Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Latent Variable Modeling for Unipolar Constructs in Health Sciences : Partially Ordered Scaling Procedure for Including Reference Population

Edward H. Ip,*† Shyh-Huei Chen †

†Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Winston-Salem, 27127, North Carolina, USA

*Corresponding author. Email: eip@wakehealth.edu

Abstract

Patient-reported outcome (PRO) data have gained increasing prominence in both medical research and FDA-related regulatory spaces. Inherited from the traditional measurement paradigm, PROs are structured to measure bipolar traits—i.e., traits that have meaning at both ends of the scale. However, in a medical context, certain constructs, such as depression and alcoholism, manifest as unipolar traits, meaning that the trait is only meaningful at one end of the distribution but not the other. For example, a low score signifies the absence of a quality (e.g., not alcoholic) rather than a relatively lower degree of that quality (i.e., less alcoholic) when compared to others. Methods such as zero inflation may not be well-suited for modeling unipolar traits because nonzero low scores, may simply indicate the absence of the condition. In this article, we explore methods for addressing unipolarity using a partially ordered set (poset) item response theory (IRT) model. This model decomposes response categories into two components: (1) the lowest category (e.g., "Never"), which is considered qualitatively distinct from the other categories, versus all other categories and (2) the remaining categories (e.g., "Rarely" to "Always") as ordered categories. Poset calibration is performed using a unidimensional IRT approach, which assumes that a common underlying latent trait drives both components. A real dataset (n=653) of breast cancer survivors containing depression data was used to illustrate the exploratory analysis. Scenarios both including and excluding "non-depressed" patients were examined using the poset IRT and the graded response model (GRM). This study also highlights the feasibility of the approach and explores potential extensions, such as incorporating a "Not Applicable" (NA) option and conducting multidimensional analyses.

Keywords: patient reported outcome, partially ordered set IRT, depression

1. Introduction

A patient-reported outcome (PRO) is a measure of a health-related construct that comes directly from a patient. When a PRO is operationalized for data analysis, certain assumptions must be made about the measure. Many PROs have been developed for specific populations of patients, and some constructs may not be directly relevant to a "healthy" subpopulation. An obvious example is perceived stigma, or perceived undesired “differentness” that deprives individual of full acceptance

(Goffman, 1963). This kind of measure is likely to be relevant to patients that have some form of noticeable chronic condition such as epilepsy or muscular atrophy (Lai et al., 2024). A more nuanced example is anxiety/depression, which is typically assessed in individuals with mood disorders. However, the measure would be more meaningful if the scale can be referenced to individuals without clinical conditions but having day-to-day “normal” fluctuations of mood. Other such examples of unipolar PROs include symptoms resulting from cancer treatment (Tran et al., 2021), alcoholism (Toner et al., 2019), drug abuse (Skinner, 1982), and pain (Revicki et al., 2009; Amtmann et al., 2010).

The relevance issue is closely related to the concept of unipolarity, where one end of a scale is not meaningful to some respondents. It is also related to the statistical challenge known as the inflated zero problem but is more complex. In unipolar constructs, a very low score (e.g., 0) reflects the absence of a quality rather than simply a relatively low score below the mean (Reise et al., 2021). This leads to measurement challenges, as the scale is not meaningful at the very low end, rendering some individuals "non-scalable." The distribution of scores tends to be quasi-continuous with a "floor effect," and measurement developers cannot construct items at the low end to adequately spread out the scale.

Apparently, there are practical implications for the construction of PRO and calibration for concepts that are deemed unipolar. For example, should the general/healthy population be excluded from the psychometric scaling process? Should all items in the PRO measure be constructed in such a way that they are only relevant to the diseased population? While one can argue that unipolarity is theoretically justifiable and real, in practice there may not be a clear and well-defined demarcation line between unipolar and non-unipolar constructs. For example, it may not be practical to only include truly depressed individuals into a sample for scaling a PRO designed to measure depression. First, depression scales are often used, for example, as a screening tool prior to a patient who may undergo a clinical assessment for depression. Second, only including clinically depressed patients would significantly narrow the applicability of the scale. There would not be a meaningful reference point as how such a narrow scale can be interpreted, as opposed to a scale that includes the general population in its calibration.

At the item level, the relevance issue persists. Some items may be applicable to the general population but others may not be meaningful when the respondent to the item is not affected by the disease. For example, in the Alcohol Dependence Scale (ADS; Skinner & Allen, 1982), the item “When you drink, do you stumble about, stagger, and weave?” may be relevant to the general population, but another item “After a period of abstinence (not drinking), do you end up drinking heavily again?” may not. We therefore will need a robust approach for scaling items that may or may not be relevant to the general population.

The purpose of this article is to explore psychometric analysis that could be used to handle the scaling process for constructs in which unipolarity is of concern. Specifically these analyses are based on a partial-order scaling method that treats ordered categories within an item as having two separate components. Consider the PROMIS emotional distress-anxiety measure (Pilkonis et al., 2014). An item such as “In the past 7 days, I felt fearful” contains 5 response categories from Never, Rarely to Always. One may want to treat “Never” as distinct from the other categories. The partial-order scaling method decomposes the response categories into “Never” vs other categories

(component 1); and then Rarely to Always as ordered categories (component 2). Calibration takes the form of a unidimensional IRT which posits that a common underlying latent trait drives both components 1 and 2. Commonly used IRT software packages can be used to estimate such models. Multidimensional extension of the poset model is also possible with the use of existing software. In this article, we use a real data example that contains a measure of depressive symptoms in a population of breast-cancer patients to illustrate the poset approach.

