2012-05-22

Multidimensional Item Response Theory in Clinical Measurement: A Bifactor Graded-Response Model Analysis of the Outcome-Questionnaire-45.2

Arjan Berkeljon
Brigham Young University - Provo

Multidimensional Item Response Theory in Clinical Measurement:

A Bifactor Graded Response Model Analysis of the

Outcome-Questionnaire-45.2

Arjan Berkeljon

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

Doctor of Philosophy

Scott A. Baldwin, Chair
Gary M. Burlingame
Joseph A. Olsen
Mikle D. South
Richard R. Sudweeks

Department of Psychology

Brigham Young University

August 2012

ABSTRACT

Multidimensional Item Response Theory in Clinical Measurement:
A Bifactor Graded Response Model Analysis of the
Outcome-Questionnaire-45.2

Arjan Berkeljon
Department of Psychology, BYU
Doctor of Philosophy

Bifactor item response theory (IRT) models are presented as a plausible structure for psychological measures with a primary scale and two or more subscales. A bifactor graded response model, appropriate for polytomous categorical data, was fit to two university counseling center datasets ($N = 4,679$ and $N = 4,500$) of Outcome-Questionnaire-45.2 (OQ) psychotherapy intake data. The bifactor model showed superior fit compared to a unidimensional IRT model. IRT item parameters derived from the bifactor model show that items discriminate well on the primary scale. Items on the OQ's subscales maintain some discrimination ability over and above the primary scale. However, reliability estimates for the subscales, controlling for the primary scale, suggest that clinical use should likely proceed with caution. Item difficulty or severity parameters reflected item content well, in that increased probability of endorsement was found at high levels of distress for items tapping severe symptomatology. Increased probability of endorsement was found at lower levels of distress for items tapping milder symptomatology. Analysis of measurement invariance showed that item parameters hold equally across gender for most OQ items. A subset of items was found to have item parameters non-invariant across gender. Implications for research and practice are discussed, and directions for future work given.

Keywords: bifactor model, clinical measurement, graded response model, item response theory, outcome questionnaire

ACKNOWLEDGMENTS

My deepest thanks and appreciation go to Dr. Scott Baldwin, my committee chair. Through this project we explored fairly undiscovered country in the field of clinical measurement. I appreciate your willingness to dive in and stick with me. Your support and encouragement means a lot. I am also indebted to my committee members: Dr. Gary Burlingame for your expertise and friendship, I am glad we are connected; Drs. Joseph Olsen and Richard Sudweeks for your kind guidance and impressive statistical knowledge; and Dr. Mikle South for your refreshing perspective and support. I would also like to thank Dr. Michael Lambert. Sharing in your knowledge of the OQ was both helpful and inspiring. Thank you for your friendship. I would also like to thank Drs. Robert Gibbons, Donald Hedeker, Kim Jong-Bae, and Eisuke Segawa at the University of Chicago's Center for Health Statistics for their generous help with initial specifications of the bifactor model.

I would also like to extend my warm gratitude to my friends in The Netherlands, Provo, and Rochester. Thank you for your support and friendship.

Most importantly my thanks go to my parents, Aad and Marion, and my sister, Marlous. Their support of my education and life in the United States has been extraordinary. I am deeply grateful.

Finally, thank you Kate. Thank you for many an afternoon spent working on our dissertations, thank you for your support, thank you for your presence in my life.

Contents

List of Figures

List of Tables

Multidimensional Item Response Theory in Clinical Measurement: A Bifactor Graded Response Model Analysis of the Outcome-Questionnaire-45.2

> *"…we wish to clearly identify the measurement area as a problem area that presents special challenges. For the vigorous and rigorous researcher who can produce creative innovations in this area, great rewards are likely to follow."*
>
> (Lambert & Garfield, 2004, p. 817)

In its most basic sense psychological measurement entails *measuring*—systematically assigning numbers to represent psychological traits, or characteristics of individuals. Psychological *tests* are the instruments by which such measurement happens. Historically, developments in psychological measurement have been driven by a theory of measurement known as True Score Theory or Classical Test Theory (Allen & Yen, 2001; Lord & Novick, 1968). Consequently, the assumptions underlying the conceptualization of measurement in CTT have guided how clinical measures are developed and evaluated (Embretson & Reise, 2000; Reise & Waller, 2009). Although CTT has led to theoretical and practical benefit in clinical measurement, problematic issues remain.

As its name suggests, CTT defines a theory of tests or measures. Specifically, CTT defines the relationship between an individual's observed score on a test and the measurement error involved in obtaining this score. A foundational assumption of CTT states that *observed scores* consist of *true scores* plus *error scores*. A true score is defined as the mean of the theoretical distribution of observed scores; a person's true score would be obtained by infinite, independent measurements using the same test. The true score thus is a theoretical construct with a fixed, but unknown value. The observed score is an attempt to measure the true score as accurately as possible, minimizing measurement error. This notion of measurement accuracy yields the familiar

notion of test reliability. More precisely, test reliability can be expressed as the ratio of true score over observed score variance. Consequently, a measure is perfectly reliable if all observed score variance reflects true score variance, and error score variance is zero. However, perfect reliability is never obtained in practice because measurement is always assumed to reflect some error, that is, error score variance is always greater than zero.

CTT assumes that this error score variance is constant across individuals for any particular test. Further, CTT assumes that all items on a particular test contribute equally to true score variance. This assumption, known as parallel measurement, is the foundation of test development from a CTT perspective. An important consequence of this assumption is that item responses may be summed to reflect total test performance. For example, imagine a hypothetical measure of the construct depression with 10 items, each rated on a 4-point Likert scale, with anchors "Never," "Rarely," "Sometimes," "Always," respectively. Imagine a patient with a total observed score of 25 obtained by rating five items at four and five items at one. Suppose a second patient also had a total score of 25, but obtained that score by rating seven items at three, two at two, and one at one. Here the question arises to what extent the same total score reflects similar distress given that the first patient's scores reflect a dichotomous presentation of symptoms and the second patient's scores reflect a more uniform presentation. One might infer that patients' level of distress is similar because their scores are equal. However, given their particular responses to the items, the quality and content of their distress may be distinct.

Next, suppose that two new patients are administered the measure from the previous example at intake and after 10 sessions of therapy. The first obtained an initial score of 40 and, following treatment, a score of 30. The second obtained a initial score of 20 and, following treatment, a score of 10. Did the

patients experience the same amount of change? Obviously in a numerical sense it is equal, 10 points. However, the same score change may have come about in different ways. For example, a change in rating on three items from 4 to 1, and a change in rating on one item from 2 to 1 equals a change of 10 points. However, a change in rating from 4 to 3 on 10 items also equals a change of 10 points. Thus, although the difference score suggests similar change, the manner in which such change occurs may be different across patients. Also, because both patients start therapy at different ends of the scale, one high, and one low, the question arises whether equal numerical change may reflect a different quality of change considering the place on the scale. That is, for a high initial level of distress, a 10 point reduction might reflect significant relief, for example from feeling depressed "Always" to feeling depressed "Sometimes." By contrast, on the lower end of the scale the difference between feeling depressed "Sometimes" and feeling depressed "Rarely," may provide some, but less significant relief.

In CTT, although these problems are acknowledged, they are usually ignored in practice. For example, although confirmatory factor analyses of clinical measures are common, the findings that different items do indeed contribute differentially to the measure as a whole do not necessarily affect how a measure is scored. Because CTT makes no discrimination on an item level in terms of each items' differential contribution (i.e. the parallel assumption), the current practice of using sum scores seems warranted. In fact, obtaining a global level or generalized index of complex information serves a useful deductive purpose in clinical decision making and reducing information is a legitimate and necessary component of effective decision making. However, when items do not meet the parallel assumption, an alternative measurement model would be useful. Such a model would have to incorporate differences on an item level and yet not sacrifice in effective decision making.

A measurement model that can satisfy this requirement is Item Response Theory (IRT; Embretson & Reise, 2000). Although IRT has been used in measurement development at least since the late 1950s, its adoption into mainstream clinical measurement has been slow (Embretson & Reise, 2000; Reise & Waller, 2009). A key difference between CTT and IRT is that whereas in CTT test performance is defined in terms of a person's true score, in IRT test performance is defined in terms of an unobserved ability or trait (Allen & Yen, 2001; Embretson & Reise, 2000). For example, for the depression measure above CTT assumes that a person's item responses depend on their true score for that measure. IRT, on the other hand, assumes that a person's responses depend on their standing on the latent trait of depression. IRT provides estimates of latent traits that may be used to interpret person and item performance. Because such interpretations are trait-based, rather than score-based, possible interpretative complications such as the ones illustrated above can be avoided.

Item Response Theory

Because of the proposed direct relationship between latent trait and test performance in IRT, this method allows for a more direct investigation of the relationship between latent trait levels and person and item performance than under CTT. Unlike CTT, where the observed score is a linear function of the true score plus error, in IRT the observed score is not a simple linear function of the latent trait. Instead, a person's response to any given item on a test is a function of their trait level and certain parameters of the item. In other words, IRT methods allow one to investigate of how items differentially contribute to a measure.

IRT methods yield two important parameters for each item on a measure, *difficulty* and *discrimination*.[1] Item difficulty is scaled on a common metric with the trait level of persons; item discrimination is a multiplier of the difference between item difficulty and a person's trait level. Thus, information about item parameters provides a means to select items that give the most accurate estimate of a person's standing on the latent trait. By definition, when trait level equals item difficulty, the probability of a person with that trait level endorsing an item with that difficulty is .5. Consequently, for persons with a certain trait level, the probability of endorsing an item of higher difficulty is less than .5, and endorsing an item of lower difficulty is greater than .5. For items with a certain difficulty, the probability of a person with higher trait level endorsing the item is greater than .5, and a person with lower trait level endorsing the item is less than .5. An item's discrimination is defined as the degree to which an item discriminates between different trait levels. In other words, for a highly discriminating item, a person who endorses the item is likely to have a trait level greater than the difficulty of that item. Conversely, a person who does not endorse the item is likely to have a trait level less than the difficulty of the item. An item that does not discriminate well provides much less information about persons' trait standing relative to the item's difficulty.

The information obtained via IRT methods can be used to improve the accuracy of measurement. In particular, IRT methods may be useful to clinical measurement development in three ways. First, Reise and Waller (2009) discuss the benefits of IRT trait estimates as opposed to using CTT-inspired sum and change scores. The authors note that although IRT trait estimates may correlate highly with CTT estimates, the information obtained from IRT analysis makes an IRT approach beneficial in spite of high correlations. A correlation is insensitive

[1]Other item parameters can be obtained using IRT methods, but difficulty and discrimination are most common.

to specific values of estimates in the sense that a similar pattern and spread yields a high correlation even though relative position on the trait scales suggests a different interpretation for different scores.

Second, Doucette and Wolf (2009) discuss the importance of analyzing item parameters from an IRT perspective. They mention three relevant analysis strategies. First, an analysis of item difficulty may be done to ensure sufficient coverage of the latent trait. Items may be located unevenly on the trait range (i.e. lumped) or lack of coverage may exist for a certain range of the trait. Second, an analysis of difficulty and discrimination parameters may be done to assess for unexpected item behavior. That is, items' actual difficulty or discrimination may not represent their intended difficulty or discrimination. Thus, an item intended to be difficult may not actually be endorsed with a higher probability by persons located higher on the latent trait. Or, an item intended to discriminate persons located high on the latent trait from persons located lower on the latent trait may not actually distinguish such persons. For example, an item such as "In your lifetime, have you ever felt depressed?" will likely be endorsed with high probability by most persons regardless of their current level of depression; the item is not difficult. An item such as "Do you currently feel depressed?" may be more difficult than the previous item in the sense that persons who feel depressed are more likely to endorse it than persons who do not feel depressed. However, this item likely does not clearly distinguish persons who are more depressed from those that are less depressed. Third, the authors discuss using polytomous models to assess the adequacy of response categories. That is, only a limited number of an item's response categories may show expected properties. For example, an item rated on a four-point scale may display infrequent use of the outer response categories. This may reflect adjacent response category

boundaries on the latent trait continuum that overlap or response category boundaries that occur at a lower or higher level on the latent trait continuum.

