2024-04-16

Distinguishing Between Symptom Presence and Severity Using a Two-Part Sequential Model

Luiza Ferreira Pradera
Brigham Young University

Distinguishing Between Symptom Presence and Severity Using a Two-Part Sequential Models

Luiza F. Pradera

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Scott A. Baldwin, Chair
Michael J. Larson
Scott R. Braithwaite

Department of Psychology

Brigham Young University

ABSTRACT

Distinguishing Between Symptom Presence and Severity Using a Two-Part Sequential Model

Luiza F. Pradera
Department of Psychology, Brigham Young University
Master of Science

Most symptom measures either implicitly or explicitly distinguish between symptom presence and symptom severity. For example, item 2 on the PHQ-9, a commonly used measure of depressive symptoms, asks respondents to rate how much they have been “feeling down, depressed, or hopeless.” The response options are 0 (Not at all), 1(Several Days), 2(More than half of the day), and 3(Nearly every day). Answering 0 indicates that the symptoms are not present, and any response greater than 0 suggests the symptom is present. Higher values indicate higher severity of the symptom. Although the response options distinguish between symptom presence and severity, most users of the PHQ-9 score it by assuming that a 0 (i.e., no symptom), lack of symptoms, is the low end of the severity spectrum. However, clinically, there is often a distinction between experiencing symptoms and how severe any of those symptoms is. Baldwin and Olsen (2023) developed a sequential item-response theory model that can be used to evaluate whether the symptom presence and symptom severity should be separated or considered part of the same construct. We applied the sequential mode to 3 datasets, a sample of 6242 participants, containing a variety of measures (e.g., State-Trait Anxiety Inventory, Penn State Worry Questionnaire). The results indicate that the Two-Part model has the best overall fit out of the three models (Two-part, Extreme Response, Unique Relationship), suggesting that symptom presence and severity should typically be considered distinct constructs. We discuss the implication for scoring and clinical use of symptom measures in light of our results.

Keywords: two-part mode, symptom, measures, gate response, item-response theory

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Distinguishing Between Symptom Presence and Severity Using a Two-Part Sequential Model

Psychometrics focuses on the performance, analysis, and design of tests, questionnaires, and other measurement instruments. In clinical psychology and other health disciplines, commonly used measures aim to assess and track patients' symptoms. These symptom measures can be used to aid diagnosis or can be used to provide an index of treatment response. In either case, symptom measures typically combine the idea of symptom presence–does the patient experience a particular symptom?–and symptom severity–how intense or frequent is the symptom. Although most symptom measures explicitly or implicitly make this distinction, the majority of psychometric models researchers use to evaluate the reliability and validity of these measures do not allow for the presence/severity distinction to be built into the model. In this thesis, I review the presence/severity distinction, present an item-response theory (IRT) model for symptom data, and apply the IRT model from five measures and N = 6242 people to evaluate whether an IRT model that distinguished between presence and severity better fits the data than a model that treats presence and severity as the same.

Two Examples of the Presence versus Severity Distinction

To illustrate the difference between presence and severity, we walk through two examples of commonly used symptom measures – The Beck Depression Inventory- Second Edition (BDI-II, Beck et al., 1996) and The State-Trait Anxiety Inventory (STAI, Spielberger, 1983). The BDI-II is a measure clinicians use to assess clients' depression symptoms (Arbisi & Farmer, 2001). The response options to the BDI-II items make the distinction between symptom presence and severity. For example, item 5 assesses guilty feelings; the response options include

"I don't feel particularly guilty" (0), "I feel guilty over many things I have done or should have done" (1), "I feel guilty most of the time" (2), and "I feel guilty all of the time" (3). Response 0 indicates guilt is not present, whereas any answer among 1, 2, and 3 indicates guilt is present. Further, responses 1, 2, and 3 are ordered with respect to severity, with 3 being the most severe response. Likewise, item 9 assesses suicidal ideation with the response options include " I don't have any thoughts of killing myself" (0), "I have thoughts of killing myself, but I would not carry them out" (1), "I would like to kill myself" (2), and "I would kill myself if I had the chance" (3). Again, answering 0 vs. 1-3 indicates that suicidal ideation is not present versus present, and responses 1 through 3 are ordered with respect to severity, with 3 being the most severe.

The STAI also makes the presence and severity distinction for anxiety symptoms. The numeric response options differ between the STAI and BDI, but the distinction each makes is identical. For example, item 12 on the "state" subscale asks if the client "feels nervous" and the response options are "not at all" (1), "somewhat" (2), "moderately" (3), and "very much" (4). Answering 1 versus 2 through 4 indicates that nervous feelings are present versus not present at all. Likewise, responses 2 through 4 are ordered with respect to severity, with 4 being the most severe. A similar pattern is present for other items on both the state and trait portions of the STAI.

Explicit vs Implicit Models

The presence and severity components of measures can be either explicitly or implicitly built into measures (Baldwin & Olsen, 2023). Explicit measures have distinct gate and severity questions. For example, an item may ask, "Did you struggle falling asleep this week?" to which the participant responds "No" (0) or "Yes" (1). The participant is then asked, "If yes, how distressing was it to you?” with response options including “Very little” (1), “Somewhat” (2), “Quite a bit” (3), and “A lot” (4).[1]

Implicit measures have the gate and severity portions combined. For example, an item may ask, “Did you struggle falling asleep this week?”. The response options include “No” (0), “A little” (1), “Somewhat” (2), “Quite a bit” (3), and “Often” (4). If the participant chooses 0, then the symptom is not present–the “gate” is not open. In contrast, if the participant chooses 1, 2, or 3, the symptom is present, and the severity of the symptom is indicated by which response was used. Measures such as the PHQ-9 (Spitzer et al., 1999) or the GAD-7 (Spitzer et al., 2006) would be considered implicit measures.

Why Do Common Reliability and Validity Analyses Combine Presence and Severity?

The majority of commonly used psychometric models–factor analysis, reliability models, and validity models–do not make a presence versus severity distinction when applied to symptom measures. For example, it is common, so much so it is nearly universal, to treat the “not at all” response option (or similarly worded option) in implicit Two-Part measures as representing the lowest form of severity. This is a reasonable decision, as not experiencing a symptom means that the severity is low. However, it is also possible that the absence of a symptom is a qualitatively distinct experience from low severity, not just an extension or lowest end of severity. Not indicating that a symptom is present suggests that a defining characteristic of a diagnostic criteria as not being met. This could, for example, suggest that the individual may not meet the diagnostic criteria, potentially impacting the choice of treatment.

Most commonly used psychometric approaches do not allow for within-item multidimensionality and, thus, do not easily make the distinction between presence and severity.

[1] If measures are delivered electronically, then we can design it so that the severity portion is only presented to the participant if the participant responds “yes” to the gate portion.

Within-item multidimensionality refers to the idea that parts of a single item provide information about two or more dimensions. Two-Part measures, explicit or implicit, have within-item multidimensionality. Consider again the implicit measure example, “Did you struggle falling asleep this week?”. The response options include “No” (0), “A little” (1), “Somewhat” (2), “Quite a bit” (3), and “Often” (4). Within this item, the response 0 vs 1-4 provides information about symptom presence, and the responses 1-4 provide information about the symptom severity.

Within-item multidimensionality is distinct from between-item multidimensionality. Between-item multidimensionality occurs when a multi-item scale has two or more dimensions. Sometimes, in between-item multidimensionality, there are multiple dimensions, but any given item loads on one and only one dimension (i.e., simple structure; Brown, 2015). However, it is also possible to have an item load on two or more dimensions–that is, for an item to have cross-loadings. This is distinct from within-item multidimensionality as we have defined it because the cross-loadings do not make a distinction between the parts of the response. Rather, the cross-loadings simply indicate that the item as a whole reflects variability from two dimensions.