2. Background

2.1 Who to include for calibrating PROs?

Biomedical researchers often develop PRO scales for a specific disease population. A dilemma that they encounter is the extent to which they would make the scale relevant to the general population, even though the items have been designed and constructed to target patients that have the disease or are affected by the disease and related treatments. Consider a clinical researcher who is developing the following two PRO measures for an adolescent and young adult (AYA) cancer population - bodily image and financial burden. Apparently, these two constructs are important for the AYA population that is affected by the disease and the treatment because of the induced bodily change due to chemo- and radiation-therapy and the high financial burden due to treatment of cancer. A key decision point for the investigators is whether to include a "healthy" sample (hereafter referred to as the general sample) when scaling the item bank. There are advantages and disadvantages to include the general population into the calibration and norming procedures. The most important advantage of inclusion of the general population is comparability of an individual's score against the general population, and the interpretation of scores in terms of its deviation from a referenced population, which could be the general population or sometimes, a subpopulation. On the other hand, including the general population may create flooring effects, lack of interpretability of some items, potential differential item functioning (DIF) across the general and the diseased population, and the aforementioned concerns that unipolarity brings. In terms of the psychometric properties, including a general population is likely to result in high values of the discrimination parameter and very little separation between graded response threshold or location parameters at one end of the scale (Reise & Waller, 2009). Additionally, specific disease-condition PROs that are scaled using only a patient population are argued to have greater face validity, credibility and responsiveness to changes in the patient's condition (Churruca et al., 2021). To summarize this discussion, perhaps PRO constructs need to be determined on a case-by-case basis about whether to include the general population for scaling. Continuing to use the above example constructs of body image and financial burden in AYA cancer patients to exemplify this principle, body image perhaps is better aligned with the scaling approach that includes the general population, whereas financial burden (due to cancer management and treatment), may be more appropriate to only include patients with cancer.

The PROMIS item banks (Reeve et al., 2007), which cover a broad range of PROs, would offer a glimpse into how investigators handle the above dilemma in practice. Consider the PROMIS 56-item depression item bank (final "official" item bank only contains 28 items; see Nolte et al. 2019). The PROMIS Calibration Studies sample included 21,133 respondents, with n = 1,532 recruited from primary research sites associated with PROMIS network sites, while the vast majority (n = 19,601) was recruited from an Internet polling company. The research sites included a diverse range of

patient populations, with a focus on individuals with chronic conditions and specific conditions such as cancer. Thus the calibration was conducted using a "hybrid" sample or mixture of both the general and cancer patients. Note that in the PROMIS language, the calibration sample is separate from the "centering" sample, which refers to the norm centered on the US Census population. Other PROMIS measures used either the general population or a clinical sample for calibration. The choice reflects the very specific nature of each targeted construct. PROMIS documents the calibration sample (e.g., general population, clinical sample, or hybrid) and the centering sample for the vast majority of the measures developed under its auspices (PROMIS reference population, 2024).

The choice to use a general population, a clinical sample, or a hybrid sample reflects the extent to which a construct is viewed as unipolar from the investigators' perspective. When only a clinical sample is used for calibration and for centering, it reflects the perspective that the corresponding PROMIS score is not relevant to the general population such that there is no need to reference an individual's score with those that are deemed normal or healthy. Note that most of the PROMIS measures indeed used the general population for both calibration and centering. Thus, perhaps with the exception of a small number of obvious cases (an example is the measure of self-efficacy in managing medication and treatment; see PROMIS reference populations, 2024), the issue of unipolarity lingers.

2.2 Item-level Unipolarity Considerations

It can be argued that the manifestation of unipolarity occurs at the item level – i.e., some items in the item bank are only relevant to a specific population and irrelevant to a general population. In the PROMIS smoking measure (Edelen et al., 2012), the item "People think less of me if they see me smoking" may be relevant to a general population, but the item "Smoking allows me to take a break from my problems for a few minutes" would not be relevant to someone who is not currently smoking. Many PRO measures use Likert scale for which the categories indicate descending or ascending level of agreement, severity, or frequency. The lowest category (e.g., "Never" for a specific symptom) may suggest a qualitatively different response that cannot be ordered together with the other responses (e.g., Sometimes, Often, and Always). One may thus argue that we need to consider how to handle specific items when it is clear from a face-validity angle that some items are, and some are not relevant to the specific population. In other words, some items may be "more scalable" than others. Indeed, one could include "Not Applicable (NA)" as a response option. Accordingly, the response categories would form a partially ordered set (poset) that can be represented in Figure 1a.

3. Method

Using real depression data, we illustrate how poset IRT can be used to calibrate items for a scale that is suspected to be unipolar. In this illustration, we focus on exploratory analysis that demonstrates feasibility and interpretation of the proposed method. Items within a measure are all treated the same way, and we did not identify items that may not be scalable. As the empirical data used here did not contain the NA option, for illustration we treated the lowest category (0) as being qualitatively distinct from the other categories. We also used latent class analysis to identify patients that can be classified as "non-depressed" for further exploring the behavior of the poset approach.