Third, Reise and Waller (2009) discuss Differential Item Functioning (DIF; referred to as measurement invariance in the CFA literature). In most research distinct groups of responders are often present (e.g. men and women, depressed and nondepressed etc.). The purpose of a DIF analysis is to ascertain to what degree item parameters are different for the same trait level for persons in distinct groups (Embretson & Reise, 2000). DIF analyses are clinically useful to justify use of uniform tests or items across groups (in the absence of DIF) or use of distinct tests or items across groups (in the presence of DIF).

IRT Models for Binary Responses

In the simple case of a dichotomous item, an IRT model describes the relationship between a person's latent trait standing and the probability of a correct response. Most standard IRT models share two assumptions: (a) unidimensionality and (b) local independence (Embretson & Reise, 2000). Unidimensionality refers to the assumption that a single trait captures the relationships between the items; this is referred to as a unidimensional latent trait. Local independence refers to the relationships between the items or persons. This assumption is satisfied if the relationship among items is fully accounted for by the item and person parameters specified in the model.

Given a unidimensional latent trait, regressing the item score onto the latent trait, θ, yields an item characteristic curve (ICC), which describes the relationship between trait level and probability of a certain response. In the traditional normal ogive model the ICC takes the shape of a normal cumulative

distribution function (CDF):

$$P_i(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i(\theta)} \exp(-y^2/2)\, dy, \tag{1}$$

where

$$z_i(\theta) = a_i(\theta - b_i), \tag{2}$$

which defines the probability of a correct score, P_i, on a dichotomous item, i, as a function of the cumulative proportion of cases below a standard score, z_i, defined on a standard scale θ with a mean of 0 and standard deviation of 1, containing item parameters a_i and b_i, representing item discrimination and difficulty as defined above, respectively. Because the model incorporates both a_i and b_i parameters, this model is often referred to as the traditional two parameter normal ogive model. Note that the relationship between trait level and item response in IRT models has also been defined in terms of the logistic rather than normal ogive cumulative distribution function. The normal ogive and logistic functions nearly coincide and thus yield similar item characteristic curves given identical item properties. The logistic analog to the two parameter normal ogive model is defined as follows:

$$P_i(\theta) = \frac{\exp(Da_i(\theta - b_i))}{1 + \exp(Da_i(\theta - b_i))}, \tag{3}$$

which defines the probability of a correct score, P_i, on a dichotomous item, i, as a logistic function of item parameters and standard scale θ. To correct for the small scaling difference between the normal ogive and logistic models, D is included as a unit scaling factor set at $D = 1.7$. In the logistic model $a_i(\theta - b_i)$ is referred to as the logistic deviate with a_i and b_i as defined above.

A graphical representation of three different item characteristic curves is shown in Figure 1. A theoretical trait level is displayed on the x-axis. On the y-axis the corresponding response probability is given. The ICCs for three items with different b_i parameters are shown. The different b_i parameters affect the location of the inflection point of the curve in relation to the x-axis. That is, as the value of b_i increases, the curves inflection point shifts to the right on the θ scale. Thus, as b_i increases, the value of θ necessary to yield an equal response probability also increases. This illustrates that b_i represents the difficulty of an item given a certain trait level. Note that when trait level equals the item difficulty (at the curve's inflection point), the probability of endorsing the item is precisely .5. For an item of average difficulty, item difficulty equals the average trait level, 0 on the standard θ scale.

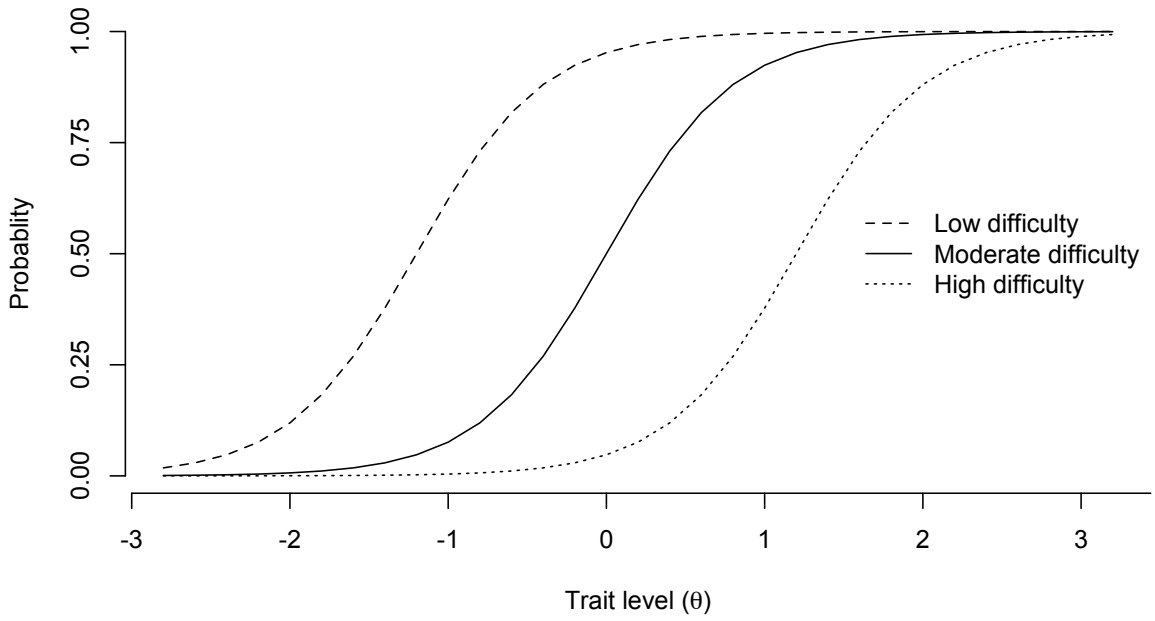


Figure 1. Item characteristic curves for items of varying difficulty. Graphic shows item response probability as a function of latent trait level for items of low, moderate, and high difficulty. The latent trait scale has a mean of zero and a standard deviation of one.

A set of ICCs with different a_i, but equal b_i parameters is shown in Figure 2. As can be seen the a_i parameters affect the slope at the inflection point of the curve (at b_i, the item's difficulty). That is, for larger values of a_i the slope of the curve at its inflection point increases; the curve is steeper. Thus, as a_i increases, relatively smaller differences in θ yield relatively larger differences in response probability. Note that the influence of a_i on response probabilities is strongest at an item's b_i and decreases as θ and b_i become more distinct. This illustrates that a_i represents the discriminating ability of an item of particular difficulty, b_i, given a certain trait level, θ. An item is most discriminating for persons whose trait level matches the item's difficulty. Note that as a_i approaches zero, an item's discriminating ability diminishes. Thus, in all cases desirable values for a_i are positive real integers.

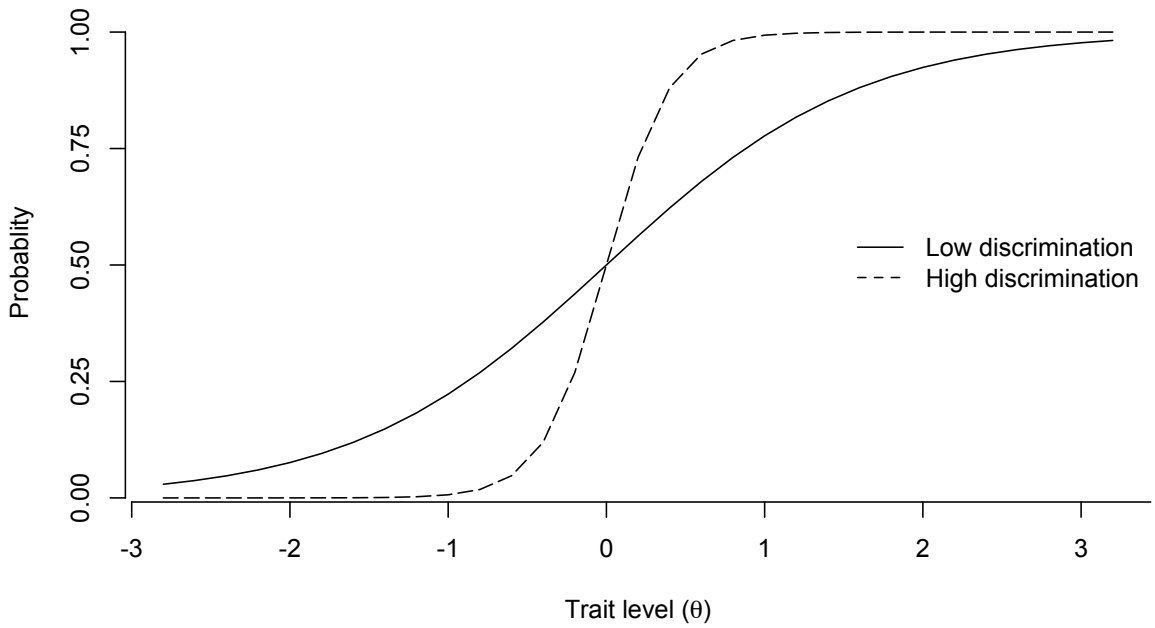


Figure 2. Item characteristic curves for items of varying discrimination. Graphic shows item response probability as a function of latent trait level for items of low and high discriminating ability. The latent trait scale has a mean of zero and a standard deviation of one.

More concretely, consider again the hypothetical measure of depression mentioned previously. We may interpret Figure 1 to reflect the level of depression on the x-axis and the probability of endorsing three of the 10 items on the y-axis. The curves represent, for each of the three items shown, the relationship between level of depression and the probability of endorsing the item in question. The curve farthest to the left represents an item of low difficulty or severity. That is, the probability of endorsing this item is high for relatively low levels of the latent trait, depression. In other words, this item reflects an aspect of depression that is present for relatively low levels of depression, for example depressed mood. By contrast, the item farthest to the right represents an item of high difficulty or severity. That is, the probability of endorsing this item is low for relatively low levels of the latent trait and only increases as the level of the latent trait increases. In other words, this item reflects an aspect of depression that is present only for relatively high levels of depression, for example blunted affect.

Similarly, Figure 2 may be interpreted as showing items of low and high discriminating ability concerning level of depression, respectively. One might conceive of an item with the content "I feel down", and imagine a low discriminating version as having but two response options, "Never" or "Always." Given these options and the frequent incidence of "feeling down" at even low levels of depression (i.e. the item is not difficult), it seems likely that the item will not clearly distinguish lower trait levels from higher trait levels. However, given a more elaborate response set such as "Never," "Rarely," "Sometimes," and "Always" it seems likely the item might be able to distinguish lower trait levels from higher trait levels more accurately than the dichotomous version of the same item.

Before describing IRT models for polytomous responses, a class of common IRT models based on the logistic function deserves mention. Historically, these models have become more or less synonymous with IRT, and although they do not exclusively define the field much of IRT's advances have come from work in this area. In a class of models known as Rasch models (Rasch, 1960) the probability of a correct answer to a dichotomously scored item given a certain trait level is represented as follows:

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}, \tag{4}$$

with terms defined as above. Note that in the traditional Rasch model given above, as well as derivations thereof, no item discrimination parameter, a_i, is modeled.

In a more generalized form, the Rasch model may be expressed as follows:

$$P_i(\theta) = \frac{\exp(a(\theta - b_i))}{1 + \exp(a(\theta - b_i))}, \tag{5}$$

where a equals 1. When a is not constrained to be 1 but is estimated as a common parameter across all items, Equation 5 defines a class of models known as one parameter logistic (1PL) models.

If the common item discrimination parameter estimated in 1PL models is freely estimated separately for each item, Equation (5) yields Equation (3) defined above. Because in such models two parameters are estimated uniquely for each item, namely item discrimination and item difficulty, these models are often referred to as two parameter logistic (2PL) models. The Rasch model can thus be seen as a special case of the 1PL model, which in turn is a special case of the 2PL model.

IRT Models for Polytomous Responses

In the polytomous case the relationship between trait level and item response is complicated by the availability of three or more response options. IRT models for ordered categorical responses have addressed this complication by treating a k-category response item as $k-1$ hypothetical dichotomous subitems. Thissen and Steinberg (1986) distinguish two broad classes of models: divide-by-total models and difference models.