Because most developers and users of symptom scales use summed values of items to create total scores, most studies using symptom measures have assumed a unidimensional (if there is a single scale) or a between-item multidimensional model (if there are subscales). Whether this assumption is a problem is an empirical question. The next step in addressing this question is to evaluate whether symptom measures are best analyzed by a model that allows for within-item multidimensionality and separates symptom presence from symptom severity. The primary goal of this paper is to evaluate this question across multiple measures and samples.

Latent-Variable Models for Two-Part Measures

Two models have been proposed, separating symptom presence and severity. Magnus et al. proposed two zero-altered factor models–a hurdle model (Magnus & Liu, 2021, p. 940) and a zero-inflated model (Magnus & Garnier-Villarreal, 2022, p. 262). These models are conceptually similar to hurdle and zero-inflated count regression models in that they use a mixture of a logistic model to handle responses that represent the absence of symptoms (i.e., the zero portion of the model) and a graded response model[2] that represents the severity of the symptom if present. Though both models have sub-models for presence versus severity, the zero-inflated and hurdle models differ in how they treat the zeros. The zero-inflated model assumes that there are two kinds of zeros in the data: (1) Structural zeros, which are for individuals who do not ever experience the symptom, and (2) sampling zeros, which are for individuals who experience the symptoms but did not within the time frame of the measurement.

Baldwin and Olsen (2023) proposed a Sequential Item-Response Theory (IRT) model for symptom data. The sequential model, like most IRT models, has its origins in educational research. It is called the sequential model because it is appropriate for items that are in an ordered sequence. For example, a common method for explaining sequences is a simple math problem like (2+5)3 = ? (Tutz, 1990). To solve this correctly, there are two steps that must be followed in a sequence. First, a person must solve 2+5 = 7. After that step has been completed, then the person must solve 7*3 = 21. More complicated problems have more steps, but the logic is the same.

[2] In the count models, the second distribution is typically a Poisson or Negative Binomial distribution. Additionally, though Magnus et al. use a graded response model to model the severity portion of the items, in principle, other models could be used–sequential, partial credit, or even continuous.

The application of a sequential model to symptom measures has a similar logic. That is, a respondent must pass through the gate of not having the symptom to having the symptom and then onwards up the severity scale. One does not need to have gone through a prior step for the sequential model to be valid. The sequence need only refer to how the response options are ordered, not specifically how they are experienced. That said, someone who experiences symptoms over a period of time will likely move up and down the sequence as a symptom waxes and wanes.

The sequential model is useful for two reasons. First, the logic of the sequential model is similar to the logic of symptom items. Second, via the introduction of constraints and multiple latent variables, we can easily compare the fit of a severity-only model (i.e., a unidimensional model) and presence and severity model (i.e., a model with within-item multidimensionality). Loadings/discriminations for all parts of the model are on the same metric and come from the same likelihood. Additionally, all the sequential models I present are nested within each other, allowing me to use likelihood ratio tests to compare the fit of one-factor and two-factor models.

Aims of this Paper

Whether symptom presence and severity should be separated when modeling symptom measures is ultimately an empirical question involving multiple steps. In this thesis, I address the first step, which is to compare the fit of models that distinguish between presence and severity and models that do not. If distinguishing symptom presence and severity is important, then the multidimensional model will fit the data the best. To this end, this thesis compared three different models: (1) the Extreme Response model, (2) the Unique Relationship model, and (3) the Two-Part sequential model. The details of each model are described in the Methods section.

I hypothesize that the Unique Relationship model will fit better than the Extreme Response model across datasets because the Unique Relationship model allows for the gate response to have a Unique Relationship with the latent construct as compared to the severity responses. In contrast, the Extreme Response model forces the gate response to have the same relationship with the latent variable as the severity responses. I also hypothesize that the Two-Part sequential model will fit better than the Unique Relationship model because the Two-Part model introduces a new latent factor representing presence, whereas the Unique Relationship model is a unidimensional model. I hypothesize that the Two-Part model will have a better fit than the other two models.

Methods

Datasets and Participants

Datasets were drawn from multiple sources. Dataset 1 is a combination of datasets reported across multiple electroencephalograph (EEG) studies, which have been discussed in multiple publications (Clawson et al., 2013; Clayson et al., 2011; Clayson & Larson, 2011a, 2011b, 2012, 2013; Larson et al., 2010; Larson et al., 2013; Larson & Clayson, 2010; Larson et al., 2012; Larson et al., 2011; Larson et al., 2013). The data are drawn from N = 792 people (443 = female, 349 male) ranging in age from 17 to 52. We used item-level data from the State-Trait Anxiety Inventory (STAI), Positive and Negative Affect Scale (PANAS), and Beck Depression Inventory Second Edition (BDI-II; see below for references and discussion of measures).

Dataset 2 is an unpublished dataset from an EEG study of learning during the menstrual cycle (Clawson et al., 2013; Clayson et al., 2013). The data are drawn from N = 78 women ranging in age from 18 to 26 (88% white, 4% mixed race, and 8% Hispanic). We used item-level data from the Penn-State Worry Questionnaire (see below). Though this study also included the STAI, BDI-II, and Profile of Mood States measure, we were not able to fit the models because there was not sufficient variability in the responses due to most participants not reporting symptoms.

Dataset 3 was drawn from a study of students taking a family studies course at a large southeastern university (Braithwaite et al., 2010). The sample consists of N= 5,372 (1354 men, 4,018 women) people (69% white, 13% African American, 10% Hispanic, and 8% other) ranging in age from 17 to 55 (mean = 20.63). We used item-level data from the Center of Epidemiological Study of Depression and PANAS. Of those participants, we used the data from N = 5263 responded to the CESD, and N = 950 responded to the PANAS.

Measures

As noted above, the questionnaires used in the data set included the State-Trait Anxiety Inventory (STAI, Spielberger, 1983), the Beck Depression Inventory- Second Edition (BDI-II, Beck & Brown, 1996), the Positive and Negative Affect Schedule (PANAS, Watson et al., 1988), Penn State Worry Questionnaire (PSWQ, Meyer et., al, 1990b), and Center for Epidemiological Studies Depression (CESD-10, Andresen et al., 1994; CESD-20, Radloff, 1977).

State-Trait Anxiety Inventory . The STAI (Spielberger, 1983) is a measure that assesses for trait and state anxiety; it is utilized in clinical settings to distinguish between anxiety and depressive symptoms. The STAI includes 40 items, 20 each for the two subscales: state and trait anxiety. The twenty items for the state anxiety scale measure “how one feels right now.” For example, a state item is “ I am tense,” and the response options are: (1) not at all, (2) somewhat, (3) moderately, and (4) very much. The twenty items for the trait anxiety scale measure “how one generally feels." For example, a trait item is "I try to avoid facing crisis or difficulty," and the response options are (1) not at all, (2) somewhat, (3) moderately, and (4) very much. Test-retest reliability for the state scale was *r* = .33 for men and *r* = .16 for women, and for the trait scale was *r* = .84 for men and *r* = .76 for women during 20 days.

The Beck Depression Inventory- Second Edition. The BDI-II is a commonly used self-report inventory that assesses the severity of depression in adolescents and adults (Carlson & Waller, 1998). The current BDI-II has a total of 21 items, and responses are rated on a 4-point scale (Carlson & Waller, 1998). For example, one of the items assessed for "crying," and the response to this statement is (0) I don't cry any more than I used to, (1) I cry more than I used to, (2) I cry over every little thing, and (3) I feel like crying, but I can't. Estimates of internal consistency for the BDI-II are $\alpha = 0.92$ (outpatient sample) and $\alpha = 0.93$ (nonclinical sample). The test-retest reliability coefficient across a period of one week was *r* = 0.93 (Arbisi & Farmer, 2001).