(a) (b)

Figure 1(a). Partially ordered set (poset) structure of response categories for a hypothetical item that contains 5 categories (Never, Sometimes, Often, Always, and Not Applicable [NA]). An arrow A⍰B indicates category A dominates (is superior) to category B. Here the categories Never and NA are treated as qualitatively different from the ordered categories Sometimes, Often, and Always. Under a general poset framework, it is also possible that the Never and NA can be lumped together to form one category. (b) poset dominance structure for two binary items each with 0/1 response.

3.1 Data

The data set was a subset of data collected from a longitudinal study of age-related differences in adjusting to breast cancer diagnosis. Details on study recruitment and eligibility are reported elsewhere (Avis et al., 2013). Female patients were recruited within 8 months of breast cancer diagnosis from two sites in New York City and Texas. Eligibility criteria included a first-time diagnosis of stage I-III breast cancer, age >18 years, and ability to read and write English. Data were collected at 5 time points: the initial baseline survey (administered within 8 months of breast cancer diagnosis), and 3, 6, 12, and 18 months following completion of the baseline survey. The self-administered questionnaire included questions on depression, symptoms, sociodemographics, health-related quality of life, and psychosocial factors. In this study, we focus on baseline data from the Beck Depression Inventory (BDI; Beck, 1961), a commonly used scale for assessing depression. The BDI contains 21 items, each rated on a 4-point scale. A sample BDI question about being disappointed contains the 4 response options: 0= I don't feel disappointed in myself; 1=I am disappointed in myself; = 2=I am disgusted with myself; and 3=I hate myself.

Out of 740 surveys mailed to women deemed eligible, 653 women completed baseline surveys. The sample used in this study thus contained n=653 participants. The average age of the sample was 54.9 years. The majority (89.6%) was White and most (87.4%) attained education beyond high school, and 71.7% were married or partnered. Most of the sample (92%) had been diagnosed with stage I or II breast cancer. Two-thirds (66.9%) had received chemotherapy, 72.3% received radiation, and 63.7% had undergone surgery.

3.2 Poset IRT and GRM

Two exploratory analyses for the BDI data were conducted. First, we applied the poset IRT to calibrate the items and examined the item parameters. Second, we explored ways to distinguish between a "healthy" population and a "depressed" population, and then used poset IRT to calibrate the items by including only the "depressed" population. We also examined item parameters and latent distributions.

The poset IRT method was described in Ip et al. (2022), and the theoretical foundation was reported in Zhang & Ip (2012), Ip et al. (2013), and Ip, Chen, & Quandt (2016). Briefly, the poset IRT is designed to calibrate a new class of response data that are partially ordered. Unlike traditional response data, poset data allows some form of ordering that may not be complete and represents a generalization to ordinal and categorical response data. A classical example would be a testlet of two binary items (Figure 1b), where the categories 01 and 10 are deemed not comparable. One example of such a top-spin like structure in PRO is symptom that contains the following two attributes: severe (yes 1, or no 0) and bothersomeness (yes 1, or no 0). It can be proved that such a poset structure can be decomposed into a chain of comparable categories 11>{10,01}>00 and an antichain of incomparable categories 01 and 10. Within a latent variable framework, analyzing such poset responses can be translated into recoding the responses into subitems. In the example, the poset decomposition would result in two subitems that is shown in Table 1. The principle applies to more complex poset structures. Accordingly, the nominal IRT model (for subitem 1) and graded response model (for subitem 2) can be applied to the recoded data. Using the recoding scheme, it is possible to analyze poset data using commonly available IRT programs such as mirt (Chalmers, 2012) and IRTPRO (Cai, Thissen, & du Toit, 2017).

Table 1. Recoding of response patterns into subitems for poset IRT analysis.

Response pattern		Subitem 1*	Subitem 2*
Y_1	Y_2	Y_c	Y_o
1	1	NA	1
1	0	1	2
0	1	0	2
0	0	NA	3

*Subitem 1 indicates the categorical nature between 01 and 10, and subitem 2 indicates the ordered relationship for the response patterns 11, {10,01}, and 00. $Y_{1 \text{ and }} Y_2$ represent the original response, and $Y_{c \text{ and }} Y_c$ rpresent the recoded subitems.

4. Result

4.1 Exploratory Analysis 1

We analyzed the BDI data using both the graded response model (GRM) model and the poset IRT. We treated 0 as forming its own category and the others as ordinal categories. Consequently, there were two subitems – one binary (0 vs others), and the other ordinal (1,2, and 3). The two-parameter (2PL) logistic model IRT model was used for the binary outcome and GRM was used for the ordered outcome. Table 2 summarizes the slope parameters for both the GRM and the poset models.

Table 2. Discrimination parameters for BDI items using GRM and poset for the entire sample (left panel) and using only the restricted sample (right panel).