In divide-by-total models such as the Partial Credit Model (PCM; Masters, 1982) and the Rating Scale Model (RSM; Andrich, 1978b, 1978a), the probability of a response in a certain category is directly estimated for each category. By contrast, in difference models such as the graded response model (GRM; Samejima, 1969, 1996), first the probability of a response in or above category j is calculated as a function of item parameters and a given trait level. Subsequently the probability of a response in category j is calculated as the difference between the probability of a response in or above that category and the probability of a response in or above the adjacent category. Thus for a k-category item the probability of responding in category j is defined as follows:

$$P_{ij}(\theta) = P^*_{ij}(\theta) - P^*_{i,j+1}(\theta), \tag{6}$$

with $P^*_{i0}(\theta) = 1$ and $P^*_{ik}(\theta) = 0$.

The normal ogive model in the polytomous case again assumes a normal CDF for the probability of a response in or above category j of item i, $P^*_{ij}(\theta)$:

$$P^*_{ij}(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{ij}(\theta)} \exp(-y^2/2)\, dy, \tag{7}$$

where

$$z_{ij}(\theta) = a_i(\theta - b_{ij}), \tag{8}$$

with item discrimination, a_i, and threshold parameter, b_{ij}, which denotes the threshold on the θ scale between adjacent categories j and $j+1$ on item i.

$P^*_{ij}(\theta)$ is defined as follows for the logistic model:

$$P^*_{ij}(\theta) = \frac{\exp(Da_i(\theta - b_{ij}))}{1 + \exp(Da_i(\theta - b_{ij}))}, \tag{9}$$

with a_i and b_{ij} defined as above. The normal ogive and logistic functions define operating characteristic curves for each $P^*_{ij}(\theta)$. Through Equation 6 these yield category response curves (CRCs) which represent the probability of a response in a given category conditional on trait level, $P_{ij}(\theta)$.

An example for a three-category item CRC is shown in Figure 3. A theoretical trait level is displayed on the x-axis. On the y-axis the corresponding response probability is given. The three different curves show the relationship between trait level and probability of response in a certain category for each of the three categories. For low levels of the latent trait, the probability of responding in category 1 is high, whereas the probability of responding in either category 2 or 3 is low. As trait level increases the probability of responding in category 1 decreases and the probability of responding in category 2 or 3 increases. As trait level increases further, the probability of responding in category 2 reaches a maximum and then decreases. The probability of responding in category 3 subsequently increases. The CRC shown thus reflects an item where persons low on the latent trait have a high probability of responding in category 1, persons with moderate levels of the latent trait have a high probability of responding in category 2, and persons high on the latent trait have a high probability of responding in category 3. Note that, similarly to

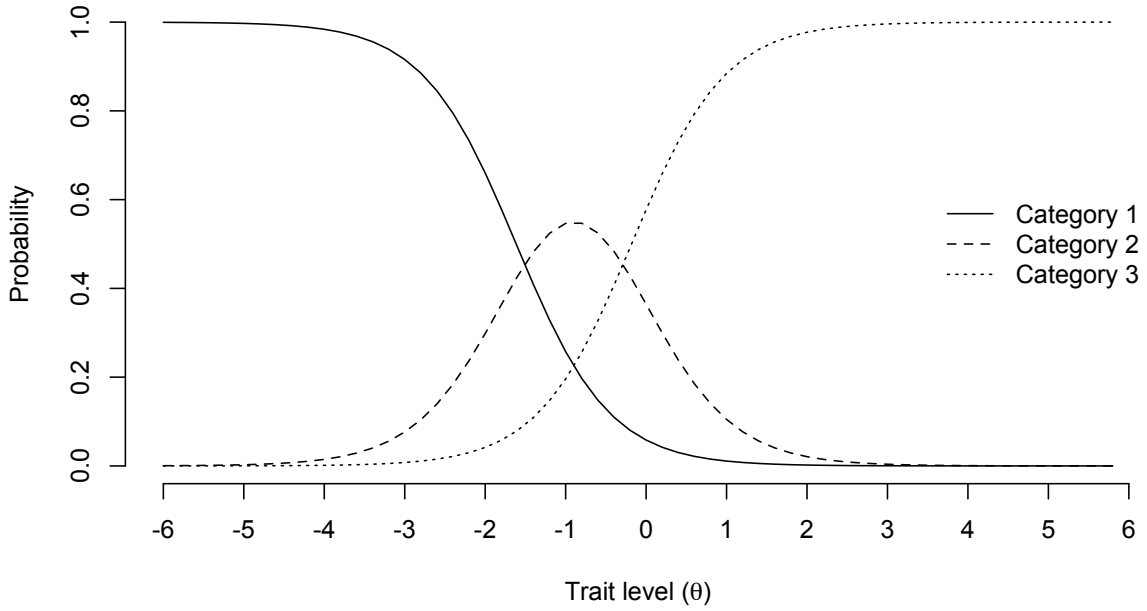


Figure 3. Category response curve for a three-category item. Graphic shows category response probability as a function of latent trait level for a three-category item of particular difficulty and discrimination. The latent trait scale has a mean of zero and a standard deviation of one.

binary IRT models, item difficulty is expressed as location of the CRCs with respect to the latent trait scale. However, item discrimination is expressed as the peakedness or slope of the curves; a higher discrimination parameter is associated with more peaked or steeper curves.

For example, on a three-category response item asking about the daily presence of depressed mood, the categories represented in Figure 3 might denote "Rarely," "Sometimes," and "Always." Thus, for low levels of the latent trait or low depression, the depressed mood rarely occurs daily. For increasing levels of depression the occurrence of daily depressed mood increases as well from often to always.

Multidimensional IRT Methods

The IRT models and applications discussed thus far rely on the assumption of unidimensionality. However, most measures, and certainly most psychological measures, are conceptually broad in that they have content heterogeneous indicators. For example, a hypothetical depression measure may have items for both mood and physiological concerns. The more content heterogeneous indicators a measure contains, the more appropriate a multidimensional representation becomes. Because of the assumption of unidimensionality in most standard IRT models, content heterogeneity has direct implications for IRT model selection and standard IRT models may not be appropriate (Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007).

Methods have been developed that allow for multidimensional IRT models in which item response are modeled as person and item properties reflecting two or more dimensions or latent traits (Embretson & Reise, 2000; Reckase, 2009). Such methods apply where two (or more) distinct dimensions are modeled. However, many psychological measures have indicators that reflect a hierarchical structure. For example, although the hypothetical depression measure mentioned previously may have items for both mood and physiological concerns, these dimensions are not independent of the broader dimension of depression. Rather, they may be considered subdimensions. In the confirmatory factor analysis literature hierarchical factor models are commonplace in modeling multidimensional constructs (Brown, 2006). In such models, layers of factors represent the heterogeneity of the broader content domain. The observed variables in the model load on one or more of these factors. One such hierarchical model is the bifactor model. In the bifactor model proportionality constraints are such that items relate to one general dimension or factor and one of two or more subdimensions or group factors. Each item is constrained to load

on the general factor and one of the group factors. This contrary to a second-order model, which is nested within the bifactor model, where the bifactor general factor resembles the second-order factor, and the bifactor group factors resemble disturbances of first-order factors. Yung, Thissen, and Mcleod (1999) describe the difference between these two factor models as the difference between the “breadth" versus the “superordination" conception, respectively. The properties of the bifactor model, specifically the restriction of nonzero loadings on the general factor and only one group factor, make it computationally attractive (Gibbons & Hedeker, 1992).

In addition to computational advantages, Chen, West, and Sousa (2006) name several advantages of the bifactor over a second-order model, three of which are relevant to the present study. First, the bifactor specification allows for inferences regarding the predictive value of the group factors over and above the general factor. Second, the bifactor model allows for a direct examination of the strength of the relationship between items and group factors. Third, measurement invariance between different groups (e.g. males vs. females), can be directly tested for the general as well as group factors. This cannot be done in a second-order model because domain specific factors are represented by disturbances. Such measurement invariance testing is analogous to the process of DIF testing in IRT.

An early bifactor IRT implementation for dichotomous response scale data is described by Gibbons and Hedeker (1992). An extension applied to categorical response data is described by Gibbons et al. (2007). They define a bifactor implementation of the GRM. For computational purposes it is convenient to define the GRM using the item intercepts, $c_{ij} = -a_i b_{ij}$ (Bock, Gibbons, & Muraki, 1988; Gibbons et al., 2007). This yields the probability of a response in or above category j of item i defined as in Equation 7, with the

following in the unidimensional case (equivalent to Equation 8):

$$z_{ij}(\theta) = c_{ij} + a_j\theta. \tag{10}$$

In the bifactor case the probability of a response in or above category j of item i is also defined as in Equation 7, but with

$$z_{ij}(\boldsymbol{\theta}) = c_i + \sum_{k=1}^{s} a_{jk}(\theta_k), \tag{11}$$

and k equal to the number of factors in the bifactor model and $k = 1$ for the primary factor and $k = 2, \ldots, s$ for the group factors. Only one of $k = 2, \ldots, s$ values of a_{jk} is nonzero in addition to a_{j1}. This imposes the bifactor model constraint that each item loads on the primary factor and on only one of the group factors.

The item intercepts span all factors in the bifactor model and are not directly interpretable in terms of any one factor. This has particular consequences for the item threshold parameters (i.e. the difficulty or location parameter in IRT terms). In the bifactor model these invariant location parameters do not exist with respect to one latent variable in the model. Rather, the thresholds pertain to an additive composite of all latent variables in the model (Bock et al., 1988).[2] Thus, contrary to unidimensional IRT models, where item thresholds are defined relative to the unidimensional latent trait, in bifactor models item thresholds are defined relative to an additive composite of all latent variables in the model. Specific to the GRM, whereas in the unidimensional case item thresholds reflect the location on the latent trait where there is a probability of .50 of selecting a response in a particular category or higher, in the bifactor case thresholds reflect this same probability, but on a scale defined by the added

[2]R. Gibbons, personal communication, February 8, 2011.

composite of latent variables in the model. Consequently, whereas in a unidimensional model item slopes and intercepts are defined uniquely for the single latent variable in the model, in the bifactor case only item slopes are uniquely defined for each of the latent variables in the model, however, item thresholds are not.

IRT in Clinical Measurement

IRT methods have been slow to appear in clinical research and useful findings are sparse (Reise & Waller, 2009). Multidimensional, and in particular bifactor IRT approaches, are even less common in spite of their potential utility in clinical measurement. Gibbons and Hedeker (1992) report a full-information bifactor analysis of the Hamilton Depression Rating Scale (HDRS), noting however, that an unrestricted five-factor model had significantly better fit than a five-dimensional bifactor model. This suggests that the four subdomains of the HDRS model are likely not fully independent as a strictly orthogonal bifactor structure presupposes. Extending the full-information bifactor approach to polytomous items Gibbons et al. report an analysis of a quality of life measure, the Quality of Life Interview for the Chronically Mentally Ill (Lehman, 1988). For this measure, with seven proposed subscales and one global item, a bifactor graded response model showed improved fit over a unidimensional graded response model. Using a similar method, Reise et al. (2007) show that after controlling for a general dimension, subscales of the Consumer Assessment of Healthcare Providers and Systems survey (a healthcare satisfaction survey) provide little measurement precision. Finally, Immekus and Imbrie (2008) test the dimensionality of an adapted version of the State Metacognitive Inventory (O'Neill & Abedi, 1996), a self-report measure of self-regulatory processes in

students. They report improved fit for a bifactor GRM over a unidimensional GRM.

To extend the understanding and application of bifactor IRT methods in clinical measurement the present study gives an implementation of a bifactor IRT model for a well-known psychotherapy measure, the Outcome Questionnaire-45.2 (OQ; Lambert et al., 1996; Lambert et al., 2004). The OQ is a 45-item, self-report questionnaire that measures patients' progress during the course of psychotherapy treatment. Progress is measured as an overall score on 45 five-point (0=Never, 1=Rarely, 2=Sometimes, 3=Frequently, and 4=Almost Always) Likert scale items as well as a score on three subscales: (a) Symptom Distress (SD); (b) Interpersonal Relations (IR); and (c) Social Role (SR). The overall score (0–180) provides a measure of overall psychological disturbance. The SD score (0–100) signifies subjective symptom distress; the IR score (0–44) indicates satisfaction/problems with interpersonal relationships; the SR score (0–36) indicates patients' dissatisfaction, experienced conflict, or feelings of inadequacy in tasks related to their employment, family life, and leisure life. The sparsity of IRT research in clinical measurement in general is reflected in the literature available for the OQ. Two IRT studies of the OQ exist in the literature.

