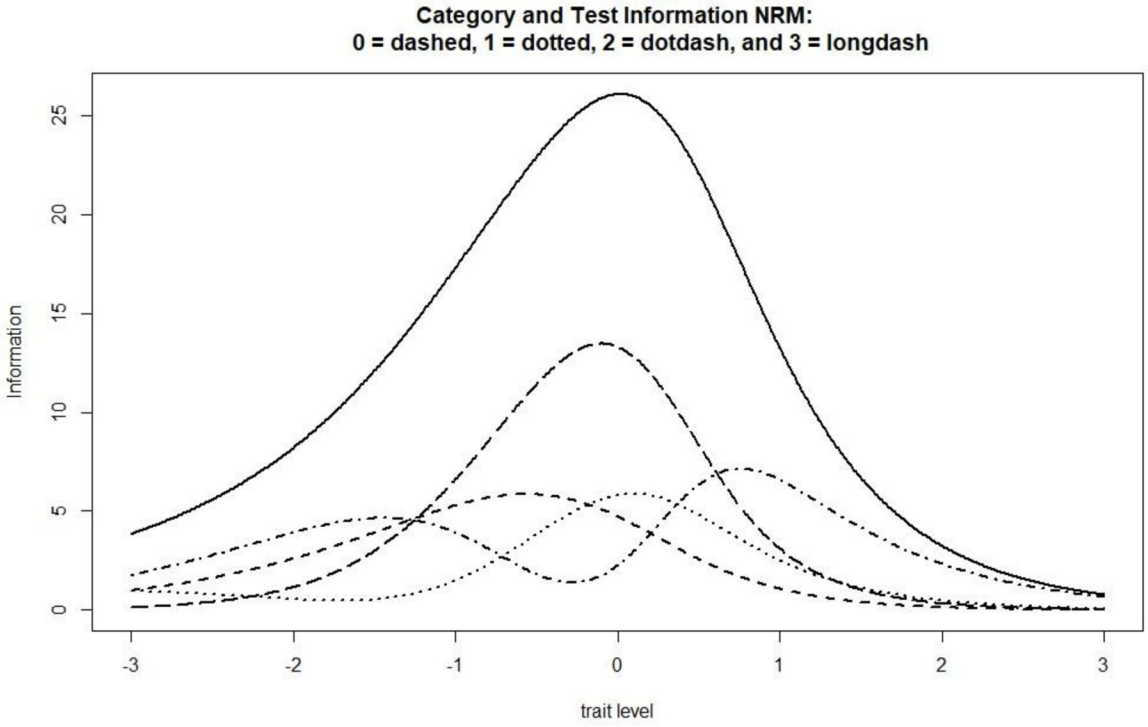
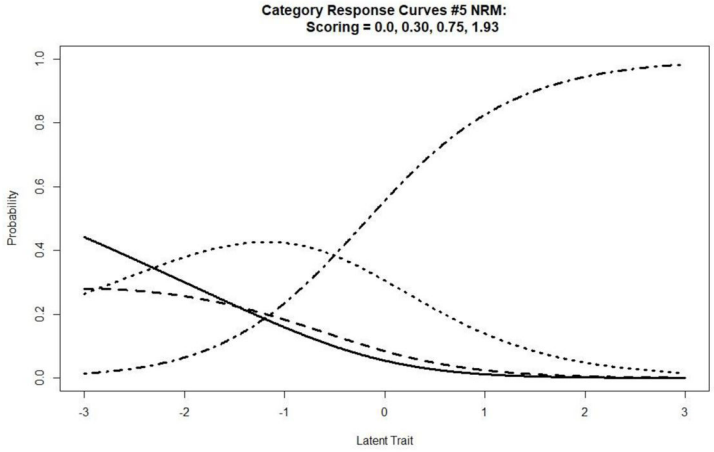



**Figure 4.**
Category and Test Information Under the Nominal Response Model.

#5 *It is easy for me to fill my medications on time*
Category 0 = solid, Category 1 = dashed, Category 2 = dotted, Category 3 = dotdash




#21 *I know I can get information I need about the pros and cons of treatments*.
Category 0 = solid, Category 1 = dashed, Category 2 = dotted, Category 3 = dotdash