	Entire Sample			Restricted Sample		
BDI item	GRM	Poset 2PL	Poset GRM	GRM	Poset 2PL	Poset GRM
	a	a_{2PL}	a_{GRM}	a	a_{2PL}	a_{GRM}
sad	2.38	2.35	2.26	1.22	0.59	2.05
future	2.09	1.95	2.31	0.98	0.65	1.78
failure	1.98	1.78	1.41	0.69	0.60	1.62
satisfy	2.85	3.36	1.85	1.06	1.17	1.52
guilty	1.78	1.62	1.7	0.60	0.47	1.41
punish	1.5	1.37	1.25	0.51	0.33	0.96
hate	1.84	1.69	2.16	0.42	0.21	1.67
blame	1.19	1.07	1.69	0.10	-0.19	1.30
kill	1.19	1.2	0.93	0.81	0.89	0.91
cry	1.62	1.56	1.67	0.60	0.13	1.31
annoy	1.07	1.17	-0.23	0.53	0.39	0.13
interest	1.86	1.88	1.35	0.71	0.52	0.86
decide	2.14	2.12	1.43	1.13	0.77	0.83
look	1.56	1.57	0.92	0.57	0.09	0.76
effort	1.84	1.99	1.3	1.27	0.48	1.99
sleep	1.00	1.02	0.69	0.51	0.26	0.58
tired	1.42	1.41	1.35	0.82	-0.03	1.38
appetite	1.12	1.05	1.21	0.79	0.60	0.72
pounds	0.68	0.66	0.07	0.48	0.28	0.79
worry	1.48	1.50	1.10	0.34	0.02	0.55
sex	1.33	1.13	1.34	0.91	0.14	1.53

The result on the left panel in Table 2 shows the poset 2PL a_{2PL} parameter is often comparable to the GRM *a* parameter. On the other hand, in some cases the poset parameter a_{GRM} is diminished. One example is the item annoy, a=1.07, a_{2PL}=1.17, and a_{GRM} = -0.23, as well as the item pounds (0=I haven't lost much weight, if any, lately; 1-3 = lost 5,10, and 15 pounds respectively). The result suggests that for these items, most of the discriminating information is available at the 0 vs other dichotomy. The ordered component does not add information. On the other hand, there are items for which a_{GRM} is "boosted" and has value higher than *a* in GRM. The items hate and blame are examples. The separation between 0 vs others and the 3 non-zero categories appears to enhance the discriminating power of these items. Note that the GRM uses a common discrimination parameter across categories, so it is possible that the enhancement is due to the "spreading out" of the item characteristic curves within the poset GRM model.

4.2 Exploratory Analysis 2

In the second exploratory analysis, we assumed that there was a subpopulation in the cancer patient population where unipolarity may be driving the responses. While this breast cancer patient

study did not include a general population (i.e., participants without cancer), we assumed here that there were study participants that behaved similarly to the general "healthy" population. The purpose of this exploratory analysis is to analyze how excluding this subpopulation would affect the calibration and also behavior of individual items.

4.2.1 Latent Class Analysis for Identifying a Restricted Sample

We first use a latent class analysis (LCA) to explore the heterogeneity across the 21 items in the entire sample. Using the BIC as criterion to select the number of classes, the LCA resulted in a 6-class model. Figure 2 shows the conditional probabilities of the item responses across the 21 BDI items for each identified class. It can be seen that the class labeled HState4 is the most depressed class, whereas HState 1 is the most non-depressed class. After inspecting the classes, we interpreted the classes HState1, HState 3, and HState 5 as relatively "normal" or non-depressed subgroups (approximately 72% of the original sample) and excluded them in the subsequent analysis. We purposefully excluded participants more liberally so if there is any psychometric impact we would be more likely to discover that.

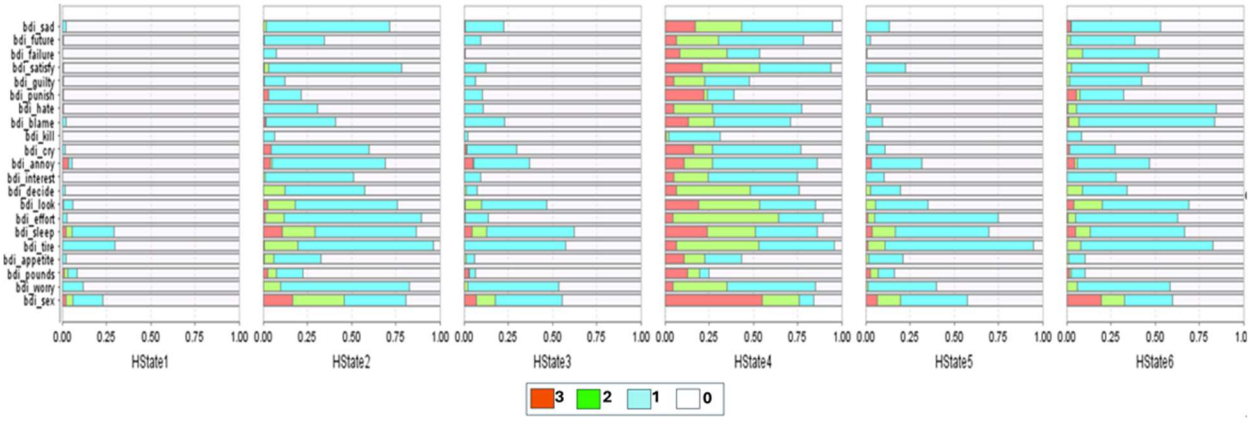

Figure 2. Conditional probabilities of six latent classes (HState 1 through 6) for BDI depression data. Each stacked bar for an item (vertical axis) represents the conditional probabilities of attaining a response value (0,1,2,3) for the item.