Pastor and Beretvas (2006) report a longitudinal Rasch study of the OQ. They address the multidimensionality of the OQ by separately fitting three unidimensional IRT models to three subscales derived using an exploratory factor analysis. Analysis of item difficulty parameters across three points shows time non-invariance for four of the 18 OQ items comprising their derived subscales. The remaining 14 items were found to be invariant across time. Although important given the OQ's use as a measure of patient change across time, Pastor and Beretvas' findings are limited because derived subscales cover only a subset of the 45 total items on the OQ. In addition, subscales were fit

separately, rather than in a model framework that captures the primary-subscale structure of the OQ.

Doucette and Wolf (2009) report an IRT analysis of the full response scale of the Life Satisfaction Questionnaire (LSQ; Lambert et al., 2003), a 30-item variant of the OQ. Both a Rasch model and 2PL model were fit, with the 2PL model showing modestly better fit over the Rasch model. Results showed that items do not provide adequate coverage for mild and high distress. The authors report a principal components analysis (PCA) of model residuals to support their assertion that the LSQ is sufficiently unidimensional to warrant fitting a standard IRT model. The first extracted component explains 53% of the variance in the data, the second 4%. Noting that three substance abuse differ notably from other items in terms of IRT parameters, the authors argue that the LSQ consists of two subscales, one with the three substance abuse items, the other with the 27 remaining items. They further report PCA results to support their claim that certain items' residuals are correlated and may be eliminated because of redundant content. The authors' approach is appropriate given that the LSQ has a proposed unidimensional structure, but it is not an appropriate general strategy for dealing with unidimensionality given the structure of many clinical measures. As argued above, because many clinical measures are structurally multidimensional, unidimensionality can often not be assumed and multidimensionality has to be explicitly addressed. Moreover, PCA, which is primarily a data reduction method, is limited as a method to analyze measurement structure compared to methods such as confirmatory factor analysis (CFA; Brown, 2006). In fact, in a recent CFA of OQ data, Bludworth, Tracey, and Glidden-Tracey (2010) report superior fit of a bifactor model over a single (unidimensional) model of the OQ. This further establishes the utility of the bifactor method in the present study.

Primary Aims

The present study proposes a bifactor IRT method to account for the multidimensionality of the OQ within an IRT framework. The proposed method serves three broad aims: (a) study the OQ's dimensional structure; (b) study the OQ's primary scale, subscale, and item behavior; and (c) study the stability of the OQ's scale and item properties across a clinically relevant group (i.e. measurement invariance with respect to gender), all within an IRT framework. For each of these aims, the following specific objectives are defined:

1. **Dimensionality:**
   (a) Address the dimensional structure of the measure, by fitting a bifactor graded response model (GRM; Gibbons et al., 2007; Samejima, 1969, 1996) to OQ data and compare model fit to a traditional unidimensional graded response model.
   (b) Test for model parsimony by comparing an unconstrained bifactor GRM to a constrained, tau-equivalent version of the model.
2. **Item behavior:**
   (a) Evaluate the OQ in terms of its item behavior (expected/unexpected difficulty and discriminatory properties).
   (b) Assess item behavior for both primary and subscale factors and assess utility of the subscales, if any, over and above the primary dimension.
3. **Measurement invariance:**
   (a) Assess measurement invariance of item behavior across gender. Specifically, measurement invariance across gender will be assessed separately for primary and subscale factors.

Methods

Participants and Procedures

Two sources of outcome data were used, one from The University of Texas at Austin Research Consortium consisting of aggregate data from 70 United States college and university counseling centers (cohort I) and one from Brigham Young University's Counseling Center (cohort II). The former included observations of subjects who presented for discrete episodes of individual therapy at various university counseling centers. This yields a total dataset of 4,679 patients seen by 488 therapists. Average attendance was 7.58 sessions ($SD = 4.25$, range = 4–38). Clients were predominantly Caucasian and female (73% and 66%, respectively) with an age range from 16–61 ($M = 23.4$, $SD = 5.77$). The average intake OQ score was 70.77 ($SD = 24.90$, range = 4–153). The latter included a subset of observations from a larger data pool. Subjects were included who presented for individual therapy, within their first episode of therapy (defined as no more than 90 days between sessions), and attended at least three sessions, but no more than 40[3]. A random subset of these data gives 4,500 clients seen by 181 therapists. Average attendance was 7.87 sessions ($SD = 6.15$, range = 3–40). Clients were predominantly Caucasian and female (87% and 62%, respectively) with an age range from 17–60 ($M = 22.66$, $SD = 3.99$). The average intake OQ score was 69.31 ($SD = 22.87$, range = 6–162).

Dimensionality and item behavior analyses proceeded on two random subsets of cohort I (cohort Ia, $N = 2,297$ and cohort Ib, $N = 2,382$) separately and the full sample of cohort II ($N = 4,500$). To balance gender representation in measurement invariance analyses subsamples of these cohorts were used. In addition, because measurement invariance testing involves many model

[3]This criterion excludes outlier observations with number of sessions below and above the bottom and top fifth percentiles, respectively.

comparisons, reducing the full data decreases computational time. Measurement invariance analyses proceeded on random subsets of cohort Ia ($n_{\text{men}} = 512$, $n_{\text{women}} = 559$), cohort Ib ($n_{\text{men}} = 560$, $n_{\text{women}} = 530$), and cohort II ($n_{\text{men}} = 1710$, $n_{\text{women}} = 1661$).

Measure

Psychotherapy outcome data was gathered using the Outcome Questionnaire-45.2 (OQ; Lambert et al., 1996; Lambert et al., 2004). The OQ shows high test-retest reliability (.84 for the Total score) and high internal consistency (Cronbach's α for the Total score is .93 for a sample of college students and a sample patients). Internal consistencies for the subscales ranged from .70–.92 in the nonpatient sample and from .71–.91 in the patient sample. The OQ also shows high concurrent validity in that correlations between the OQ Total score and instruments such as the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the State-Trait Anxiety Inventory (Spielberger, 1983), and Social Adjustment Scale (Wiessman & Bothwell, 1976) are moderate to high.

Statistical Analyses

Bifactor and unidimensional graded response models were fit using Mplus software, version 6 (Muthen & Muthen, 2010). Measurement invariance was assessed with a series of constrained bifactor GRMs in Mplus. All other analyses were performed R using version 2.13 (R Development Core Team, 2011). To address the aims of the present study I performed the following analyses:

- **Dimensionality:**

    1. A bifactor graded response model (*BFGRM*; see Figure 4) was fit to OQ data using robust maximum-likelihood (MLR) estimation with a

probit link (Muthen & Muthen, 2010). Robust maximum-likelihood estimation, like standard maximum-likelihood estimation, is a full-information estimation approach and yields optimal estimates given sufficient data. In addition, robust maximum-likelihood is appropriate for use with non-normal data. Thus, MLR is a desirable estimator for the naturalistic clinical data used in the present study.

2. A unidimensional IRT implementation of the graded response model (*UGRM*; see Figure 5) was fit to OQ data using robust maximum-likelihood estimation with a probut link. This analysis thus provides a basis for comparing the bifactor graded response model to a unidimensional graded response model.
3. A constrained version of the bifactor graded response model was fit to OQ data using robust maximum-likelihood estimation with a probit link:
   (a) Tau-equivalent bifactor GRM with factor loadings constrained to be equal within factors (*BFGRMte*; see Figure 6). That is, factor loadings on the primary factor are constrained to be equal and factor loadings within each of the subscale factors are constrained to be equal, respectively.

A Satorra-Bentler scaled χ^2-difference or log-likelihood ratio test, which is appropriate for robust maximum-likelihood estimation, was used to compare nested models (Muthen & Muthen, 2010; Satorra, 2000).

- **Item behavior:**

4. Item properties (i.e. difficulty and discrimination) are reported as estimated by Mplus. Subscale utility was quantified by calculating subscale reliability using the following formula (Reise et al., 2007, p.

26):

$$\text{Reliability} = (N - \sum \text{SE}^2)/N, \tag{12}$$

where SE denotes the standard errors of the estimated factor scores. This reliability estimate is a measure of the degree of precision with which individuals can be assessed on a subscale factor controlling for the primary factor.

- **Measurement invariance:**

5. Measurement invariance (analogous to differential item functioning or DIF in an IRT framework) was studied with respect to gender with men as reference group and women as focal group.

Historically, studies of measurement invariance in a factor analysis framework and an IRT framework have been separate, however, the methodologies frequently overlap (Embretson & Reise, 2000; Holland & Wainer, 1993). For the purposes of this paper measurement invariance was tested in a nested model approach as is common in the measurement invariance literature. In their extensive review of the measurement invariance literature, Vandenberg and Lance (2000) describe the following recommended practices for measurement invariance testing:

(a) Test of configural invariance

(b) Test of metric or weak factorial invariance (invariance of factor loadings across groups)

(c) Test of scalar or strong invariance (invariance of intercepts or thresholds across groups)

(d) Tests of partial invariance (of individual items, provided that full metric or scalar invariance do not hold)

These tests are sequential. That is, a test for metric invariance takes into account results of a test of configural invarance, a test of scalar invariance proceeds takes into account results of a test of metric invariance, and so on (Vandenberg & Lance, 2000).

Invariance testing was evaluated via nested model comparisons using a weighted least squares estimator with mean and variance adjustment (WLSMV) under a theta parameterization.[4]. Like MLR, WLSMV is robust against non-normal data. Unlike MLR, which is a full-information estimator, WLSMV is a limited-information estimator. Thus, compared to MLR WLSMV estimates are suboptimal. However, MLR as currently implemented in Mplus (version 6.1 at time of writing) introduces several complexities into measurement invariance testing, two of which are relevant here: (a) because of the numerical integration required, MLR estimation is substantially slower (by a factor of approximately eight for the current models and data) than WLSMV estimation; (b) modification indices are not offered for MLR, but are offered for WLSMV. Addressing these complexities is beyond the scope of the current study and therefore WLSMV is used an estimator for measurement invariance testing in what follows.

Model comparisons are done using the Mplus DIFFTEST procedure for nested model comparisons under WLSMV. This procedure yields appropriately scaled χ^2-difference tests for nested models. Partial invariance testing is done using modification indices to suggest removal of equality constraints between groups that contribute most to local model misfit. Modification indices give the approximate overall model χ^2 decrease after removal of an equality constraint (Brown, 2006). Equality constraints were removed iteratively based on modification indices suggesting a significant χ^2 decrease at $p = .05$.