The Positive and Negative Affect Schedule. The PANAS (PANAS, Watson et al., 1988) is a self-report measure with two subscales, positive and negative affect. There are 20 items, 10 for each subscale. The positive affect (PA) scale measures how much a person feels enthusiastic, active, and alert (Watson et al., 1988). A low score on the PA scale would suggest that the participant is sad and has low energy. An example of an item from the PA scale is, "indicate the extent you have felt this way over the past week- enthusiastic," to which the participant would respond, (1) very slightly or not at all, (2) a little, (3) moderately, (4) quite a bit, (5) extremely. In contrast, the negative affect (NA) scale measures the extent to which a person feels anger, contempt, disgust, guilt, fear, and nervousness. A low NA scale score suggests that the participant tends to be calm. An example of an item for the NA scale is, “indicate the extent you have felt this way over the past week-hostile,” to which the participant would respond, (1) very slightly or not at all, (2) a little, (3) moderately, (4) quite a bit, (5) extremely. Estimates of internal consistency for the PA scale ranged from $\alpha =$.86 to $\alpha =$ .90, and the NA scale ranged from $\alpha =$.84 to $\alpha =$.87 (Watson et al., 1988, as cited in Crawford & Henry, 2004).

The Penn State Worry Questionnaire. The PSWQ (PSWQ, Meyer et al., 1990a) is a self-report questionnaire on worry. The 16 items for the PSWQ were developed from a factor analysis of a large pool of items, and overall, the measure correlated highly with the trait anxiety (Meyer et al., 1990a). An example of a question in the PSWQ is item 7, which states, “ I am always worrying about something”, which the respondent then is asked to the extent that this item represents using the following options: (1) not at all typical of me, (2) not very typical of me, (3) somewhat typical of me, (4) fairly typical of me, or (5) very typical of me. Regarding the reliability of the measure, the test-retest reliability coefficient across a period of 8 to 10 weeks was. $r =$ 0.92 (Meyer et al., 1990a). Estimates of internal consistency for the PSWQ were $\alpha =$ 0.93. Furthermore, the PSWQ correlated significantly with the Cognitive Somatic Anxiety Questionnaire (CSAQ, Schwartz et al., 1978) total $r(46) = .69, P < 0.001.$ In a study that had 34 clients who had all been diagnosed with Generalized Anxiety Disorder (GAD), the PSWQ was administered alongside a couple of others depression and anxiety related measures (Meyer et al., 1990a) prior to clients receiving treatment, results indicated that after cognitive therapy there was a great reduction in PSWQ scores. These findings indicated that the construct of worry is independent of the construct of anxiety and depression.

Center for Epidemiological Studies Depression. The CESD is a self-report scale that assesses mood, somatic complaints, interaction with others, and motor functioning (Eaton et al., 2004). Dataset 3 includes both the 10-item and the 20-item versions. An example of a question in the CESD-10 and CESD-20 is, “During the past week: I had trouble keeping my mind on what I was doing,” with response options of (1) Rarely or none of the time, (2) Some or a little of the time, (3) Occasionally or a moderate amount of time, (4) Most or all the time. The Cronbach’s alpha for the CESD-20 was $\alpha = .91$, and α=.86 for the CESD-10 (Miller et al., 2007). The test-retest reliability coefficient across a period of two weeks for the CESD-20 was ICC=.87 (95% CI 0.79-0.93) and ICC=0.85 for the CESD-10 (95% CI 0.75-0.92) (Miller et al., 2007).

Data Analysis

I fit the sequential IRT models using binary confirmatory factor analysis (CFA). All analyses were conducted using the `gsem` command in Stata 18 (StataCorp, 2023). To estimate the sequential model with binary CFA, each item’s responses must be transformed into a set of pseudo-items representing contrasts among sequential levels. We used what is called a forward continuation ratio, inverse odds model because the contrasts represent the increasing steps of symptom items. See Baldwin and Olsen (2023) for more details about other coding options.

To illustrate the contrasts, consider the Beck Depression Inventory- Second Edition, which has response options ranging from 0 to 3. The four options will be represented by three dummy variables or pseudo-items, which themselves represent the sequence of responses. Consider item 5 from the BDI-II again, which assesses feelings of guilt and has response options ranging from 0 to 3 (0 is the lowest). Dummy 1 for item 5 (d_{15}) is 0 if a person answers 0 and 1 if a person answers 1-3. Dummy 2 for item 5 (d_{25}) is 0 if a person answers 1, 1 if a person answers 2-3, and missing otherwise. Dummy 3 for item 5 (d_{35}) is 0 if a person answers 2, 1 if a person answers 3, and missing otherwise. Table 1 shows the dummy values for two hypothetical responses to item 5 for the BDI-II. Though the number of dummy indicators will vary based on response options, the structure is the same: 0 for the response, 1 for anything above the response, and missing for anything below the response.

Table 1. *Example dummy variables for all possible responses to the BDI-II Item 5.*

Response 1	Dummy 1	Dummy 2	Dummy3
0	0	.	.
1	1	0	.
2	1	1	0
3	1	1	1

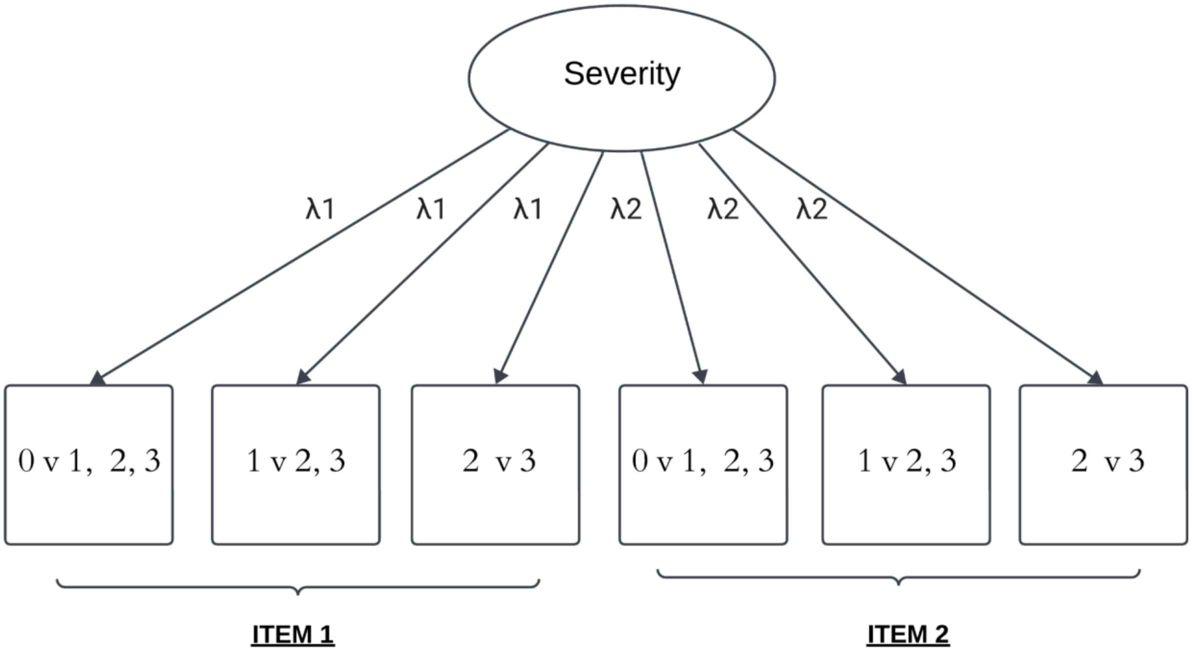
Note. Hypothetical responses to Item 5 on the BDI-II. Response options range from 0 to 3. Cells with a period "." are coded as missing.

Once the dummy variables have been created, we can then fit the model as a binary confirmatory factor analysis.[3] The dummy variables are used in place of the original response variable. We considered three models for each measure: (a) Extreme Response, (b) Unique Relationship, and (c) Two-Part. Figures 1-3 are the path diagrams for the three models. In the Extreme Response model, the gate response (i.e., not experiencing the symptom) is not considered unique but just the most Extreme Response on severity (on the low end). The Extreme Response model includes just a single latent variable that loads on the dummy variables

[3] Binary factor analysis parameters can be converted to traditional item-response theory parameters (i.e., difficulty and discrimination). In this manuscript, the research question is primarily addressed by the relative fit of the models rather than the specific values of the parameters themselves. Consequently, for convenience we report all parameters in the factor analysis metric (i.e., thresholds and loadings).