To provide more information about the depression profile of the excluded classes, Figure 3 shows the distributions of the BDI total score (range 0-63) across the 6 classes. The medians of the 3 excluded classes are all below 10. The two horizontal lines correspond to thresholds for "mild mood disturbance" and "borderline clinical depression". Both cutoffs at 10 and 17 were investigated for screening, and evidence for screening was stronger at cutoff=10 (Norris et al., 1987, Edelstein et al., 2010). The majority of the 3 excluded classes can be characterized as "the ups and downs are considered normal". While these labels are not necessarily universal – e.g., different clinical populations have different thresholds, we used the thresholds here as a support for our selection of classes to exclude. Hereafter we refer to the subsample (HStates 1,3,5 excluded) being analyzed as the restricted sample.

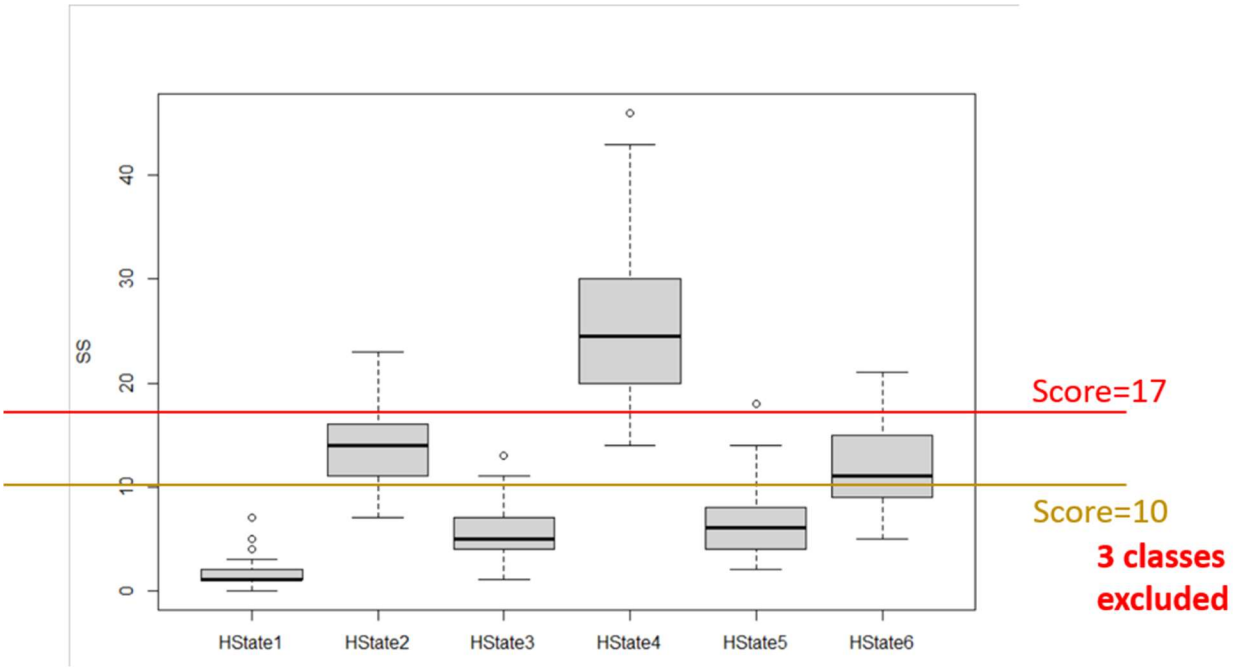

Figure 3. Boxplots of distributions of BDI score by latent class. The subsample for which the three classes (HState 1,3, and 5) was referred to as the restricted sample.

4.2.2 Poset IRT and GRM Analysis for the Restricted Sample

We compared two calibration procedures- the GRM and the poset IRT. The right panel in Table 2 shows the slope parameters for both calibrations for the restricted population. Not surprisingly, the *a* parameter in the GRM is much lower than the a parameter for the entire sample. As Reise et al. (2020) observed, "the more non-cases are over sampled, the more IRT slope parameters are inflated." Comparing a_{2PL} and a_{GRM} is revealing. The 2PL slope in poset IRT is much closer to 0. With the exception of the item Satisfy, all a_{2PL} values are less than 1.0.

Figures 4 and 5 respectively show the item characteristics curves (ICCs) of the poset and the GRM for the restricted sample. Consistent with the result from Table 2, Figure 4 shows that for the items worry, sex, look, and tired, the ICC in the binary component are flat, suggesting that there for these items those responded at the level of 0 (e.g., for the worry item 0="I am no more worried about my health than usual") have more or less the same overall score as those that responded with other options. Such symptoms appear to be more related to everyday ups and downs that are experienced by the general population and not specific to the more depressed population, and unlikely to be depressive symptoms that are directly disease-related. Surprisingly, the binary ICC of the item blame (0= "I don't feel I am any worse than anybody else") has a negative trend compared to other items. It may be that for the cancer population, the blame item is more reflective of self-esteem or self-confidence and not directly related to the overall depression severity, which is reflective in other symptoms that are caused by the disease or treatment.

We also conducted a multigroup (MG) calibration procedure (Bock & Zimowski, 1997) in which the entire sample was included for joint calibration using GRM. MG calibration differs from single-group calibration in that it allows the less depressed classes (HStates 1,3,5) and the more depressed

classes (HStates 2,4,6) to have different means and variances in their latent distributions. The GRM item parameters from the MG analysis (details not shown) were quite similar to that for the joint calibration without allowing such differences. Interesting, the mean (1.15) and variance (0.72) of the more depressed group is substantially different from those for the less depressed (reference) group, which are respectively set to 0 and 1. The result shows that mean latent depression score for the more depressed group, which was identified by the LCA, is more than 1 standard deviation from the reference group of the "general" population and has a lower dispersion.