[4] Under a theta parameterization residual variances of latent response variables are parameters of the model. This parameterization is preferred for multiple group analyses, particularly of categorical data, because it allows three sources of group differences to be distinguished, factor loadings, factor variances, and residual variances (Muthen & Asparouhov, 2002)

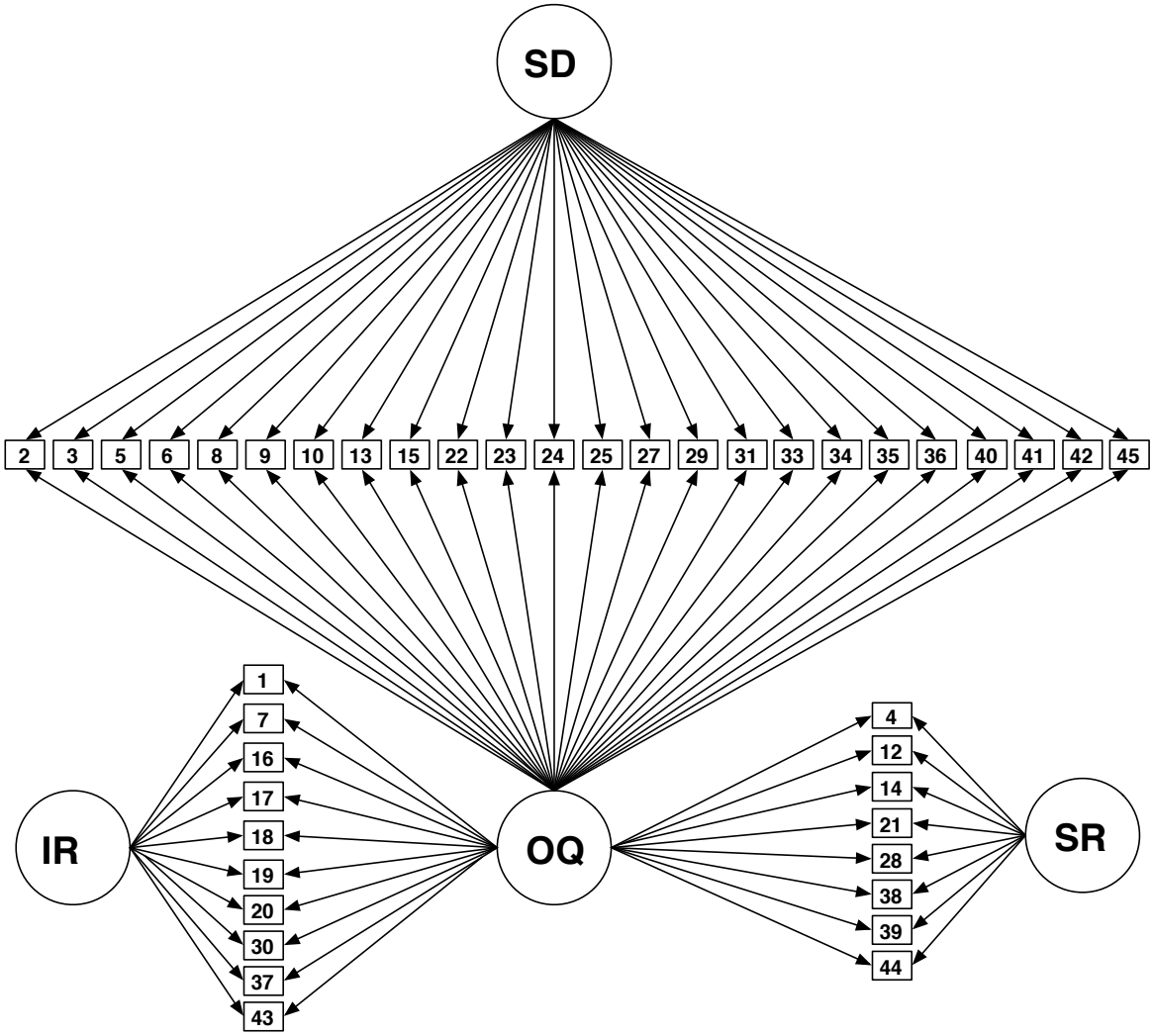


Figure 4. *BFGRM*: full bifactor graded response model. Graphic shows an unrestricted bifactor graded response model of the OQ. Primary and group factors are represented as circles and correspond to the OQ's general dimension and subscales. OQ items are represented as squares. Item factor loadings are represented as arrows and are all estimated. Factor means are fixed at zero and factor variances at one. Factors are defined orthogonal to each other.

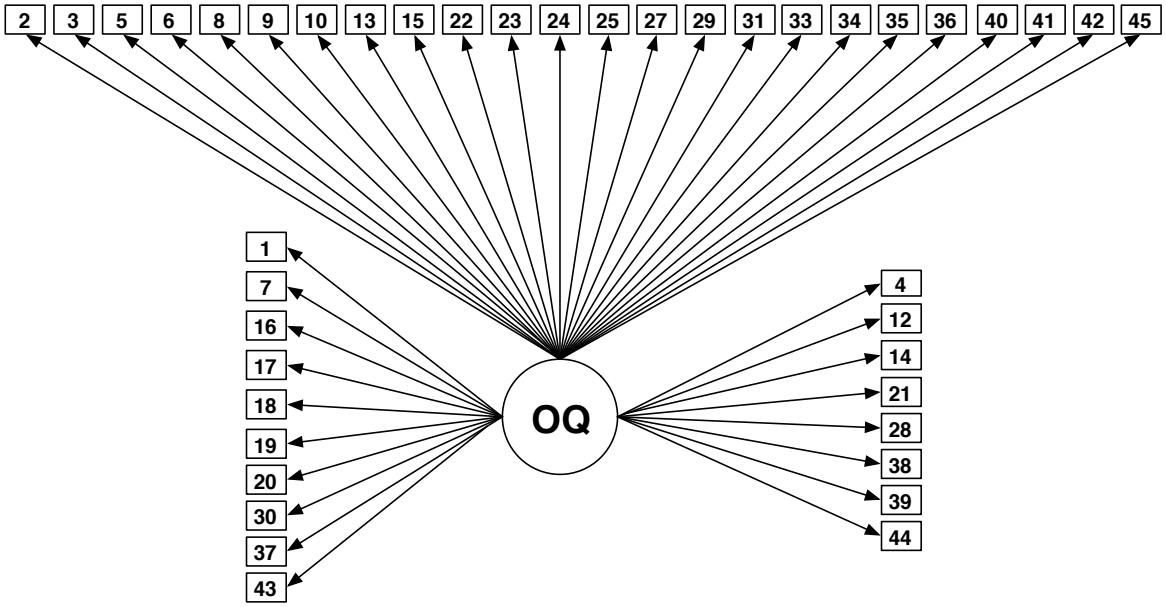


Figure 5. *UGRM*: unidimensional graded response model. Graphic shows an unrestricted unidimensional graded response model of the OQ. Latent trait is represented as a circle and OQ items are represented as squares. Item factor loadings are represented as arrows and are all estimated. Factor means are fixed at zero and factor variances at one.

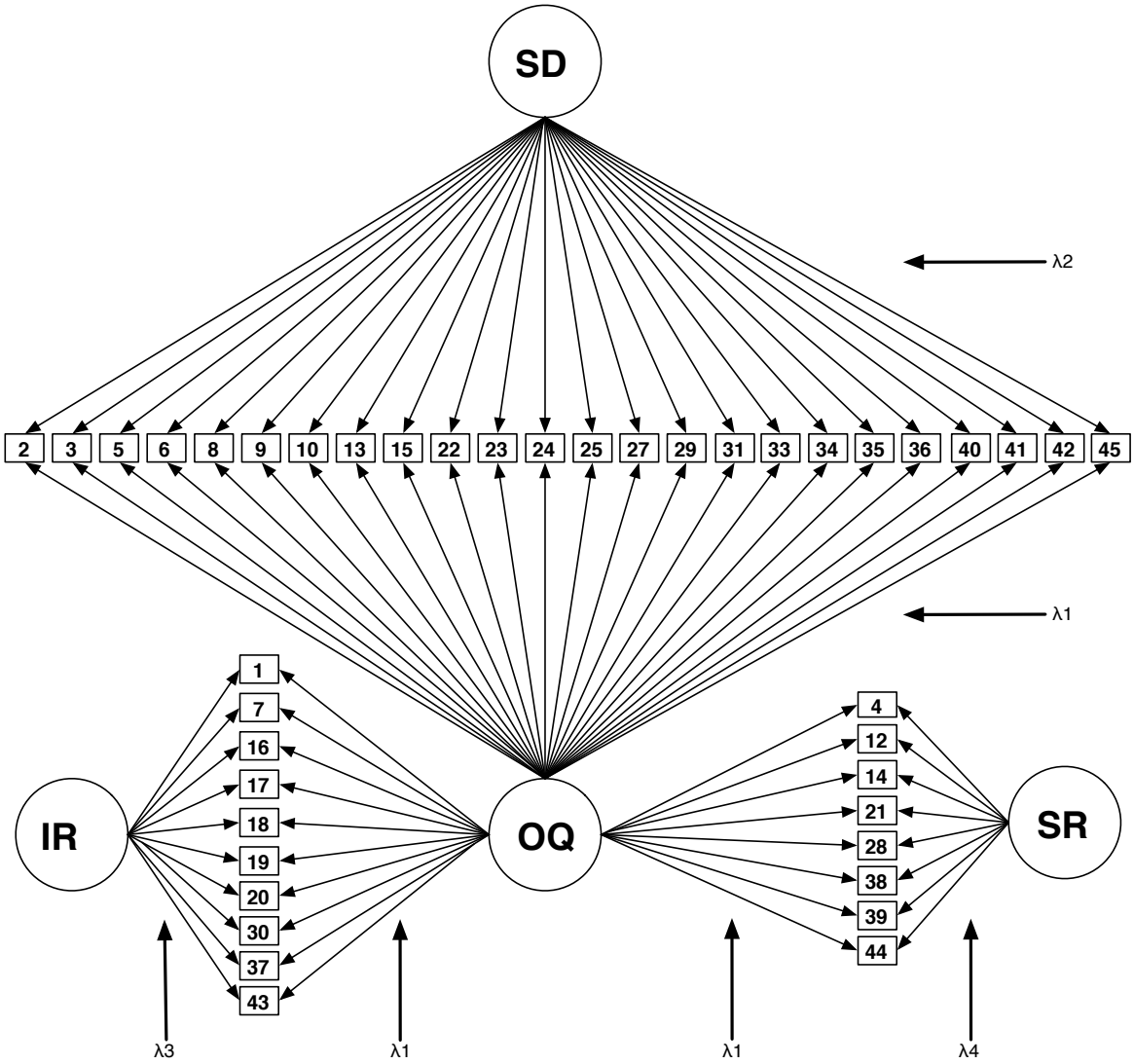


Figure 6. *BFGRMte*: tau-equivalent bifactor graded response model. Graphic shows a restricted bifactor graded response model of the OQ. Primary and group factors are represented as circles and correspond to the OQ's general dimension and subscales. OQ items are represented as squares. Item factor loadings are represented as arrows and are restricted to be equal within each factor (such that loadings are equal within each λ_k, with $k = 1, 2, 3, 4$). Factor means are fixed at zero and factor variances at one. Factors are defined orthogonal to each other.

Results

Preliminary Analyses

Preliminary and exploratory data analysis revealed two relevant concerns.[5] First, consistent with previous literature the OQ's 5-point Likert scale yields infrequent responses in the extreme categories (Pastor & Beretvas, 2006). Second, also consistent with previous literature the OQ's three substance abuse items (11, 26, and 32) show significant skew in that they are endorsed relatively infrequently (Doucette & Wolf, 2009). To reduce computational convergence concerns and interpretation problems introduced by these skewed response patterns two data management decisions were implemented. First, the lowest and highest categories (i.e. 0=Never and 1=Rarely, and 3=Frequently and 4=Almost Always, respectively) were collapsed for all observations yielding three effective categories. Second, the three substance abuse items were excluded from all subsequent analyses. This is in line with previous research by Doucette and Wolf (2009), who suggest that the substance abuse items may be best regarded as forming a separate subset of items apart from the other items on the OQ.[6] Final data analyses are presented for data from cohort's I and II with response categories collapsed and substance abuse items removed.

[5]Feasibility and soundness of model specifications were assessed with two independent random samples of 200 observations from cohort I. Assessment proceeded using a subset of OQ items to reduce model complexity and to minimize computational difficulty. Three items were chosen from SD and SR subscales, and two items from the IR subscale. In addition, all three substance abuse items were included to assess sensitivity of the implementation to anomalous items. This yielded a final subset consisting of items 1, 4, 5, 8, 11, 26, 30, 32, 35, 39, and 44. The bifactor GRM in Mplus, POLYBIF (Gibbons & Hedeker, n.d.), and IRTPRO beta (Cai, du Toit, & Thissen, n.d.) software all converged, although inclusion of substance abuse items introduced convergence problems for Mplus using the MLR estimator. Factor loadings and item thresholds were comparable across software programs apart from scaling. This substantiates the soundness of the bifactor GRM as implemented in the present study. Further details available from the author upon request.

[6]The developers of the OQ acknowledge the problematic nature of the substance abuse items, however the items were included as requested by a funding contributor in spite of their divergent properties (Michael Lambert, personal communication, June 24, 2010).

Dimensionality

Initial analyses show superior fit for the bifactor graded response model (*BFGRM*) on a random subset ($N = 2297$) of cohort I's (I_a) data compared to the unidimensional graded response model *UGRM*, $\chi^2_{diff}(42) = 2313.94$, $p < .001$. In addition, the *BFGRM* shows superior fit compared to the tau-equivalent model (*BFGRMte*), $\chi^2_{diff}(80) = 3048.34$, $p < .001$.