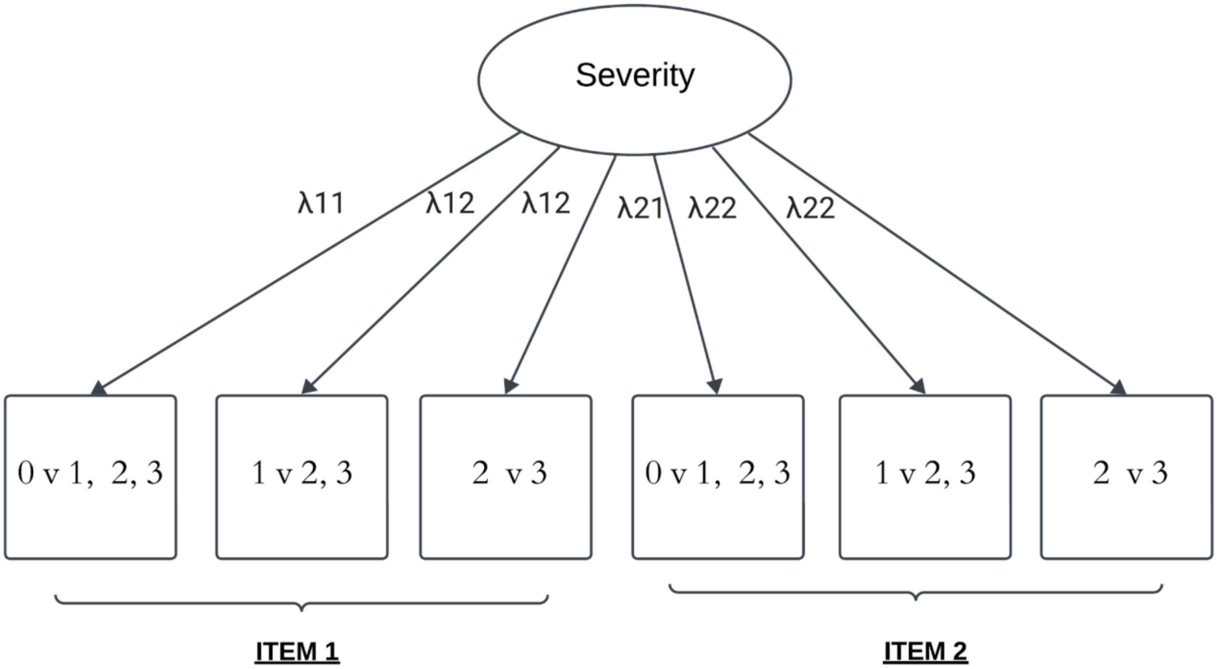
for all relevant items, and the latent variable is best thought of as an index of severity. Additionally, within an item, all loadings for the dummy variables are constrained to be equal. In other words, all sequences represented by the dummy variables provide the same information about the latent variable.

Figure 1. Extreme Response Model. The gate response is represented by the box 0 v 1, 2, 3. Loadings are represented by the λ sign.



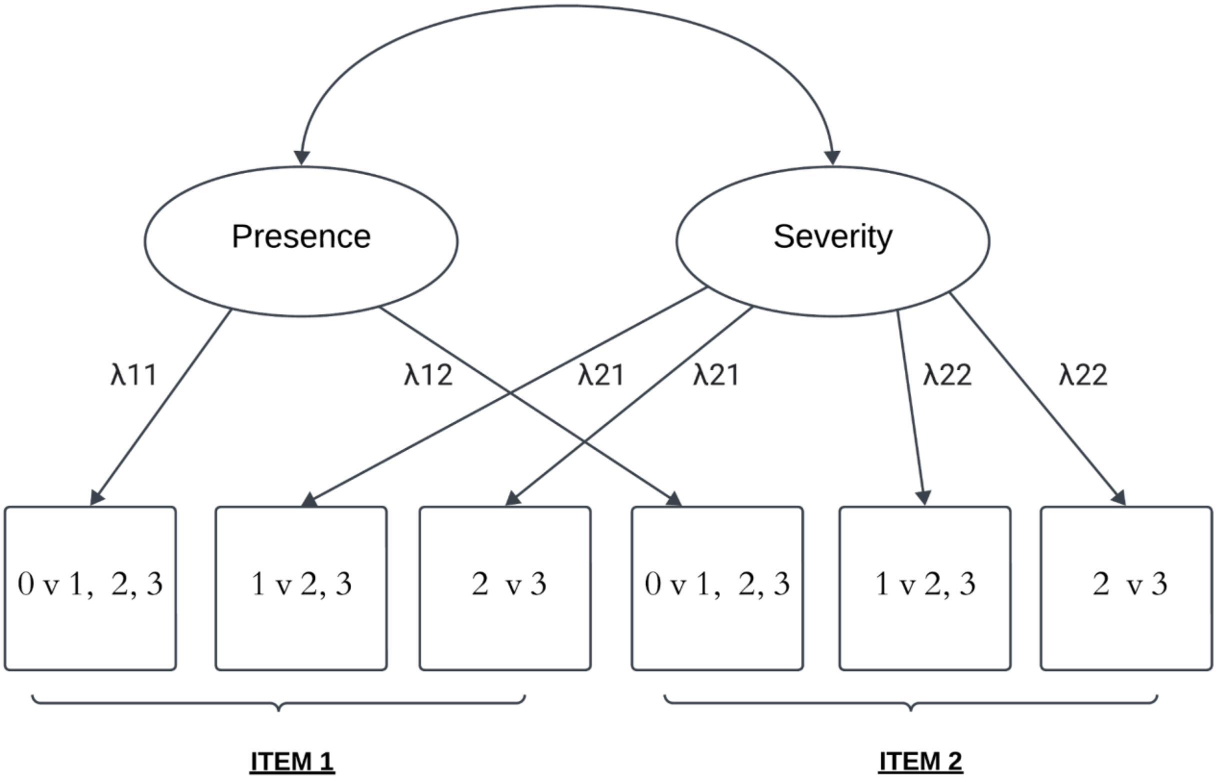
In the Unique Relationship model, there is also only one latent variable that loads on all dummy variables. However, within an item, the gate dummy variable is allowed to have a unique relationship with the latent variable as compared to the dummy variables that represent symptom presence. Thus, this model allows for the gate response to provide unique information about severity, but it is still an index of severity.

Figure 2. Unique Relationship Model. The gate response is items are represented by the box 0 v 1, 2, 3. Loadings are represented by the λ sign.



The Two-Part model includes two latent variables, one for the gate portion of items and one for the severity portion of items. The gate latent variable loads on the first dummy variable for all items and loadings are allowed to be unique across items. The severity portion loads on all remaining dummy variables. In this part of the model, loadings are allowed to vary across items, but within an item, loadings are constrained to be equal.[4]

Figure 3. Two-Part Model. The gate response is items are represented by the box 0 v 1, 2, 3. Loadings are represented by the λ sign.



[4] Baldwin and Olsen (2023) considered versions of the Unique Relationship and Two-Part models that did not constrain the within-item loadings. They did not perform any better than the constrained version so we have limited our analysis to the constrained version.

For each dataset, we fit all three models. Because the models are nested within one another, we can compare the fit of the model with the likelihood ratio test. The likelihood ratio test is distributed χ^2 with degrees of freedom equal to the difference in parameters across the two models. The null hypothesis is that the more complicated model does not fit the data better than the constrained model, and the α-level for this test was set at 0.05. We compared the Extreme Response to the Unique Relationship model–significance indicates that the Unique Relationship model fits better than the Extreme Response model. We also compared the Unique Relationship to the Two-Part model–significance indicates that the Two-Part model fits better than the Unique Relationship model.

To supplement the likelihood ratio test, I also examined Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) because both fit indices help correct for the increased complexity of the Unique Relationship model as compared to the Extreme Response and Two-Part model as compared to the Unique Relationship. Smaller values of the AIC and BIC are better, and as there is no threshold for considering something "significantly" smaller, I treated any smaller value as indicating a better fit.