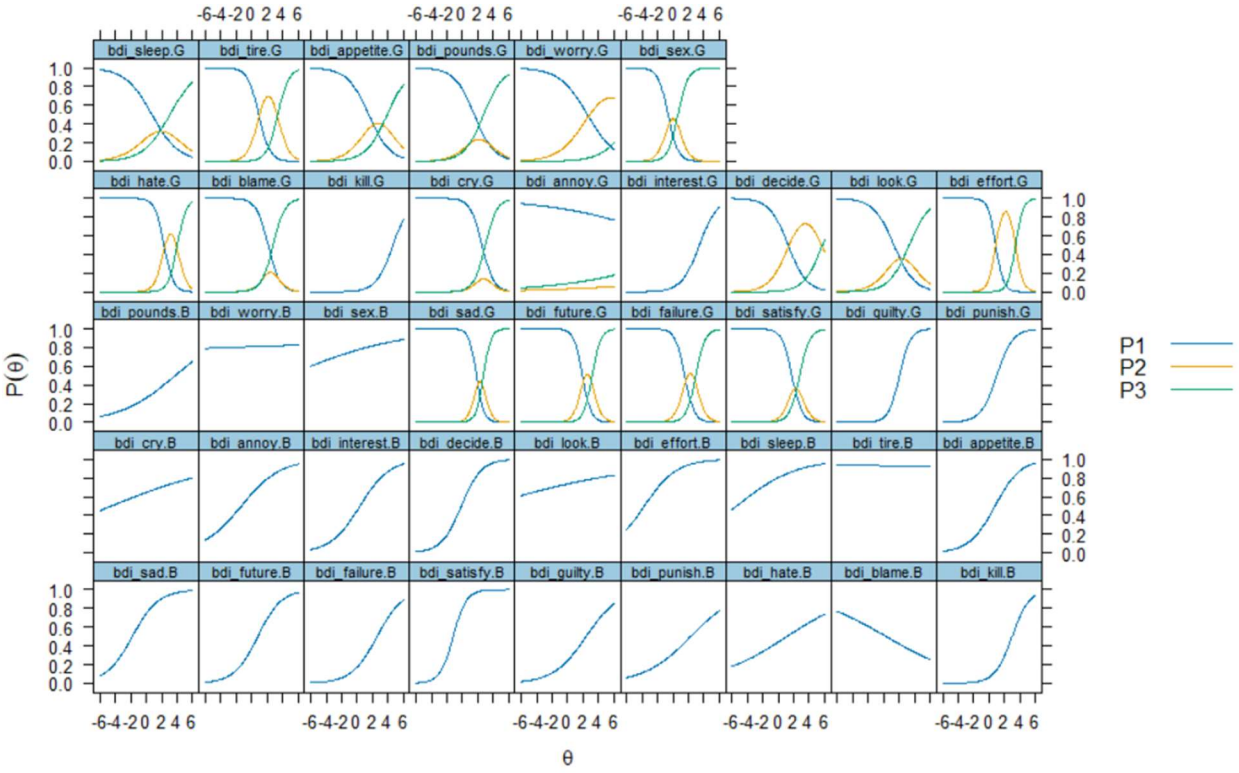

Figure 4. The poset item characteristic curves (ICCs) for a restricted sample of depressed participants. The labels with _G indicates graded response and _B indicates binary response. P1, P2, and P3 indicate the ordered categories 1,2, and 3.

5. Discussion

Using the poset IRT approach, this exploratory study demonstrates a method for analyzing scales and items suspected of exhibiting unipolarity while exploring different modeling options. The concept of unipolarity may be better understood as a continuum rather than a binary characteristic. However, in practice, developers of PROs must make a definitive decision regarding whether to include the general population in scaling.

PROMIS serves as a useful reference in this context, as most of its PRO measures (>300) use the general population as both a reference and a centering point. This decision is primarily driven by the need to ensure broad applicability of the scale and to facilitate comparisons between diseased individuals and the general population. However, including the general population often introduces

the unipolarity dilemma—enhancing the scale's general relevance at the cost of incorporating individuals for whom the construct is not as meaningful.