Cross-validation of these results on the remaining observations from cohort I's (I_b) data and cohort II's full data yielded similar results. In cohort I_b's data ($N = 2382$) the *BFGRM* shows superior fit compared to the *UGRM*, $\chi^2_{diff}(42) = 2346.10$, $p < .001$. In addition, the *BFGRM* shows superior fit compared to the tau-equivalent model (*BFGRMmte*), $\chi^2_{diff}(80) = 3176.14$, $p < .001$. In cohort II's data ($N = 4500$), the *BFGRM* shows superior fit compared to the *UGRM*, $\chi^2_{diff}(42) = 5718.03$, $p < .001$. In addition, the *BFGRM* shows superior fit compared to the tau-equivalent model (*BFGRMte*), $\chi^2_{diff}(80) = 9228.09$, $p < .001$. Taken together these results provide support for the utility of a bifactor model over a unidimensional model for the OQ. In addition, a full, unrestricted version of the bifactor model achieves better fit than models where item factor loadings are restricted within factors or scales.

Item Behavior

Item behavior in traditional polytomous IRT models is evaluated in terms of discrimination and item category difficulty. The bifactor IRT implementation presented here provides parameters that are analogous, although not strictly identical to, traditional IRT parameters: (a) analogous to item discrimination, items are evaluated with respect to item factor loadings on the primary as well as group factors; (b) analogous to item category difficulty, items are evaluated with respect to item category thresholds. However, because under the bifactor model

item thresholds are not directly interpretable in terms of any one factor, they cannot be interpreted as the invariant location with respect to a unidimensional latent trait like in unidimensional IRT models. Rather, thresholds pertain to an additive composite of all latent variables in the model and as a such are not separated between primary and group factors. In what follows, discrimination and difficulty refer to parameters as they apply in the bifactor IRT case.

Item factor loadings. Figure 7 shows, for each of the OQ's subscales separately, a comparison of primary loadings versus group loadings across datasets. Each item's coordinates reflect its relative discrimination on the primary versus group factor. Positive values reflect an items' discriminating ability, that is the information an item provides about persons latent trait standing. Values close to zero reflect poor discriminating ability, that is persons high and low on the latent trait have comparable probabilities of response endorsement. Negative values reflect negative discriminating ability such that on these items persons at higher trait levels have a lower probability of response endorsement and persons at lower trait levels have a higher probability of response endorsement. Because items with discrimination close to zero and smaller provide little to no useful information about persons latent trait standing, such items are considered undesirable. Based on Figure 7 four observations can be made. First, as required given the bifactor structure, item discrimination (loadings) is higher on the primary than on group dimension. That said, there is definite separation of items across both dimensions in that some items discriminate more highly on both dimensions, some discriminate more highly on primary, less so on group dimension, some discriminate less highly on primary, more highly on group dimension, and some items do not

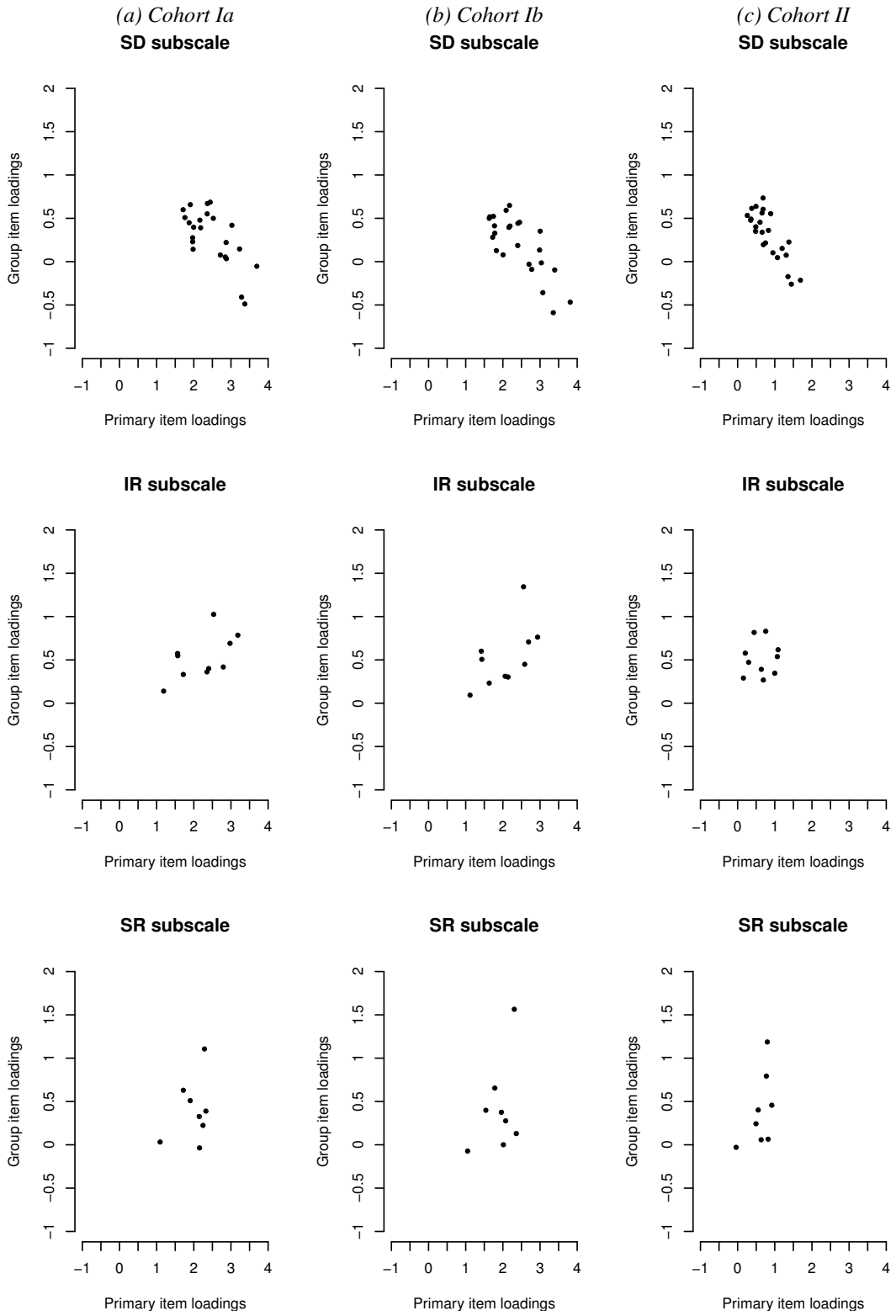


Figure 7. Primary and subscale item factor loading estimates. Graphic shows the magnitude of item factor loadings on the OQ's primary versus subscale dimensions. Axes denote each dimension's respective latent trait scale, in both cases with a mean of zero and standard deviation of one.

discriminate as highly on either dimension. Second, across different datasets items on the SD subscale are more clustered on both primary and group dimensions than items on either the IR or SR subscales. Third, for items on the IR subscale the range of discrimination appears larger across the group dimension than across the primary dimension compared to the SD or SR subscales across datasets. Finally, items on the SR subscale appear to have among the lowest discriminatory power across dimensions across datasets.

Table 1 gives the exact magnitude of item factor loadings shown in Figure 7. Although differences in absolute magnitude of loadings are evident across datasets, a comparison of relative magnitude suggest a similar pattern of relative item discrimination behavior across datasets: Spearman's correlations of items ranked by item factor loading magnitude are high on the primary dimension (.84–.94), on the SD subscale (.95–.98), on the IR subscale (.76–.98), and on the SR subscale (.90–.98). The OQ's primary dimension appears well-suited to its task of distinguishing those with higher distress from those with lower distress; all items have moderate to high loadings on the primary factor. The three most discriminating items were 31 ("I am satisfied with my life"), 15 ("I feel worthless", and 24, ("I like myself") in cohort I, and items 31, 13 ("I am a happy person"), and 42 ("I feel blue") in cohort II. The three least discriminating items were 14 ("I work/study too much"), 16 ("I am concerned about family troubles"), and 7 and 17 ("I feel unhappy in my marriage/ significant relationship" and "I have an unfulfilling sex life"), in cohorts I and II respectively.

Item factor loadings on the group dimensions denote the discriminating ability of an OQ item on its respective subscale, controlling for the primary dimension. The three most discriminating items on the Symptom Distress (SD) subscale in cohorts I and II were 36 ("I feel nervous"), 29 ("My heart

Table 1
Primary and subscale item factor loading estimates.

	Cohort Ia				Cohort Ib				Cohort II			
Item	**OQ**	SD	IR	SR	**OQ**	SD	IR	SR	**OQ**	SD	IR	SR
1	2.40		0.40		2.14		0.30		0.69		0.27	
2	2.18	0.39			1.73	0.28			0.60	0.46		
3	2.71	0.08			2.71	-0.03			1.07	0.05		
4	1.90			0.51	1.54			0.40	0.56			0.40
5	1.98	0.14			2.01	0.08			0.75	0.22		
6	1.97	0.23			1.83	0.13			0.69	0.20		
7	1.56		0.57		1.43		0.51		0.44		0.82	
8	2.84	0.05			2.78	-0.09			0.95	0.10		
9	3.02	0.42			3.00	0.35			0.89	0.55		
10	2.53	0.50			2.45	0.45			0.69	0.60		
12	2.33			0.39	1.96			0.37	0.92			0.46
13	3.28	-0.41			3.08	-0.36			1.45	-0.26		
14	1.09			0.03	1.05			-0.07	-0.04			-0.03
15	3.70	-0.05			3.40	-0.10			1.31	0.08		
16	1.19		0.14		1.12		0.09		0.16		0.29	
17	1.57		0.55		1.42		0.60		0.21		0.58	
18	2.79		0.42		2.59		0.45		1.00		0.35	
19	1.72		0.33		1.63		0.23		0.30		0.47	
20	3.18		0.79		2.94		0.76		1.07		0.54	
21	2.16			-0.03	2.02			0.00	0.82			0.07
22	1.97	0.28			1.78	0.33			0.66	0.34		
23	2.88	0.03			3.04	-0.01			1.20	0.15		
24	3.37	-0.49			3.35	-0.59			1.36	-0.17		
25	2.00	0.40			2.16	0.40			0.48	0.35		
27	1.91	0.66			1.74	0.52			0.38	0.61		
28	1.72			0.63	1.78			0.65	0.77			0.79
29	2.37	0.67			2.09	0.59			0.49	0.64		
30	2.36		0.36		2.06		0.31		0.64		0.39	
31	4.29	-0.52			3.81	-0.47			1.69	-0.21		
33	2.16	0.48			2.40	0.44			0.66	0.56		
34	1.76	0.51			1.64	0.50			0.26	0.53		
35	2.36	0.55			2.19	0.41			0.35	0.48		
36	2.44	0.69			2.18	0.65			0.68	0.73		
37	2.53		1.03		2.56		1.34		0.76		0.83	
38	2.29			1.11	2.31			1.56	0.80			1.19
39	2.15			0.33	2.07			0.28	0.49			0.24
40	2.87	0.22			2.40	0.19			0.83	0.36		
41	1.87	0.45			1.77	0.41			0.49	0.40		
42	3.23	0.15			2.99	0.13			1.38	0.23		
43	2.98		0.69		2.69		0.71		1.09		0.62	
44	2.25			0.22	2.36			0.13	0.63			0.06
45	1.71	0.60			1.65	0.52			0.36	0.49		

pounds too much"), and 27 ("I have an upset stomach"). The three least discriminating items were 31 ("I am satisfied with my life"), 24 ("I like myself"), and 13 ("I am a happy person"). Thus, these items discriminate highly on the primary factor, but not on their group factor. On the Interpersonal Relations (IR) subscale the three most discriminating items were 37 ("I feel my love relationships are full and complete"), 20 ("I feel loved and wanted"), 43 ("I am satisfied with my relationships with others") and 7, in cohorts I and II respectively. The three least discriminating items were 16 ("I am concerned about family troubles"), 19 ("I have frequent arguments"), and 1 ("I get along well with others"), 18 ("I feel lonely"), and 30 ("I have trouble getting along with friends and close acquaintances"), in cohorts I and II respectively. On the Social Role (SR) subscale the three most discriminating items were 38 ("I feel that I am not doing well at work/school"), 28 ("I am not working studying as well as I used to"), and 4 ("I feel stressed at work school") and 12 ("I find my work/school satisfying"), in cohorts I and II respectively. The three least discriminating items were 21 ("I enjoy my spare time"), 14 ("I work/study too much"), and 44 ("I feel angry enough at work/school to do something I might regret"). From the above it is clear that most items have substantial loadings on the primary factor and less substantial loadings on the group factors. Particularly, pronounced factor loading differences between primary and group factors are seen for items 13, 24, and 31 on the SD subscale. That is, these items discriminate highly on the primary factor, but not on their group factor. In other words, they suggest distinct differences in content and measurement precision between the primary construct measured by the OQ and the specific construct measured by the SD subscale.