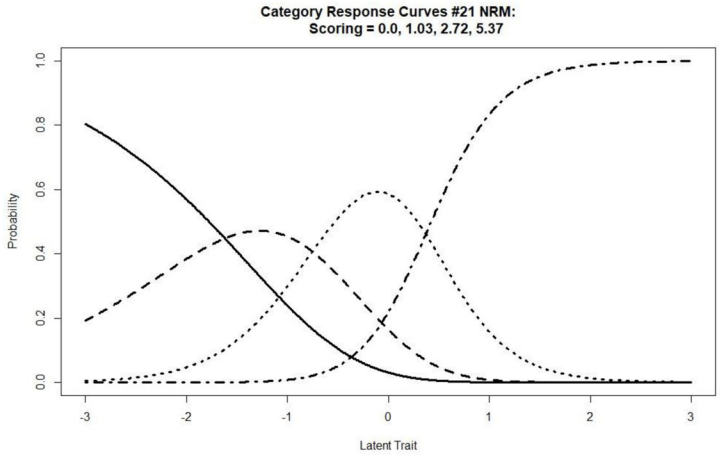
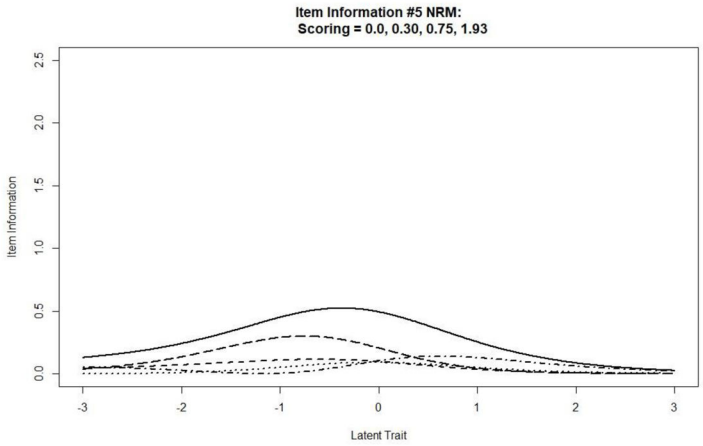



**Figure 5.**
Category Response Curves for Items #5 and #21.

#5 *It is easy for me to fill my medications on time.*
Category 0 = dashed, Category 1 = dotted, Category 2 = dotdash, Category 3 = longdash
Item Information = solid




#21 *I know I can get information I need about the pros and cons of treatments*.
Category 0 = dashed, Category 1 = dotted, Category 2 = dotdash, Category 3 = longdash
Item Information = solid

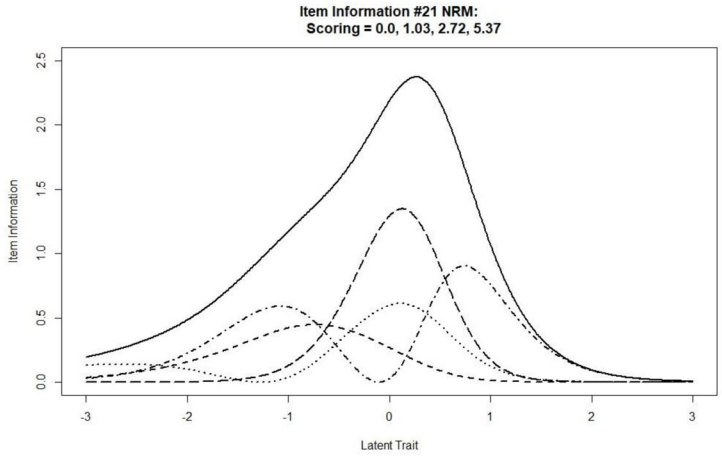



**Figure 6.**
Item Information Curves for Items #5 and #21.

**Table 1.**

Item Content for the 23-item Healthcare Engagement Measure (HEM).