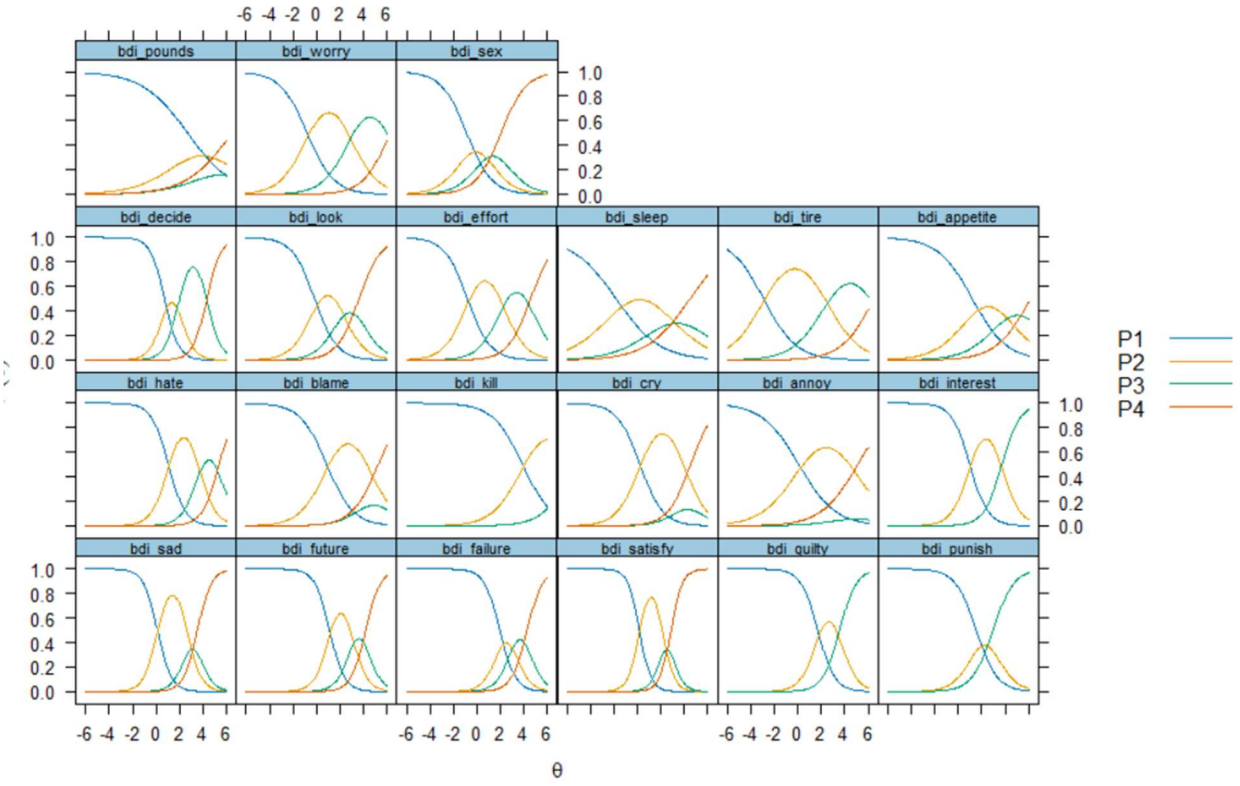

Figure 5. The GRM item characteristic curves (ICCs) for a restricted sample of depressed participants. P1, P2, P3, and P4 indicate the ordered categories 0,1,2, and 3.

This dilemma can also be understood at the item level. For example, in the context of anxiety and depression within a cancer population, some items are relevant to the general population (e.g., those reflecting normal fluctuations in mood), whereas others are more specific to the disease population due to the impact of illness or treatment (e.g., pain, fatigue, and changes in appearance).

The poset approach has several advantages for item banks for unipolar constructs. First, it conceptually distinguishes the categories. Without the NA option, the poset can separate the lowest response such as Never, which is typical for a general population, from the other options. Second, the poset approach provides a highly flexible framework for analyzing PRO measures that may exhibit unipolarity at the item level. Although this study applied poset IRT to all items in the BDI, the model can be selectively applied to a subset of items. For example, if a particular item is deemed relevant only to the disease population, the poset response model can be applied exclusively to that item. As discussed in the Background section, distinguishing between relevant and irrelevant items for the general population can be achieved through various strategies within the poset formulation, including the inclusion of an NA option, treating 0 as a distinct category, or folding the 0 option into NA. While a limitation of this study is its reliance on a unidimensional IRT framework, wherein both the dichotomous subitem and the ordered subitem are assumed to be driven by a single latent trait, the flexible poset framework can be extended to relax the assumption using a multidimensional poset IRT approach.

In addition to alternative IRT parameterizations, such as the log-logistic model proposed by Reise et al. (2020) for addressing unipolarity, recent work has explored modeling unipolar constructs as multidimensional processes. Magnus and Garnier-Villarreal (2022) described a multidimensional zero-inflated (MZI) GRM to handle symptom data that are related to the unipolar phenomenon. According to the model, two distinct but correlated latent variables underlie ordinal item responses; one represents susceptibility to the construct, whereas the other represents severity. Like the poset formulation, the MZI uses a 2PL IRT model for the component 0/1 (susceptibility), and the GRM for the ordinal component (severity). Wall, Park, & Moustaki, (2015) proposed a two-part mixture model for accommodating zero-inflated data. Strachan and Ip (2021) used a bivariate latent variable model for modeling longitudinal attitude data – one latent variable drives repeated response of the same attitudinal variable and the other drives consistency in same variable over time. This bivariate model can be adapted for a poset formulation for alternative modeling processes.

Acknowledgement

We acknowledge Dr. Nancy Avis for sharing the cancer survivor data.

Funding Statement This research was supported by grants from the National Science Foundation grant number SES 2120174 , and National Institute of Health grant number 5P30CA012197-50.

Competing Interests None.

Reference

Amtmann, D., Cook, K.F., Jensen, M.P., Chen, W., Choi, S., Revicki, D., Cella, D., Rothrock, N., Keefe, F., Callahan, L., & Lai, J. (2010). Development of a PROMIS item bank to measure pain interference, *Pain, 150*, 173-182.Avis, N.E., Levine, B., Naughton, M.J., Case, D.L., Naftalis, E., & Van Zee K.J. (2013). Age-related longitudinal changes in depressive symptoms following a breast cancer diagnosis and treatment. *Breast Cancer Research And Treatment, 139(1),* 199-206.

Beck, A.T., Ward, C.H., Mendelson, M.,Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archive of General Psychiatry, 4*, 561-71.