Estimates of the differences in measurement precision between primary and subscale dimensions can be calculated and provide information akin to the

CTT notion of reliability. The notion of factor reliability used here derives from Reise et al. (2007) who suggest a calculation method using the standard errors of expected a posteriori factor scores . They write: "the size of these reliability estimates indicates the degree to which individuals could be precisely assessed on the group factors, sans the general" (pp. 26–27). The reliability estimates for the primary scale and subscales are shown in Table 2. Estimates for the primary scale are comparable to values of Cronbach's α in the data used (.98, .98, and .92 for Cohort Ia, Ib, and II, respectively). For the subscales, controlling for the general factor, precision is poor by common standards of reliability (Nunnally, 1978). Naturally, given the bifactor structure, reduced precision on subscales is expected after controlling for the general dimension. In other words, measurement precision obtained at the general level, is, to a degree, at the expense of measurement precision on the subscale level.

Item category thresholds. Item category thresholds or difficulty, contrary to item discrimination, cannot be analyzed separately for primary and group factors under the bifactor model. What remains is a nonetheless informative interpretation akin to that in unidimensional IRT difference models such as Samejima's GRM (1969, 1996), with the caveat that difficulty is not related to a single underlying dimension, but rather to an additive composite of both primary and group dimensions. Recall that in difference models item

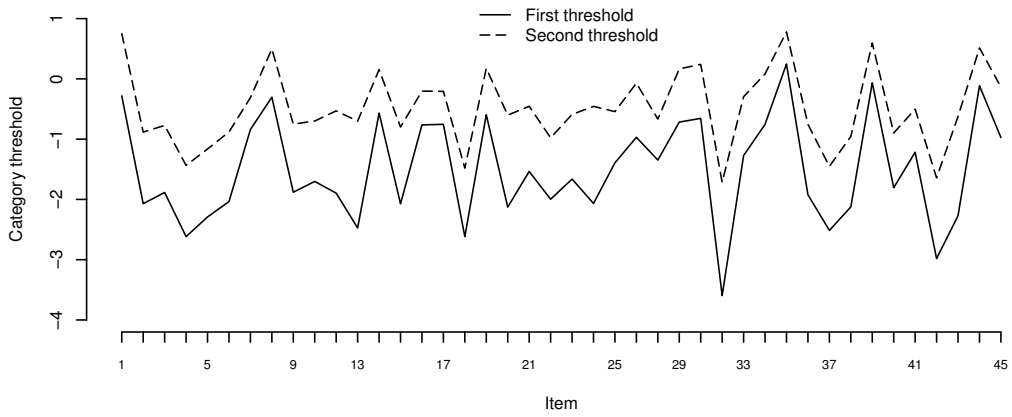
Table 2
Reliability estimates for OQ primary scale versus subscales (controlling for the general dimension)

		Reliability estimates		
		Cohort Ia	Cohort Ib	Cohort II
Primary		.99	.99	.92
Subscale	SD	.76	.74	.67
	IR	.71	.73	.60
	SR	.66	.70	.55

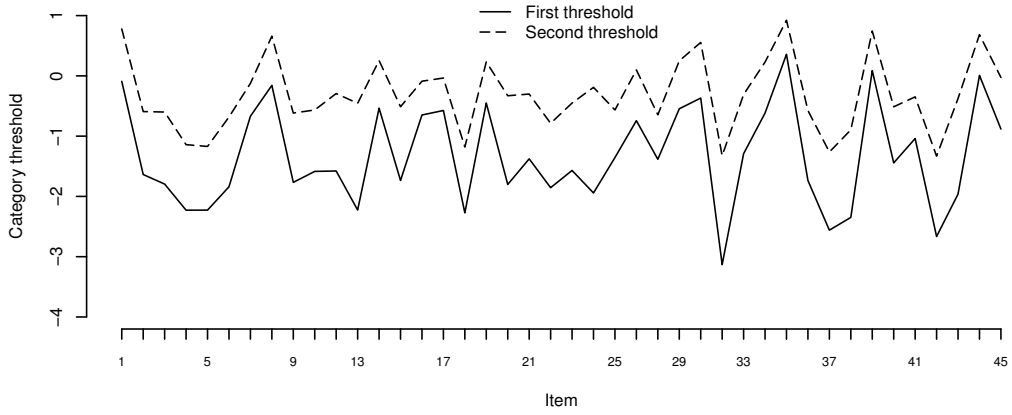
thresholds refer to the location on the latent trait (or in the bifactor case, the location with respect to the additive composite of latent variables in the model) at which persons have a .5 probability of responding in a particular category or higher. Recall also that the OQ's five-point Likert scale was collapsed to a three-point scale for the purposes of this study. Therefore, the first threshold reflects the point on the latent trait scale where persons have a .5 probability of responding in collapsed category 1 or higher (i.e. original OQ anchors "Never" and "Rarely"). The second threshold reflects the point on the latent trait scale where persons have a .5 probability of responding in collapsed category 2 or higher (i.e. original OQ anchors "Sometimes," "Frequently," and "Always"). With respect to the OQ, the latent trait scale may be conceptualized as reflecting less or more, respectively, levels of psychological distress. Thus, an item's threshold may be seen as reflecting its severity. Low threshold items are less severe, meaning that they are endorsed when a relatively lower level of psychological distress is present. High threshold items are more severe, meaning that they are not endorsed until a relatively higher level of psychological distress is present. First thresholds capture initial or lower anchor endorsement, whereas second thresholds capture subsequent or higher anchor endorsement. Thus, items with a high first threshold are severe items endorsed minimally at high levels of distress. Items with low first thresholds are less severe items, endorsed at lower levels of distress. Items with a high second threshold are items with higher anchor endorsement at higher levels of distress. Items with a low second threshold are items with higher anchor endorsement at lower levels of distress.

Figure 8 shows a comparison of first and second category thresholds across datasets. Table 3 gives the exact magnitude of item category thresholds shown in this figure. Although differences in which items appear to have highest and lowest thresholds are evident between datasets, overall relative item

(a) Cohort Ia



(b) Cohort Ib



(c) Cohort II

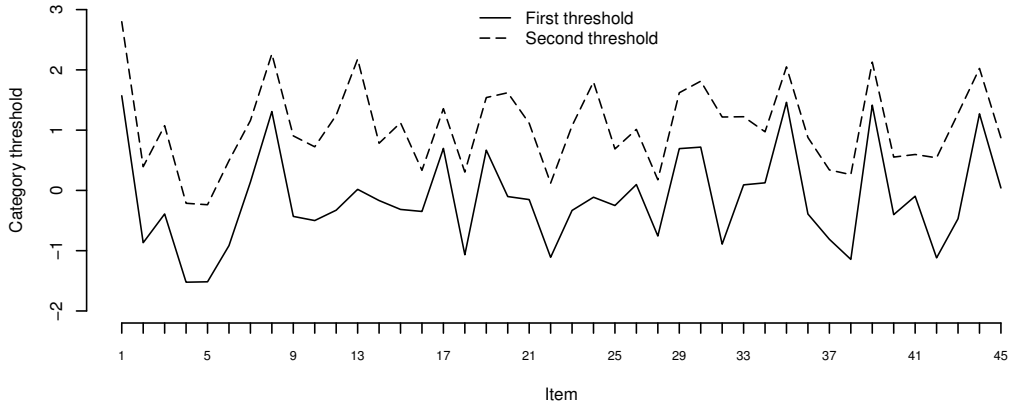


Figure 8. First and second item category threshold estimates. Graphic shows the magnitude of first and second category thresholds. The horizontal axis denotes OQ items; the vertical axis denotes the OQ's composite latent trait scale (the additive composite of both primary and group dimensions) with a mean of zero and standard deviation of one.

Table 3
First and second item category threshold estimates.

	Cohort Ia		Cohort Ib		Cohort II	
Item	**First**	Second	**First**	Second	**First**	Second
1	-0.28	0.75	-0.09	0.78	1.57	2.80
2	-2.07	-0.88	-1.64	-0.59	-0.87	0.39
3	-1.88	-0.77	-1.80	-0.60	-0.39	1.07
4	-2.62	-1.44	-2.23	-1.14	-1.52	-0.21
5	-2.29	-1.16	-2.23	-1.17	-1.52	-0.24
6	-2.04	-0.88	-1.84	-0.68	-0.92	0.49
7	-0.84	-0.31	-0.67	-0.12	0.14	1.16
8	-0.30	0.50	-0.16	0.66	1.31	2.26
9	-1.88	-0.75	-1.77	-0.62	-0.43	0.90
10	-1.70	-0.70	-1.59	-0.57	-0.50	0.72
12	-1.90	-0.53	-1.58	-0.29	-0.33	1.25
13	-2.47	-0.70	-2.23	-0.46	0.02	2.18
14	-0.57	0.15	-0.54	0.25	-0.17	0.78
15	-2.07	-0.80	-1.74	-0.51	-0.32	1.12
16	-0.76	-0.20	-0.65	-0.09	-0.35	0.33
17	-0.75	-0.21	-0.57	-0.04	0.70	1.36
18	-2.62	-1.48	-2.27	-1.18	-1.07	0.30
19	-0.60	0.17	-0.45	0.23	0.67	1.54
20	-2.13	-0.60	-1.80	-0.33	-0.10	1.62
21	-1.54	-0.45	-1.38	-0.30	-0.15	1.11
22	-2.00	-0.98	-1.86	-0.79	-1.11	0.12
23	-1.66	-0.59	-1.57	-0.45	-0.33	1.07
24	-2.07	-0.46	-1.94	-0.19	-0.11	1.80
25	-1.39	-0.54	-1.36	-0.57	-0.25	0.69
27	-0.97	-0.07	-0.74	0.10	0.10	1.01
28	-1.35	-0.67	-1.38	-0.65	-0.76	0.18
29	-0.72	0.17	-0.55	0.25	0.69	1.62
30	-0.66	0.24	-0.37	0.55	0.72	1.81
31	-3.60	-1.72	-3.13	-1.32	-0.89	1.22
33	-1.27	-0.30	-1.29	-0.31	0.09	1.22
34	-0.75	0.08	-0.62	0.23	0.13	0.97
35	0.25	0.78	0.36	0.92	1.46	2.05
36	-1.92	-0.75	-1.74	-0.58	-0.39	0.88
37	-2.51	-1.45	-2.56	-1.26	-0.81	0.34
38	-2.12	-0.95	-2.35	-0.90	-1.14	0.26
39	-0.07	0.60	0.09	0.74	1.41	2.13
40	-1.81	-0.90	-1.44	-0.51	-0.40	0.55
41	-1.22	-0.50	-1.04	-0.35	-0.10	0.60
42	-2.98	-1.64	-2.67	-1.33	-1.12	0.54
43	-2.27	-0.62	-1.96	-0.38	-0.47	1.28
44	-0.11	0.51	0.01	0.68	1.27	2.02
45	-0.97	-0.14	-0.88	-0.03	0.04	0.87

threshold behavior appear similar. As expected Spearman's correlation of first item category thresholds ranked by magnitude between cohort Ia and Ib was .98. Between cohort Ia and Ib, and cohort II correlations were .83 and .86, respectively. Second item category thresholds correlated at .97 between cohort Ia and Ib. Between cohort Ia and Ib, and cohort II correlations were .68 and .73, respectively. Inspection of item rankings by quantile reveals that difference between datasets predominantly occur at quantile boundaries, which similarly suggests that ranking differences are small.