| # | Item Content |
|---|---|
| 1 | I know I can always follow my doctor's instructions. |
| 2 | I always know who to contact when I have a health issue. |
| 3 | Learning more about my health issues helps me manage them better. |
| 4 | Even if I am tired or in pain, I know I can stick to my treatment plan. |
| 5 | It is easy for me to refill medications on time. |
| 6 | I know I can get the health care services I need, even if I must arrange it myself. |
| 7 | It is easy to find the health care resources I need (such as classes, support groups). |
| 8 | I have clear goals to improve my health. |
| 9 | Monitoring how well my treatments are working helps me get the most out of my care. |
| 10 | I know I can get a provider to deal with my main health concerns. |
| 11 | I can make sure my concerns are fully addressed before I leave appointments. |
| 12 | I know I can find a way to get in touch with my provider or care team when I need to. |
| 13 | When I need information about my care, like test results, I can get it easily. |
| 14 | If I think my treatment plan needs to change, I have no problem bringing it up with my provider. |
| 15 | I have a provider who I can trust to act in my best interests. |
| 16 | When I need more information, I ask, even when my provider is in a rush. |
| 17 | I make sure I understand all of my test results. |
| 18 | I know I can think through the pros and cons when I need to make a choice about my health. |
| 19 | I know I can get myself to keep doing the things that keep me healthy, even when life gets challenging. |
| 20 | I can get the care I need without getting discouraged. |
| 21 | I know I can get the information I need about the pros and cons of treatments. |
| 22 | I know I can express my doubts, even when my provider might disagree. |
| 23 | If I didnt' think a treatment was working, I would tell my provider. |

*Note*: Original response category labels are: 0 = *Not at all true*, 1 = *A little bit true*, 2 = *Somewhat true*, 3 = *Mostly true*, and 4 = *Very true*.

**Table 2.**

Nominal Response Model When Parameterized Using Five Response Categories.

| | $a^*$ | $ak_0$ | $ak_1$ | $ak_2$ | $ak_3$ | $ak_4$ | $c_0$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.57 | 0 | −0.71 | 0.34 | 1.83 | 4 | 0 | 1.19 | 2.88 | 4.54 | 4.18 |
| 2 | 0.90 | 0 | 0.12 | 0.80 | 2.04 | 4 | 0 | 0.88 | 2.12 | 3.44 | 3.29 |
| 3 | 0.62 | 0 | −0.25 | 0.67 | 1.96 | 4 | 0 | 1.45 | 3.38 | 4.66 | 4.84 |
| 4 | 0.69 | 0 | 0.15 | 0.85 | 2.18 | 4 | 0 | 0.88 | 1.78 | 2.31 | 1.14 |
| 5 | 0.48 | 0 | −0.10 | 0.57 | 1.52 | 4 | 0 | 0.46 | 1.39 | 2.68 | 3.28 |
| 6 | 0.53 | 0 | −1.13 | −0.16 | 1.13 | 4 | 0 | −0.39 | 0.81 | 1.70 | 1.47 |
| 7 | 0.78 | 0 | 0.31 | 1.13 | 2.13 | 4 | 0 | 0.48 | 1.20 | 1.20 | 0.22 |
| 8 | 0.70 | 0 | 0.68 | 1.38 | 2.45 | 4 | 0 | 1.28 | 2.30 | 2.63 | 1.91 |
| 9 | 0.75 | 0 | 0.11 | 0.75 | 2.08 | 4 | 0 | 1.34 | 2.69 | 3.58 | 3.15 |
| 10 | 1.08 | 0 | 0.01 | 0.74 | 1.84 | 4 | 0 | 0.59 | 1.86 | 3.08 | 2.81 |
| 11 | 1.22 | 0 | 0.04 | 0.68 | 2.05 | 4 | 0 | 0.86 | 2.57 | 4.27 | 3.66 |
| 12 | 0.92 | 0 | 0.18 | 0.89 | 1.84 | 4 | 0 | 0.72 | 1.82 | 2.92 | 3.05 |
| 13 | 0.94 | 0 | 0.33 | 0.96 | 2.08 | 4 | 0 | 0.80 | 2.00 | 2.95 | 2.70 |
| 14 | 0.72 | 0 | −0.49 | 0.29 | 1.79 | 4 | 0 | 0.07 | 1.53 | 2.88 | 3.17 |
| 15 | 0.78 | 0 | 0.09 | 0.74 | 1.78 | 4 | 0 | 0.36 | 1.59 | 2.81 | 3.45 |
| 16 | 0.89 | 0 | 0.23 | 0.85 | 2.07 | 4 | 0 | 0.73 | 2.04 | 3.43 | 3.42 |
| 17 | 0.99 | 0 | 0.11 | 0.92 | 2.13 | 4 | 0 | 1.05 | 3.05 | 4.74 | 5.04 |
| 18 | 0.69 | 0 | −0.25 | 0.70 | 1.91 | 4 | 0 | 0.89 | 2.52 | 3.61 | 3.45 |
| 19 | 0.72 | 0 | 0.21 | 1.00 | 2.25 | 4 | 0 | 1.32 | 2.58 | 3.4 | 2.43 |
| 20 | 1.36 | 0 | 0.25 | 0.88 | 2.13 | 4 | 0 | 1.02 | 2.31 | 3.33 | 2.10 |
| 21 | 1.36 | 0 | 0.05 | 0.81 | 2.06 | 4 | 0 | 1.29 | 3.23 | 4.50 | 3.52 |
| 22 | 0.89 | 0 | 0.11 | 1.07 | 2.30 | 4 | 0 | 0.85 | 2.57 | 3.68 | 3.52 |
| 23 | 0.89 | 0 | 0.08 | 0.86 | 2.02 | 4 | 0 | 1.31 | 3.26 | 5.45 | 6.77 |
| Mean | 0.85 | 0 | 0.01 | 0.77 | 1.98 | 4 | | | | | |