Results

Model Comparisons

Table 2 presents the fit statistics for each dataset and model we used. Table 2 is organized by study (Column 1) and measures nested within the study (Column 2). For every dataset and measure, there are three rows of data. Row 1 is for the Extreme Response model, Row 2 the Unique Relationship model, and Row 3 for the Two-Part model (Column 3). Column 4 provides the χ^2 values for the likelihood ratio test. The χ^2 for Row 2 compares the Extreme Response and Unique Relationship models, and the χ^2 for Row 3 compares the Unique Relationship to the Two-Part model. Columns 4 and 5 provide the AIC and BIC values, respectively.

Table 2.

Comparison of the Extreme Response, Unique Relationship Model, and the Two-Part model.

Dataset	Measure	Model	Chi-square	AIC	BIC
1. Dataset 1	STAI-S	ER	NA	28381.7	28755.66
		UN	311.93(20)*	28109.77	28577.22
		TP	267.11(1)*	27844.66	28316.79
2. Dataset 1	STAI-T	ER	NA	29060.79	29434.75
		UN	51.77(20)*	29049.02	29516.47
		TP	233.85(1)*	28817.17	29289.3
3. Dataset 1	BDI-II	ER	NA	21354.6	21742.58
		UN	57.18(20)*	21339.42	21825.57
		TP	92.41(1)*	21249.01	21739.84
4. Dataset 1	PANAS (PA)	ER	NA	11241.97	11447.43
		UN	22.36(10)*	11239.61	11486.17
		TP	26.44(1)*	11215.18	11465.84
5. Dataset 1	PANAS (NA)	ER	NA	10048.94	10254.4
		UN	41.22(10)*	10027.72	10274.27
		TP	57.51(1)*	9972.212	10222.88

6. Dataset 2	PSWQ	ER	NA	2978.251	3166.788
		UN	14.81(16)	2995.44	3221.684
		TP	24.18(1)*	2973.257	3201.858
7. Dataset 3	CESD-20	ER	NA	131536.4	132061.9
		UN	683.30(20)*	130893.1	131549.9
		TP	415.37(1)	130479.7	131143.1
8. Dataset 3	CESD-20	ER	NA	116262.1	116778.5
		UN	815.33(20)*	115486.7	116132.3
		TP	746.50(1)*	114742.2	115394.2
9. Dataset 3	CESD-20	ER	NA	103473.2	103982.7
		UN	989.50(20)*	102523.7	103160.6
		TP	771.65(1)*	101754.1	102397.3
10. Dataset 3	PANAS (NA)	ER	NA	19858.71	20101.53
		UN	61.39(10)*	19817.32	20108.7
		TP	85.97(1)*	19773.35	20029.59
11. Dataset 3	PANAS(NA)	ER	NA	19255.55	19496
		UN	75.90(10)*	19199.65	19488.2
		TP	257.77(1)*	18943.88	19237.24

12. Dataset 3	PANAS(NA)	ER	NA	17707.12	17945.2
		UN	77.35(10)*	17649.76	17935.46
		TP	222.38(1)*	17429.38	17719.83
13. Dataset 3	PANAS(NA)	ER	NA	18877.29	19119.58
		UN	40.06(10)*	18857.23	19147.98
		TP	128.50(1)*	18730.73	19026.33
14. Dataset 3	PANAS(NA)	ER	NA	17749.48	18033.7
		UN	55.66(10)*	17758.81	18045.88
		TP	221.07(1)*	17539.74	17831.6
15. Dataset 3	PANAS(NA)	ER	NA	17737.92	17975.65
		UN	113.11(10)*	17644.81	17930.88
		TP	255.01(1)	17391.8	17681.83
16. Dataset 3	PANAS (PA)	ER	NA	24935.05	25177.87
		UN	19.21(10)* (p>0.0377)	24935.84	25227.23
		TP	225.62(1)*	24712.22	25008.47
17. Dataset 3	PANAS (PA)	ER	NA	23059.1	23299.55
		UN	64.67(10)*	23014.43	23302.97

		TP	223.81(1)*	22792.62	23085.97
18. Dataset 3	PANAS (PA)	ER	NA	21666.98	21905.06
		UN	32.52(10)*	21654.47	21940.16
		TP	361.38	21295.09	21585.55
19. Dataset 3	PANAS (PA)	ER	NA	23987.84	24230.19
		UN	42.11(10)*	23965.74	24256.55
		TP	203.57(1)*	23764.16	24059.83
20. Dataset 3	PANAS (PA)	ER	NA	21907.47	22146.69
		UN	60.75(10)*	21866.71	22153.78
		TP	170.93(1)*	21697.78	21989.63
21. Dataset 3	PANAS (PA)	ER	NA	21238.71	21476.44
		UN	43.52(10)*	21215.19	21500.47
		TP	244.60(1)*	20972.59	21262.63

Note. * $p < 0.05$; ER = Extreme Response Model; UN = Unique Response Model; TP = Two-Part Model; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; NA = not applicable.

To assist the reader in interpreting Table 2, I will interpret the χ^2, AIC and BIC for Dataset 1 STAI-S rows. The likelihood ratio test comparing the Extreme Response (ER) and Unique Relationship (UN) models was significant $\chi^2(20) = 311.93, p < 0.05$. Likewise, the AIC and BIC for the Unique Relationship model (28109.77 and 28577.22) were smaller than the AIC and BIC for the Extreme Response model (28381.7 and 28755.66). These comparisons suggest that estimating a unique loading for the gate response significantly improves model fit over constraining the gate response to be equal with the severity responses.

The likelihood ratio test comparing the Unique Relationship (UN) and Two-Part (TP) models was also significant$\chi^2(1) = 267.11, p < 0.05$. Likewise, the AIC and BIC for the Two-Part model (27844.66 and 28316.79) were smaller than the AIC and BIC for the Extreme Response model (28109.77 and 28577.22). These comparisons suggest that introducing a separate latent variable for the gate improves model fit over just estimating a unique loading for the gate response. The rest of the entries in Table 2 can be interpreted in a similar manner.

The results presented in Table 2 suggest that gate responses have a different relationship to the latent constructs than severity responses. The likelihood ratio test comparing the Extreme Response model and the Unique Relationship model was statistically significant 20 out of 21 times. The AIC of the Unique Relationship model was smaller than that of the Extreme Response model 18 out of 21 times. The BIC of the Unique Relationship model was smaller than the Extreme response in 7 out of 21 times. One possible explanation for these findings is that the BIC penalizes complexity more than the AIC. Thus, the additional complexity introduced in the Unique Relationship model by assuming that the 'gate' response loads differently than other responses, though it improved fit, did so at the expense of parsimony. On the whole, my findings suggest that the Unique Relationship model fits the data better than the Extreme Response model, but the additional complexity of the Unique Relationship model may overcomplicate it.

Table 2 also suggests that adding the presence of a latent variable improved fit as compared to just estimating a unique loading for the gate response. The likelihood ratio test comparing the Unique Relationship to the Two-Part model was significant 21 times. Likewise, the AIC and the BIC for the Two-Part model were smaller than the AIC and the BIC for the Unique Relationship model all 21 times.