Bock, R.D., & Zimowski, M.F. (1997). Multiple Group IRT. In: van der Linden, W.J., Hambleton, R.K. (eds) *Handbook of Modern Item Response Theory*. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2691-6_25

Cai, L., Thissen, D., & du Toit, S.H.C. (2017). IRTPRO 4.2 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48(6),* 1–29.

Churruca, K., Pomare, C., Ellis, L.A., Long, J.C., Henderson, S.B., Murphy, L.E.D., Leahy, C.J., & Braithwaite, J. (2021). Patient-reported outcome measures (PROMs): A review of generic and condition-specific measures and a discussion of trends and issues. *Health Expectations, 24(4):*1015-1024. doi: 10.1111/hex.13254. Epub 2021 May 5. PMID: 33949755; PMCID: PMC8369118.

Edelen, M.O., Tucker, J.S., Shadel, W.G., Stucky, B.D., & Cai, L.. (2012). Toward a more systematic assessment of smoking: Development of a smoking module for PROMIS. *Addictive Behaviors, 37*,1278–1284. doi: 10.1097/01.mlr.0000245251.83359.8c.

Edelstein, B.A., Drozdick, L.W., & Ciliberti, C.M. (2010). Assessment of Depression and Bereavement in Older Adults. In P. A. Lichtenberg (Eds.) *Handbook of Assessment in Clinical Gerontology* (Second Edition) pp. 3-43, Academic Press.

Goffman E. (1963). *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs, NJ: Prentice Hall.

Ip, E.H., Chen, S., Bandeen-Roche, K., Speiser, J., Cai, L., & Houston, D. (2020). Longitudinal Partially Ordered Data Analysis for Preclinical Sarcopenia. *Statistics in Medicine, 39*, 3313-3328. PMCID: PMC8386024

Ip, E.H., Chen, S., Quandt, S. (2016). Analysis of multiple partially ordered responses to belief items with Don't Know option. *Psychometrika*. *81*, 483-505. PMC4458241

Ip, E.H., Zhang, Q., Rejeski, J., Harris, T., & Kritchevsky, S. (2013). Partially ordered mixed hidden Markov model for the disablement process of older adults. *Journal of the American Statistical Association, 108*, 370-384.

Lai, J.S., Nowinski, C., Rangel, S.M. et al. (2024). Development of the PROMIS pediatric stigma and extension to the PROMIS pediatric stigma: skin item banks. *Quality of Life Research, 33,* 865–873 (2024). https://doi.org/10.1007/s11136-023-03574-z

Magnus, B. E., & Garnier-Villarreal, M. (2022). A multidimensional zero-inflated graded response model for ordinal symptom data. *Psychological Methods, 27(2),* 261–279.

Nolte, S., Coon, C., Hudgens, S. et al. (2019). Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *Journal of Patient Reported Outcomes, 3*, 46. https://doi.org/10.1186/s41687-019-0127-0

Norris, J. T., Gallagher, D. E., Wilson, A., & Winograd, C. H. (1987). Assessment of depression in geriatric medical outpatients: The validity of two screening measures. *Journal of the American Geriatrics Society, 35(11),* 989–995. https://doi.org/10.1111/j.1532-5415.1987.tb04001.x

Pilkonis, P.A., Yu, L., Dodds, N.E., Johnston, K.L., Maihoefer, C.C., & Lawrence, S.M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study. *Journal of Psychiatry Research, 56*, 112–119 (2014).

PROMIS reference populations. (April 29, 2024)

https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/reference-populations

Reeve B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A. , Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Cella, D, & group obotPc. (2007). Psychometric evaluation and calibration of Health-Related Quality of Life item banks: Plans for the

Patient-Reported Outcome Measurement Information System (PROMIS®). *Medical Care, 45*, S22–S31.

Reise, S. P., Du, H., Wong, E. F., Hubbard, A. S., & Haviland, M. G. (2021). Matching IRT models to patient-reported outcomes constructs: The graded response and log-logistic models for scaling depression. *Psychometrika, 86(3),* 800–824.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?. *Psychological Methods, 8*, 164-84.

Revicki, D.A., Chen, W., Harnam, N., Cook, K.F., Amtmann, D., Callahan, L.F., Jensen, M.P., & Keefe, F.J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain,146 (1–2),*158-169.

Skinner, H. A., & Allen, B. A. (1982). Alcohol dependence syndrome: Measurement and validation. *Journal of Abnormal Psychology, 91(3),* 199-209.

Strachan, T., & Ip, E.H. (2021). Bivariate Model for Latent Variables Jointly Assessing Attitude and Attitudinal Stability. *Multivariate Behavioral Research, 56*, 724–738.

Tran, T.X.M., Park, J., Lee, J. et al. (2021). Utility of the Patient-Reported Outcomes Measurement Information System (PROMIS) to measure primary health outcomes in cancer patients: a systematic review. *Support Care Cancer, 29*, 1723–1739. https://doi.org/10.1007/s00520-020-05801-6

Toner, P., Böhnke, J.R., Andersen, P., & McCambridge J. (2019). Alcohol screening and assessment measures for young people: A systematic review and meta-analysis of validation studies. *Drug and Alcohol Dependence, 202*, 39-49.

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

Zhang, Q., & Ip, E.H. (2012). Generalized Linear Model for Partially Ordered Data. *Statistics in Medicine, 31*, 56-68.