Note:

$a^*$ is the item slope, $ak_0 \ldots ak_4$ are scoring coefficients, $c_0 \ldots c_4$ category intercepts; response categories 0 = *Not at all*, 1 = *A little bit true*, 2 = *Somewhat true*, 3 = *Mostly true*, 4 = *Very true*.

**Table 3.**

Descriptive Statistics for the Healthcare Engagement Measure (HEM) Scored Using Four Response Categories.

| | | | | Factor | | Response Proportions for Category | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Item #** | **r.drop** | **mean** | **sd** | **Loading** | $h^2$ | ***0*** | ***1*** | ***2*** | ***3*** |
| 1 | 0.54 | 2.2 | 0.83 | .62 | .38 | .05 | .12 | .44 | .39 |
| 2 | 0.66 | 2.1 | 1.00 | .74 | .55 | .11 | .14 | .32 | .42 |
| 3 | 0.56 | 2.2 | 0.86 | .64 | .41 | .04 | .14 | .34 | .48 |
| 4 | 0.57 | 1.6 | 1.02 | .63 | .40 | .19 | .24 | .36 | .21 |
| 5 | 0.49 | 2.3 | 0.94 | .59 | .35 | .08 | .10 | .28 | .54 |
| 6 | 0.56 | 1.9 | 1.06 | .64 | .41 | .15 | .17 | .32 | .37 |
| 7 | 0.59 | 1.4 | 1.11 | .67 | .45 | .28 | .25 | .25 | .22 |
| 8 | 0.54 | 1.7 | 1.04 | .61 | .37 | .17 | .25 | .32 | .25 |
| 9 | 0.60 | 2.0 | 0.98 | .68 | .46 | .10 | .20 | .35 | .35 |
| 10 | 0.70 | 2.0 | 1.05 | .79 | .62 | .13 | .15 | .30 | .42 |
| 11 | 0.71 | 2.1 | 0.95 | .80 | .64 | .09 | .15 | .37 | .39 |
| 12 | 0.65 | 2.1 | 1.05 | .74 | .55 | .13 | .14 | .28 | .46 |
| 13 | 0.66 | 2.0 | 1.05 | .74 | .55 | .14 | .17 | .30 | .39 |
| 14 | 0.61 | 2.2 | 0.97 | .71 | .50 | .09 | .13 | .29 | .49 |
| 15 | 0.62 | 2.2 | 1.00 | .72 | .52 | .10 | .12 | .25 | .54 |
| 16 | 0.63 | 2.1 | 0.98 | .73 | .53 | .10 | .14 | .32 | .45 |
| 17 | 0.65 | 2.3 | 0.89 | .75 | .56 | .06 | .12 | .31 | .51 |
| 18 | 0.59 | 2.1 | 0.94 | .67 | .45 | .08 | .17 | .35 | .41 |
| 19 | 0.57 | 1.8 | 0.96 | .64 | .41 | .12 | .22 | .40 | .26 |
| 20 | 0.74 | 1.8 | 1.05 | .81 | .66 | .16 | .19 | .33 | .31 |
| 21 | 0.74 | 1.9 | 0.97 | .81 | .66 | .10 | .19 | .36 | .35 |
| 22 | 0.64 | 2.0 | 0.98 | .72 | .52 | .10 | .17 | .32 | .41 |
| 23 | 0.59 | 2.6 | 0.75 | .73 | .53 | .03 | .06 | .23 | .68 |

Note: r.drop is the item-test correlation with the item dropped; $h^2$ is the item communality.