Areas of Strain

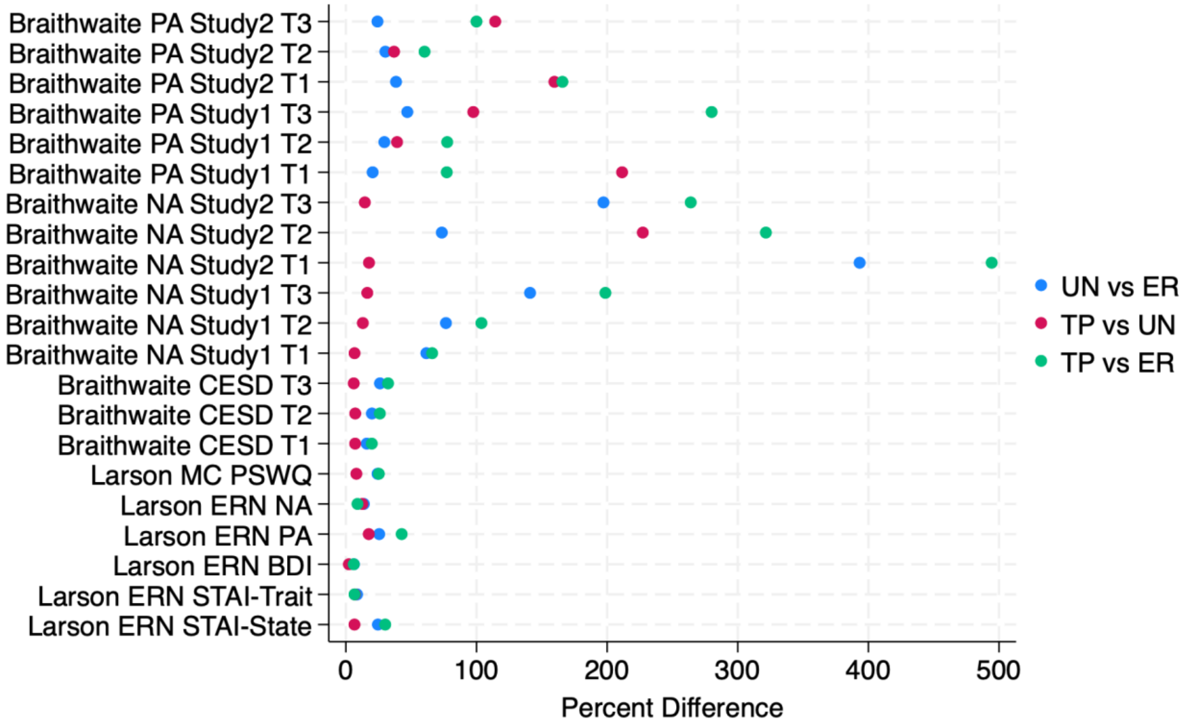
Strain in the context of IRT (or factor analysis generally) refers to how constraints in a model create problems in the fit. For example, it is common in a confirmatory factor analysis to, by default, assume that residual errors are uncorrelated (i.e., the residual correlation is constrained to 0). However, if two items share the same item stem, there may be a correlation between the items that are unaccounted for by the latent variable alone. Thus, forcing the residual correlation to be 0 “strains” the model. In the sequential models, forcing the gate response to have the same loading as the severity responses in the Extreme Response model could produce strain in the model fit. Likewise, including just one latent variable in the Unique Relationship model, which is like estimating both a presence and severity latent variable but forcing the correlation to be 1, could also produce strain in the model.

Strain in the model can affect model parameters, such as the size of loadings. Consequently, we compared the size of the loadings across models. When comparing the size of the loadings, it is important to keep in mind that the Unique Relationship model generally fits the data better than the Extreme Response model, and the Two-Part model fits better than the Unique Relationship model. Thus, we can have more confidence in the loading in the better-fitting model than the poorer-fitting model. To examine the possible strain that comes about from treating the gate response as the floor of severity, I attended to the degree to which models where the gate response was unconstrained (Unique Relationship and Two-Part) had loadings that differed from the Extreme Response model where the gate response was constrained. I examined the effect of the constraint on both the estimate of the loading for the gate response option as well as the loading for the severity response options.

Figures 4 and 5 present the percent change (absolute value) when comparing the Unique Relationship to Extreme Response models (blue dot), the Two-Part to the Unique Relationship models (red dot), and the Two-Part to Extreme Response models (green dot). The y-axis indicates the dataset and outcome, and the x-axis is the percent change.

If the blue and green dots produce the biggest discrepancies between models, then that is consistent with the idea that the constraint in the Extreme Response model introduces strain as compared to the Unique Relationship and Two-Part Models. If the red and green dots produce the biggest discrepancies, then that is consistent with using a single latent variable as the source of the strain. As illustrated in Figure 4, which is focused on the gate response option, the largest discrepancy is between either the blue or green dots or the red dots, which indicates that including a distinct loading (Unique Relationship model) or a presence latent factor (Two-Part model) helped alleviate strain in the Extreme Response model. Although the Two-Part model and the Unique Relationship model showed some discrepancy, it was relatively small for most models.

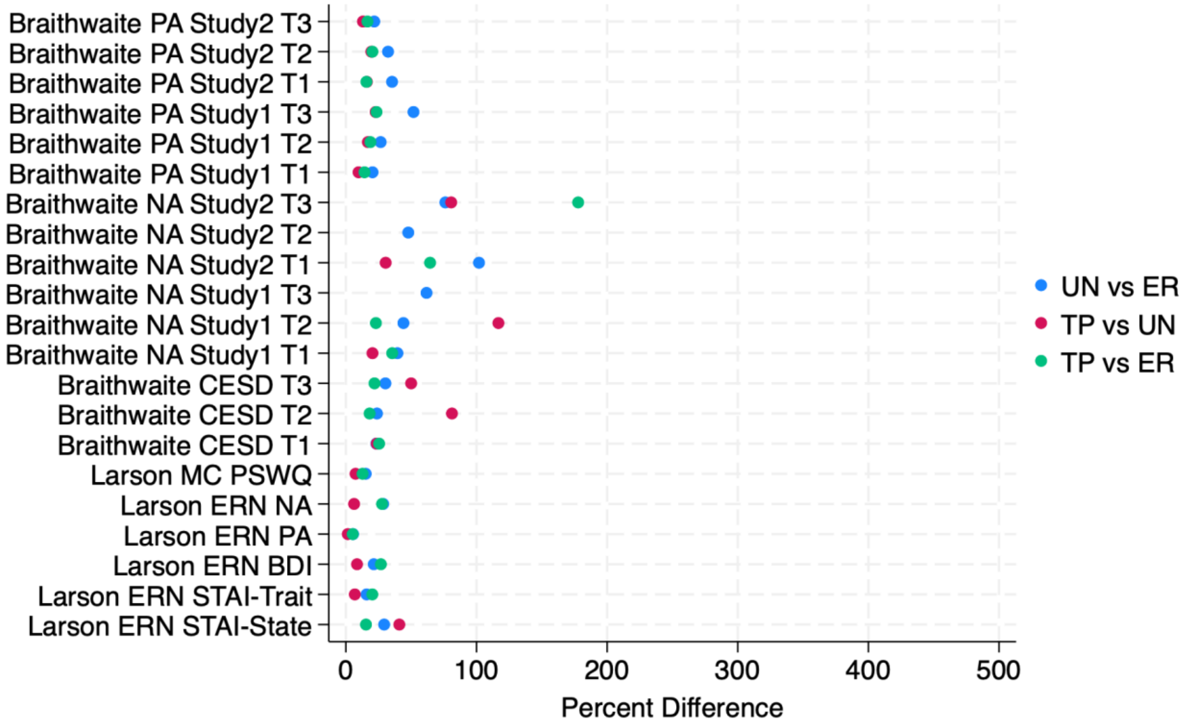
Figure 4. Absolute value of the percent change in the gate response loadings among the three models.



Note. * Gate; ER = Extreme Response Model; UN = Unique Relationship Model; TP = Two-Part Model; PA = PANAS Positive Affect; NA= PANAS Negative Affect; CESD= Center for Epidemiologic Studies Depression Scale; PSWQ= Penn State Worry Questionnaire; STAI= The State-Trait Anxiety Inventory

As seen in Figure 5, the loadings for the severity response showed less change among the models than the gate response. The Two-Part model and the Unique Relationship model appear to have the biggest discrepancies, followed by the Two-Part and Extreme Response models, which suggests that separating severity from presence in the latent structure of the model affected the size of the loadings.

Figure 5. Absolute value of the percent change in the severity responses loadings among the three models.