**Table 4.**

Partial Credit Model Scored Using Four Response Categories.

| | $a^*$ | $ak_0$ | $ak_1$ | $ak_2$ | $ak_3$ | $c_0$ | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.19 | 0 | 1 | 2 | 3 | 0 | 2.17 | 4.04 | 3.72 |
| 2 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.37 | 2.70 | 2.72 |
| 3 | 1.19 | 0 | 1 | 2 | 3 | 0 | 2.47 | 3.98 | 4.20 |
| 4 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.07 | 1.67 | 0.49 |
| 5 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.43 | 3.16 | 3.75 |
| 6 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.12 | 2.17 | 1.98 |
| 7 | 1.19 | 0 | 1 | 2 | 3 | 0 | 0.56 | 0.57 | -0.33 |
| 8 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.26 | 1.71 | 0.88 |
| 9 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.75 | 2.73 | 2.41 |
| 10 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.15 | 2.31 | 2.36 |
| 11 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.64 | 3.03 | 2.82 |
| 12 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.15 | 2.35 | 2.64 |
| 13 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.25 | 2.24 | 2.19 |
| 14 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.57 | 2.94 | 3.31 |
| 15 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.36 | 2.74 | 3.42 |
| 16 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.44 | 2.85 | 2.98 |
| 17 | 1.19 | 0 | 1 | 2 | 3 | 0 | 2.00 | 3.62 | 4.04 |
| 18 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.89 | 3.15 | 3.07 |
| 19 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.56 | 2.49 | 1.55 |
| 20 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.10 | 1.95 | 1.40 |
| 21 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.70 | 2.74 | 2.35 |
| 22 | 1.19 | 0 | 1 | 2 | 3 | 0 | 1.62 | 2.79 | 2.77 |
| 23 | 1.19 | 0 | 1 | 2 | 3 | 0 | 2.22 | 4.42 | 5.66 |
| Mean | 1.19 | 0 | 1 | 2 | 3 | | | | |

Note:

[a]$a^*$ is the item slope, $ak_0 \ldots ak_3$ are scoring coefficients, $c_0 \ldots c_3$ category intercepts; response categories 0 = *Not at all/A little bit true*, 1 = *Somewhat true*, 2 = *Mostly true*, 3 = *Very true*.

**Table 5.**

Generalized Partial Credit Model Scored Using Four Response Categories.

| | $a^*$ | $ak_0$ | $ak_1$ | $ak_2$ | $ak_3$ | $c_0$ | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.05 | 0 | 1 | 2 | 3 | 0 | 1.94 | 3.69 | 3.40 |
| 2 | 1.36 | 0 | 1 | 2 | 3 | 0 | 1.55 | 2.97 | 2.95 |
| 3 | 1.05 | 0 | 1 | 2 | 3 | 0 | 2.24 | 3.64 | 3.87 |
| 4 | 0.89 | 0 | 1 | 2 | 3 | 0 | 0.79 | 1.31 | 0.35 |
| 5 | 0.80 | 0 | 1 | 2 | 3 | 0 | 0.90 | 2.33 | 2.92 |
| 6 | 0.86 | 0 | 1 | 2 | 3 | 0 | 0.76 | 1.64 | 1.57 |
| 7 | 0.94 | 0 | 1 | 2 | 3 | 0 | 0.37 | 0.37 | -0.33 |
| 8 | 0.82 | 0 | 1 | 2 | 3 | 0 | 0.90 | 1.25 | 0.66 |
| 9 | 1.08 | 0 | 1 | 2 | 3 | 0 | 1.61 | 2.52 | 2.23 |
| 10 | 1.67 | 0 | 1 | 2 | 3 | 0 | 1.66 | 3.08 | 3.00 |
| 11 | 1.88 | 0 | 1 | 2 | 3 | 0 | 2.47 | 4.29 | 3.90 |
| 12 | 1.36 | 0 | 1 | 2 | 3 | 0 | 1.31 | 2.61 | 2.86 |
| 13 | 1.33 | 0 | 1 | 2 | 3 | 0 | 1.38 | 2.44 | 2.34 |
| 14 | 1.18 | 0 | 1 | 2 | 3 | 0 | 1.52 | 2.88 | 3.25 |
| 15 | 1.28 | 0 | 1 | 2 | 3 | 0 | 1.45 | 2.89 | 3.56 |
| 16 | 1.29 | 0 | 1 | 2 | 3 | 0 | 1.54 | 3.01 | 3.11 |
| 17 | 1.52 | 0 | 1 | 2 | 3 | 0 | 2.44 | 4.33 | 4.72 |
| 18 | 1.09 | 0 | 1 | 2 | 3 | 0 | 1.74 | 2.92 | 2.86 |
| 19 | 0.97 | 0 | 1 | 2 | 3 | 0 | 1.31 | 2.15 | 1.33 |
| 20 | 1.89 | 0 | 1 | 2 | 3 | 0 | 1.78 | 2.91 | 2.01 |
| 21 | 1.99 | 0 | 1 | 2 | 3 | 0 | 2.62 | 4.08 | 3.39 |
| 22 | 1.27 | 0 | 1 | 2 | 3 | 0 | 1.69 | 2.90 | 2.85 |
| 23 | 1.53 | 0 | 1 | 2 | 3 | 0 | 2.76 | 5.31 | 6.62 |
| Mean | 1.26 | 0 | 1 | 2 | 3 | | | | |

Note:

[a]$a^*$ is the item slope, $ak_0 \ldots ak_3$ are scoring coefficients, $c_0 \ldots c_3$ category intercepts; response categories 0 = *Not at all/A little bit true*, 1 = *Somewhat true*, 2 = *Mostly true*, 3 = *Very true*.

**Table 6.**

Nominal Response Model Scored Using Four Response Categories (top) and Scoring Weights and Category Boundary Discriminations (bottom).

| | $a^*$ | $ak_0$ | $ak_1$ | $ak_2$ | $ak_3$ | $c_0$ | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.87 | 0 | 0.60 | 1.57 | 3 | 0 | 1.42 | 3.08 | 2.72 |
| 2 | 1.18 | 0 | 0.54 | 1.49 | 3 | 0 | 0.87 | 2.19 | 2.04 |
| 3 | 0.87 | 0 | 0.61 | 1.53 | 3 | 0 | 1.71 | 2.99 | 3.17 |
| 4 | 0.89 | 0 | 0.57 | 1.59 | 3 | 0 | 0.55 | 1.08 | -0.09 |
| 5 | 0.64 | 0 | 0.46 | 1.16 | 3 | 0 | 0.44 | 1.73 | 2.32 |
| 6 | 0.8 | 0 | 0.28 | 1.11 | 3 | 0 | 0.29 | 1.18 | 0.94 |
| 7 | 0.99 | 0 | 0.74 | 1.53 | 3 | 0 | 0.24 | 0.24 | -0.75 |
| 8 | 0.82 | 0 | 0.76 | 1.67 | 3 | 0 | 0.77 | 1.10 | 0.38 |
| 9 | 0.98 | 0 | 0.50 | 1.51 | 3 | 0 | 1.11 | 2.00 | 1.57 |
| 10 | 1.44 | 0 | 0.54 | 1.37 | 3 | 0 | 0.81 | 2.03 | 1.76 |
| 11 | 1.61 | 0 | 0.48 | 1.51 | 3 | 0 | 1.33 | 3.02 | 2.42 |
| 12 | 1.18 | 0 | 0.58 | 1.32 | 3 | 0 | 0.69 | 1.79 | 1.92 |
| 13 | 1.19 | 0 | 0.59 | 1.47 | 3 | 0 | 0.82 | 1.77 | 1.52 |
| 14 | 1.03 | 0 | 0.40 | 1.45 | 3 | 0 | 0.80 | 2.16 | 2.44 |
| 15 | 1.02 | 0 | 0.51 | 1.30 | 3 | 0 | 0.69 | 1.9 | 2.54 |
| 16 | 1.14 | 0 | 0.54 | 1.49 | 3 | 0 | 0.92 | 2.30 | 2.29 |
| 17 | 1.29 | 0 | 0.63 | 1.56 | 3 | 0 | 1.68 | 3.37 | 3.67 |
| 18 | 0.97 | 0 | 0.62 | 1.49 | 3 | 0 | 1.29 | 2.37 | 2.21 |
| 19 | 0.93 | 0 | 0.64 | 1.62 | 3 | 0 | 1.03 | 1.84 | 0.87 |
| 20 | 1.74 | 0 | 0.55 | 1.53 | 3 | 0 | 0.97 | 1.99 | 0.75 |
| 21 | 1.79 | 0 | 0.58 | 1.52 | 3 | 0 | 1.67 | 2.94 | 1.96 |
| 22 | 1.17 | 0 | 0.75 | 1.69 | 3 | 0 | 1.35 | 2.46 | 2.30 |
| 23 | 1.16 | 0 | 0.59 | 1.47 | 3 | 0 | 1.67 | 3.87 | 5.19 |
| Mean | 1.12 | 0 | 0.57 | 1.48 | 3 | | | | |