Note. *Severity; ER = Extreme Response Model; UN = Unique Relationship Model; TP = Two-Part Model; PA = PANAS Positive Affect; NA= PANAS Negative Affect; CESD= Center for Epidemiologic Studies Depression Scale; PSWQ= Penn State Worry Questionnaire; STAI= The State-Trait Anxiety Inventory

In summary, as it was hypothesized the Two-Part model had the best overall fit of the three models. This suggests that having a separate latent variable for the gate response improves the fit of the model. Furthermore, my results indicated that there was a discrepancy between the Two-Part model and the Extreme Response and Unique relationship models in terms of gate response loadings. These findings suggest that the single latent variable is a source of the strain on the fit of the model. Together, the results suggest that the Two-Part Model outperformed the

Extreme Response and Unique Relationship models, demonstrating the importance of the introduction of a separate latent variable for the gate response.

Discussion

Review of Results

Our findings suggest that the Two-Part model had the best overall fit across the twenty-one data sets when compared to the Extreme Response model and the Unique Relationship model. These results indicate that the gate response has a different relationship to the latent constructs, and by creating a new latent construct for the gate response, the model fits improve. Results also suggest that constraints in the gate response, as seen in the Extreme Response model, introduce strain in the model and affect factor loadings. Additionally, the results suggest that combining both symptom presence and symptom severity into a single variable also introduces strain in the model.

Potential Changes to Measures

My findings support the idea of separating presence from severity when using symptom measures. Consequently, developers of symptom measures could consider adopting an explicit two-part structure for their measures. That is, respondents can initially specify what symptoms are present and then rate the severity of the symptoms accordingly. For instance, item 17 on the STAI is “I am worried.” As it is currently structured, respondents rate this statement from 1 to 4, with 1 being the gate response and the other options representing severity if the symptoms are present. Revising the STAI to an explicit structure would mean that respondents are prompted to indicate whether worry is present as a symptom by selecting (1) yes or (0) no. If the symptom is present, the respondent would then assess the severity of the symptom using a scale, with options ranging from (1) Somewhat, (2) Moderately, and (3) Very much. The explicit structure could help respondents more carefully consider whether they are experiencing a symptom at all and hopefully provide clinicians with a more robust picture of what patients are experiencing.

Clinical Implications

Clinicians utilize the Diagnostic and Statistical Manual of Mental Disorders (5th ed; DSM-5; American Psychiatric Association, 2013) to attempt to categorize symptoms for each specific mental health disorder. The diagnostic criteria outlined by the DSM-5 form the basis for developing psychometric measurements, which are specifically designed to assess the distinct symptoms associated with various mental health disorders. These psychometric measurements designed for diagnostic purposes are constructed based on the symptoms associated with each specific mental health disorder, taking into account both symptom presence and severity criteria. The Two-Part model considers both the presence and severity of symptoms diagnostic criteria, thereby improving the accuracy of psychometric measurements in diagnosis. For example, the GAD-7 (Spitzer et al., 2006) is a seven-item anxiety scale that assesses for symptoms of Generalized Anxiety Disorder (GAD). The measurement has a strong criterion validity associated with a diagnosis of GAD (Spitzer et al., 2006), indicating its effectiveness in accurately assessing symptoms associated with GAD. By implementing the Two-Part model to a measure such as the GAD-7, clinicians would have two columns in each item, one that would indicate the symptom presence- (0) no, and (1) yes, and one that would indicate symptom severity through a Likert scale- if yes how severe (1-4). Thus the Two-Part model would address the need to assess for both the symptom presence and severity.

The Two-Part model not only enhances the diagnostic process but also has the potential to improve treatment efficacy. Implementing the Two-part model into a measurement like the GAD-7 would yield test results that not only list all present symptoms and their severities but also generate a hierarchy of symptoms from the most to the least severe. This hierarchy enables clinicians to prioritize and categorize which symptoms should be addressed first in treatment. For example, if someone indicates a score of 15 or greater on the GAD-7, which indicates severe anxiety, the measure would not only indicate which symptoms were present but would also provide a ranking of which symptom of anxiety is the worst for the individual, thus providing concrete information that the clinician can utilize when considering an effective treatment plan. For example, in the case of the GAD-7, if a client reports that their most severe symptom is "feeling afraid, as if something awful might happen," clinicians who follow a Cognitive Behavior Therapy (CBT) framework could prioritize exposures as their first intervention to help the client deal with their fear. A hierarchy of symptoms could provide critical information for overall treatment. Future research is necessary to determine whether implementing this model in psychodiagnostic measures would improve overall treatment success.

Potential Future Research

Furthermore, future research should also focus on broadening symptom pathology. For this study, we only focused on measures that assess for Anxiety and Depression. Future research should concentrate on applying these measures to more severe psychopathology to determine the applicability of the model across various forms of psychopathology. There are several measures that assess for disorders such as eating disorders, OCD, and Bipolar that could potentially benefit from a separation of the symptom presence from symptom severity. For example, the Young Mania Rating Scale (YMRS), which assesses Bipolar disorder, asks in the fourth item about sleep; the respondent chooses on a Likert scale between 0 to 4 what answer best resonates with their current situation: (0) reports no decrease in sleep, (1) Sleeping less than normal amount by up to one hour, (2) sleeping less than normal by more than one hour, (3) reports decreased need for sleep, and (4) Denies need for sleep. A response of 0 signifies the absence of the Bipolar symptom related to needing less sleep, indicating that the symptom is simply not present. On a scale of 1-4, a higher score indicates the presence and increased severity of the symptom, reflecting the person’s experience. Future research could investigate if implementing the Two-Part models to measurements such as this would improve overall assessment capability in these different measurements.

Another area of future research could focus on how to implement the Two-Part Sequential model in a clinical setting. Future research could help sort out how best to use a Two-Part framework for clinicians still using paper versions of measures, as well as the potential benefits of using electronic versions of measures or, at least, electronic methods of scoring measures.

Need for Validation

Future research should aim to validate whether symptom severity and symptom presence measure different constructs. Does the distinction between questions about symptom presence and questions about symptom severity have any predictive or diagnostic significance? For example, with common diagnoses such as depression, how would distinguishing between presence and the severity affect diagnostic decisions or predictions about prognosis? Diagnostic criteria for most disorders involve not just whether a symptom is present but also a judgment about how severe the symptom is. Will explicitly separating presence from severity in our measures and models assist with these decisions, or will it be irrelevant?

Another area of construct validity to explore is the extent to which presence and severity are differentially correlated with key outcomes. For example, does symptom severity and presence have differential relationships with co-morbid conditions? Are people with high levels of symptoms the most or the least likely to have co-morbid conditions? Is severity the key question when it comes to co-morbidity? Treatment is another place we can consider differential relationships. Does therapy or medication reduce the number of symptoms or just the severity of them? If it is just the severity, how might that affect patients' expectations for treatment and what life with a disorder is? These are just some examples of potential areas of research as we explore whether the presence-severity distinction is important or much ado about nothing.

Potential Constraints to the Model

A potential limitation of the Two-Part model is attitude measures. When measuring for psychological constructs such as depression, there is an assumption that there will be a presence or an absence of a particular symptom. This differs in attitude measure constructs because the items in said measure usually don't focus on identifying the presence of a construct but rather on describing the range of said construct. Attitude measures range from very unfavorable to very favorable; this suggests that there will always be a midpoint in each item that represents neutrality or indifference instead of absence or lack of a construct (Blanton & Jaccard, 2006). The Two-Part model only fits if there is first a "gate," suggesting symptom presence, introducing the severity estimate—an attitude scale only measures the person's attitude or belief about a specific construct. Attitude measures would include any Likert scale survey with questions and responses of a neutral nature. Unlike psychometric screeners, these attitude measures do not assess for symptom presence, "gate," before determining symptom severity. These attitude measures are not attempting to answer whether something is there but attempting to determine how a person feels about a particular construct. For example, an attitude measure may ask, "How do you feel about vanilla ice cream?" The responses would include answers such as 3 (I love it), 2(I like it a lot), 1(I like it), 0 (neutral), -1 (I am not a fan), -2(I don't like it), -3(I hate it).