| | Category Scoring Weights | | | | *CBDs* | | |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | $a_1^*$ | $a_2^*$ | $a_3^*$ |
| 1 | 0.00 | 0.52 | 1.38 | 2.62 | 0.52 | 0.85 | 1.25 |
| 2 | 0.00 | 0.64 | 1.75 | 3.53 | 0.64 | 1.11 | 1.78 |
| 3 | 0.00 | 0.53 | 1.34 | 2.62 | 0.53 | 0.80 | 1.28 |
| 4 | 0.00 | 0.51 | 1.42 | 2.68 | 0.51 | 0.92 | 1.26 |
| 5 | 0.00 | 0.30 | 0.75 | 1.93 | 0.30 | 0.45 | 1.19 |
| 6 | 0.00 | 0.23 | 0.89 | 2.41 | 0.23 | 0.67 | 1.51 |
| 7 | 0.00 | 0.74 | 1.52 | 2.98 | 0.74 | 0.78 | 1.46 |
| 8 | 0.00 | 0.62 | 1.37 | 2.46 | 0.62 | 0.74 | 1.09 |
| 9 | 0.00 | 0.49 | 1.49 | 2.95 | 0.49 | 1.00 | 1.46 |
| 10 | 0.00 | 0.77 | 1.96 | 4.31 | 0.77 | 1.19 | 2.35 |

| | $a^*$ | $ak_0$ | $ak_1$ | $ak_2$ | $ak_3$ | $c_0$ | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 11 | 0.00 | 0.77 | 2.43 | 4.82 | 0.77 | 1.66 | 2.39 | | |
| 12 | 0.00 | 0.69 | 1.56 | 3.55 | 0.69 | 0.87 | 1.99 | | |
| 13 | 0.00 | 0.70 | 1.75 | 3.56 | 0.70 | 1.05 | 1.82 | | |
| 14 | 0.00 | 0.41 | 1.49 | 3.08 | 0.41 | 1.07 | 1.59 | | |
| 15 | 0.00 | 0.53 | 1.33 | 3.07 | 0.53 | 0.80 | 1.74 | | |
| 16 | 0.00 | 0.62 | 1.71 | 3.43 | 0.62 | 1.08 | 1.72 | | |
| 17 | 0.00 | 0.81 | 2.01 | 3.87 | 0.81 | 1.20 | 1.86 | | |
| 18 | 0.00 | 0.60 | 1.44 | 2.90 | 0.60 | 0.84 | 1.46 | | |
| 19 | 0.00 | 0.60 | 1.50 | 2.78 | 0.60 | 0.91 | 1.28 | | |
| 20 | 0.00 | 0.95 | 2.65 | 5.21 | 0.95 | 1.70 | 2.56 | | |
| 21 | 0.00 | 1.03 | 2.72 | 5.37 | 1.03 | 1.69 | 2.65 | | |
| 22 | 0.00 | 0.87 | 1.98 | 3.51 | 0.87 | 1.11 | 1.53 | | |
| 23 | 0.00 | 0.68 | 1.71 | 3.48 | 0.68 | 1.03 | 1.78 | | |
| Mean | 0 | 0.64 | 1.66 | 3.35 | 0.64 | 1.02 | 1.70 | | |

Note:

$a^*$ is the item slope, $ak_0 \ldots ak_3$ are scoring coefficients, $c_0 \ldots c_3$ are category intercepts; response categories 0 = *Not at all/A little bit true*, 1 = *Somewhat true*, 2 = *Mostly true*, 3 = *Very true*; $a_1^* \ldots a_3^*$ are *CBD*s.