The attitude scale has a "neutral" option, which allows the person to respond in a nature that is neither positive nor negative. It differs from the Two-Part-sequential model that depends on whether a symptom is present or absent. The zero in a psychological disorder screener suggests that the symptom is either not present or, if it is characteristic of the symptom, is of minimal distress. The zero in the attitude measure assesses whether the individual finds themselves "unsure" or "neutral" about a specific construct. Thus, the zero in the attitude measure significantly differs from the zero in psychological measurements, such as the PHQ-9, which assesses if the symptom is present. Whereas the zero in the attitude scale represents neutrality, the one in the PHQ-9 represents the absence of symptoms. The Two-Part sequential model would not fit well within attitude scales because attitude scales simply do not ask about symptom presence or symptom severity.

Limitations

The samples I used in my study were a mixture of college students and some patients. It is possible that different results could arise with the measures used in this study if I had an exclusively clinical population. If I had utilized a clinical population, a broader range of severity responses would have been present, potentially favoring the two-part model more. Another limitation of the data led to challenges in fitting the models. Because many of the participants were healthy, there was limited variability in the responses (i.e., many people not reporting any symptoms). This affects the ability of the model to converge to a solution (i.e., the data is functionally a constant).

Conclusion

In conclusion, the Two-Part model had the best overall fit when compared to the Extreme Response model and the Unique Relationship model. This finding is significant because it could affect the way clinicians assess symptoms. By distinguishing between the presence and severity of symptoms, clinicians could potentially gain a more precise insight into which symptoms should be prioritized for intervention, potentially enhancing treatment outcomes. Future research endeavors should prioritize understanding the impact of distinguishing between symptom presence and severity on treatment outcomes, thereby enhancing our ability to optimize therapeutic interventions.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Andresen, E. M., Malmgren, J. A., Carter, W. B., & Patrick, D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventive Medicine*, *10*(2), 77–84. https://doi.org/10.1016/s0749-3797(18)30622-6

Arbisi, P.A., & Farmer, R. F. (2001). Beck Depression Inventory-Second Edition. *The Fourteenth Mental Measurements Yearbook.*

Baldwin, S.A. & Olsen, J. A. (2023). Two-part sequential measurement models for distinguishing between symptoms presence and symptom severity. Unpublished Manuscript

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory–II. *PsycTESTS Dataset*. https://doi.org/10.1037/t00742-000

Blanton, H., & Jaccard, J. (2006). Tests of multiplicative models in psychology: A case study using the unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 113*(1), 155–166. https://doi.org/10.1037/0033-295x.113.1.155

Braithwaite, S. R., Lambert, N. M., Fincham, F.D., & Pasley, K. (2010). Does college-based relationship decrease extradyadic involvement in relationships? *Journal of Family Psychology*, *24*(6),740-745. https://doi.org/10.1037/a0021759

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Carlson, J.F., & Waller, N. G. (1998) Beck Depression Inventory [1993 Revised]. *The Thirteenth Mental Measurements Yearbook.*

Clawson, A., Clayson, P. E., & Larson, M. J. (2013a). Cognitive control adjustments and conflict adaptation in major depressive disorder. *Psychophysiology*, *50*(8), 711–721. https://doi.org/10.1111/psyp.12066

Clawson, A.*, Clayson, P.E., & Larson, M.J. (2013b). Dopaminergic influences on performance monitoring across the menstrual cycle. Poster presented at the 41st Annual Meeting of the International Neuropsychological Society, Waikoloa, Hawaii.

Clayson, P.E., Clawson, A.*, & Larson, M.J. (2013c). The effects of dopamine on reinforcement learning across the menstrual cycle: An electrophysiological investigation. Poster presented at the 41st Annual Meeting of the International Neuropsychological Society, Waikoloa, Hawaii.

Clayson, P. E., & Larson, M. J. (2011a). Conflict adaptation and Sequential Trial Effects: Support for the conflict monitoring theory. *Neuropsychologia*, *49*(7), 1953–1961. https://doi.org/10.1016/j.neuropsychologia.2011.03.023

Clayson, P. E., & Larson, M. J. (2011b). Effects of repetition priming on electrophysiological and behavioral indices of conflict adaptation and cognitive control. *Psychophysiology*, *48*(12), 1621–1630. https://doi.org/10.1111/j.1469-8986.2011.01265.x

Clayson, P. E., & Larson, M. J. (2012). Cognitive performance and electrophysiological indices of cognitive control: A validation study of conflict adaptation. *Psychophysiology*, *49*(5), 627–637. https://doi.org/10.1111/j.1469-8986.2011.01345.x

Clayson, P. E., & Larson, M. J. (2013). Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict n2 event-related potentials. *Psychophysiology*, *50*(12), 1209–1219. https://doi.org/10.1111/psyp.12138

Clayson, P. E., Clawson, A., & Larson, M. J. (2011). Sex differences in electrophysiological indices of conflict monitoring. *Biological Psychology*, *87*(2), 282–289. https://doi.org/10.1016/j.biopsycho.2011.03.011

Crawford, J. R., & Henry, J. D. (2004). The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43(3), 245–265.* https://doi.org/10.1348/0144665031752934

Eaton, W. W., Smith, C., Ybarra, M., Muntaner, C., & Tien, A. (2004). Center for Epidemiologic Studies Depression Scale—revised. *PsycTESTS Dataset*. https://doi.org/10.1037/t29280-000

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ern) and post-error positivity (PE): The role of number of trials. *Psychophysiology*, *47*(6): 1167-1171. https://doi.org/10.1111/j.1469-8986.2010.01022.x

Larson, M. J., & Clayson, P. E. (2010). The relationship between cognitive performance and electrophysiological indices of performance monitoring. *Cognitive, Affective, Behavioral Neuroscience*, *11*(2), 159–171. https://doi.org/10.3758/s13415-010-0018-6

Larson, M. J., Clawson, A., Clayson, P. E., & Baldwin, S. A. (2013). Cognitive conflict adaptation in generalized anxiety disorder. *Biological Psychology*, *94*(2), 408–418. https://doi.org/10.1016/j.biopsycho.2013.08.006

Larson, M. J., Steffen, P. R., & Primosch, M. (2013). The impact of a brief mindfulness meditation intervention on cognitive control and error-related performance monitoring. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00308

Magnus, B. E., & Garnier-Villarreal, M. (2022). A multidimensional zero-inflated graded response model for ordinal symptom data. *Psychological Methods*, *27*(2), 261–279. https://doi.org/10.1037/met0000395

Magnus, B. E., & Liu, Y. (2021). Symptom presence and symptom severity as unique indicators of psychopathology: An application of Multidimensional Zero-inflated and hurdle graded response models. *Educational and Psychological Measurement*, *82*(5), 938–966. https://doi.org/10.1177/00131644211061820

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990a). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495. https://doi.org/10.1016/0005-7967(90)90135-6

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990b). Penn State Worry Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t01760-000

Miller, W. C., Anton, H. A., & Townson, A. F. (2007). Measurement properties of the CESD scale among individuals with spinal cord injury. *Spinal Cord*, *46*(4), 287–292. https://doi.org/10.1038/sj.sc.3102127

Radloff, L. S. (1977). The CES-D scale. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Spielberger, C. D. (1983). State-Trait Anxiety Inventory for adults. *PsycTESTS Dataset*. https://doi.org/10.1037/t06496-000

Spitzer, R. L., Williams, J. B., & Kroenke, K. (1999). Patient health questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t02598-000

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). Generalized anxiety disorder 7. *PsycTESTS Dataset*. https://doi.org/10.1037/t02591-000

StataCorp. 2023. *Stata Statistical Software: Release 18*. College Station, TX: StataCorp LLC.

StataCorp. (2023). *Stata 18 structural equation modeling reference manual.* College Station, TX: Stata Press.

Schwartz, G. E., Davidson, R. J., & Goleman, D. J. (1978). Cognitive-Somatic Anxiety Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t20915-000

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063