For cohort Ia and Ib items with the highest first thresholds were 35 ("I feel afraid of open spaces, of driving, or being on buses, subways, and so forth"), 39 ("I have too many disagreements at work/school"), and 44 ("I feel angry enough at work/school to do something I might regret"), revealing that initial endorsement requires a relatively higher level of the composite of latent variables than the remaining items. For cohort II item 1 ("I get along well with others") showed a high first threshold in addition to items 35 and 39. Items with high second thresholds (i.e. subsequent endorsement) for cohort Ia and Ib included items 35, 1, and 39. For cohort II items 1, 8 ("I have thoughts of ending my life"), and 13 ("I am a happy person"), had highest second thresholds. Items with a low first threshold, reflecting that initial endorsement is likely at relatively low levels of the composite latent trait, were 31 ("I am satisfied with my life"), 42 ("I feel blue"), and 18 and 37 ("I feel lonely" and "I feel my love relationships are full and complete") for cohort Ia and Ib. For cohort II items 4 ("I feel stressed at work/school"), 5 ("I blame myself for things"), and 38 ("I feel that I am not doing well at work/school") had lowest first thresholds. Items with low second thresholds (i.e. subsequent endorsement) for cohort Ia and Ib included items 31, 42, and 18 and 37. For cohort II items 4, 5, and 22 had the lowest second thresholds.

Measurement Invariance

Findings reported thus far assume that item factor loadings and item category thresholds are invariant across possibly clinically relevant domains such as gender. This assumption, known as measurement invariance, is tested through fitting a succession of measurement invariance models across a specified group. In what follows measurement invariance for the bifactor model of the OQ is examined across gender. Table 4 shows fit statistics for a configural invariance model, with item factor loadings and item category thresholds equal between men, the reference group, and women, the focal group. Results indicate acceptable fit across datasets, which suggest that further exploration of measurement invariance is feasible.

Item factor loading (metric) invariance. Invariance of item factor loadings was examined in a model with constrained item factor loadings. In the reference group factor means and variances were fixed at zero and one, respectively. In the focal group factor means and variances fixed at zero and estimated, respectively. Item-residual variances were fixed at one in both reference and focal groups for model identification. Item category thresholds were estimated in both groups separately. A χ^2-difference test reveals measurement invariance of item factor loadings (of primary and group factors) with respect to gender in cohort Ia, $\chi^2_{diff}(80) = 97.45$, p $= .090$ ($n_{\text{men}} = 512$, $n_{\text{women}} = 559$) and cohort Ib $\chi^2_{diff}(80) = 87.31$, p $= .270$ ($n_{\text{men}} = 560$,

Table 4
Fit statistics configural measurement invariance model in each of the three datasets.

Data	χ^2(DF)	CFI	TLI	RMSEA
Cohort I_a	3618.60 (1554)	.914	.905	.050
Cohort I_b	3885.33 (1554)	.914	.905	.052
Cohort II	9303.97 (1554)	.922	.914	.054

$n_{\text{women}} = 530$), and non-invariance in cohort II, $\chi^2_{diff}(80) = 154.01$, p $< .001$ ($n_{\text{men}} = 1710$, $n_{\text{women}} = 1661$).

Partial invariance of item factor loadings with respect to gender was examined for cohort II. Successive removal of equality constraints of item factor loadings between groups revealed non-invariance for primary loadings of items 19 ("I have frequent arguments"), 33 ("I feel that something bad is going to happen"), 10 ("I feel fearful"), 16 ("I am concerned about family troubles"), and 34 ("I have sore muscles"). Removing further item factor loading equality constraints no longer yielded significantly improved fit compared to the configural model. Thus, the remaining item factor loadings were considered invariant across gender in cohort II's data. Table 5 shows standardized item factor loadings and model comparisons of non-invariant items for both men and women. Results indicate that items were 33, 10, and 16 were significantly more discriminating for men than for women on the primary latent factor of the OQ. Items 19 and 34 were significantly more discriminating for women than for men.

Item category threshold (scalar) invariance. Invariance of item category thresholds was examined in a model with constrained item thresholds. In the

Table 5
Non-invariant item factor loadings for men vs. women in cohort II

Item	Men	Women	χ^2_{diff}	p
19	.219	.329	135.24	.000
33	.609	.461	121.63	.001
10	.615	.476	107.96	.011
16	.213	.117	101.17	.028
34	.250	.280	99.04	.033

Note. Results shown reflect order in which equality constraints were removed based on largest modification index ($\chi^2 > 3.42$, p $< .05$). Item factor loadings shown are standardized estimates from a model with equality constraints removed for all non-invariant items. Resulting χ^2-difference tests (with one degree of freedom) are shown between models with successively unconstrained item factor loadings. Associated p-values are given.

reference group factor means and variances were fixed at zero and one, respectively. In the focal group factor means and variances were estimated. Item-residual variances were fixed at one in both reference and focal groups for model identification. Item factor loadings were constrained in both groups for cohort I. For cohort II, item factor loadings found to be non-invariant above were estimated separately between groups. All other item factor loadings were constrained between groups. A χ^2-difference test reveals measurement non-invariance of item category thresholds with respect to gender in cohort Ia $\chi^2_{diff}(80) = 231.47$, p $< .001$, cohort Ib $\chi^2_{diff}(80) = 334.98$, p $< .001$, and cohort II $\chi^2_{diff}(80) = 759.78$, p $< .001$.

Partial invariance of item category thresholds with respect to gender was examined for cohort Ia. Non-invariance was found for the first category threshold of items 2 ("I tire quickly"), 17 ("I have an unfulfilling sex life"), 27 ("I have an upset stomach"), and 42 ("I feel blue"), and the first and second category thresholds of item 45 ("I have headaches"). No further item category thresholds suggested model fit improvement at a .05 significance level. Thus, remaining thresholds were considered invariant across gender in cohort Ia.

Table 6 shows standardized item category thresholds and model comparisons of non-invariant items for both men and women. Results indicate that first thresholds of items 2, 27, 42, and 45 are higher for men than for women. In other words, these items are more readily endorsed at lower levels of distress initially by women than by men. The same is true for endorsement in higher categories for item 45. Item 17 ("I have an unfulfilling sex life") on the other hand, has a first threshold that is higher for women than for men. In other words, this item is more readily endorsed at lower levels of distress initially by men than by women.

Table 6
Non-invariant item category thresholds for men vs. women in cohort Ia

Item		Men	Women	χ^2_{diff}	p
(threshold)				.	
45	(1)	0.329	-0.077	211.68	.000
45	(2)	1.026	0.650	196.38	.000
27	(1)	0.381	0.055	179.57	.000
2	(1)	-0.413	-0.790	160.77	.000
17	(1)	-0.088	0.202	146.59	.000
42	(1)	-0.650	-0.963	136.69	.000

Note. Results shown reflect order in which equality constraints were removed based on largest modification index ($\chi^2 > 3.42$, $p < .05$). Item category thresholds shown are standardized estimates from a model with equality constraints removed for all non-invariant items. Resulting χ^2-difference tests (with one degree of freedom) are shown between models with successively unconstrained item category thresholds. Associated p-values are given.

Partial invariance of item category thresholds with respect to gender was also examined for cohort Ib. Non-invariance was found for the first category threshold of items 2, 5 ("I blame myself for things"), 17, 35 ("I feel afraid of open spaces, of driving, or being on buses, subways, and so forth"), and 40 ("I feel that something is wrong with my mind"). In addition, second category thresholds were non-invariant for items 4 ("I feel stressed at work/school") and 13 ("I am a happy person"). Both thresholds were non-invariant for items 21 ("I enjoy my spare time"), 27, 34 ("I have sore muscles"), and 45. No further item category thresholds suggested model fit improvement at a .05 significance level. Thus, remaining thresholds were considered invariant across gender in cohort Ib.

Table 7 shows standardized item category thresholds and model comparisons of non-invariant items for both men and women. Results indicate that first thresholds of items 2, 5, 27, 34, 35, and 45 are higher for men than for women. In other words, these items are more readily endorsed at lower levels of distress initially by women than by men. The same is true for endorsement in higher categories for item 27, 34, and 45. Items 17 and 40, on the other hand,

Table 7
Non-invariant item category thresholds for men vs. women in cohort Ib

Item		Men	Women	χ^2_{diff}	p
(threshold)				.	
45	(2)	1.186	0.613	307.01	.000
45	(1)	0.395	-0.037	284.50	.000
4	(2)	-0.085	-0.385	267.69	.000
17	(1)	-0.076	0.203	253.63	.000
27	(1)	0.400	0.065	242.56	.000
40	(1)	-0.217	0.007	231.65	.000
13	(2)	0.489	0.859	218.98	.000
21	(1)	-0.352	-0.118	209.11	.000
35	(1)	1.465	1.135	201.48	.000
2	(1)	-0.395	-0.683	191.08	.000
34	(2)	1.116	0.844	183.08	.000
27	(2)	1.076	0.783	174.46	.000
21	(2)	0.499	0.758	165.55	.000
34	(1)	0.347	0.131	157.42	.000
5	(1)	-0.798	-1.099	148.17	.000

Note. Results shown reflect order in which equality constraints were removed based on largest modification index ($\chi^2 > 3.42$, $p < .05$). Item category thresholds shown are standardized estimates from a model with equality constraints removed for all non-invariant items. Resulting χ^2-difference tests (with one degree of freedom) are shown between models with successively unconstrained item category thresholds. Associated p-values are given.

have first thresholds that are higher for women than for men. In other words, these item are more readily endorsed at lower levels of distress initially by men than by women. Similarly, items 4, 13, and 21 are more readily endorsed in higher categories at lower levels of distress by men than by women.

Finally, partial invariance of item category thresholds with respect to gender was examined for cohort II. The first category threshold of item 44 ("I feel angry enough at work/school to do something I might regret") was non-invariant. Additionally, second category thresholds were non-invariant for items 5 ("I blame myself for things"), 10 ("I feel fearful"), and 34 ("I have sore muscles"). Both thresholds were non-invariant for items 2 ("I tire quickly"), 4 ("I feel stressed at work/school"), 6 ("I feel irritated"), 15 ("I feel worthless"), 16 ("I

am concerned about family troubles"), 17 ("I have an unfulfilling sex life"), 18 ("I feel lonely"), 24 ("I like myself"), 25 ("Disturbing thoughts come into my head that I cannot get rid of"), 27 ("I have an upset stomach"), 30 ("I have trouble getting along with friends and close acquaintances"), 35 ("I feel afraid of open spaces, of driving, or being on buses, subways, and so forth"), 40 ("I feel something is wrong with my mind"), 42 ("I feel blue"), and 45 ("I have headaches"). No further item category thresholds suggested model fit improvement at a .05 significance level. Thus, remaining thresholds were considered invariant across gender in cohort II.

Table 8 shows standardized item category thresholds and model comparisons of non-invariant items for both men and women. Results indicate that first and second thresholds of items 2, 4, 6, 15, 18, 24, 27, 30, 35, 42, and 45 are higher for men than for women. In other words, women more readily endorse these items at lower levels of distress than men. Second thresholds of items 10 and 34 are higher for men than for women, indicating that women more readily endorse these items in higher categories than men. Items 17, 25, and 40 have first and second thresholds higher for women than for men. In other words, these item are more readily endorsed initially by men than by women. Item 44 has a higher first threshold for women than for men, indicating that men more readily endorse this item at lower levels of distress.

Table 8
Non-invariant item category thresholds for men vs. women in cohort II

Item (threshold)		Men	Women	χ^2_{diff}.	p
16	(2)	0.609	0.145	888.43	.000
16	(1)	-0.088	-0.474	798.67	.000
25	(1)	-0.305	-0.003	726.92	.000
25	(2)	0.516	0.830	654.79	.000
45	(2)	1.153	0.724	607.86	.000
17	(1)	0.383	0.736	557.51	.000
17	(2)	0.924	1.361	493.87	.000
45	(1)	0.397	0.047	449.58	.000
27	(1)	0.447	0.118	408.35	.000
2	(1)	-0.418	-0.735	378.52	.000
2	(2)	0.595	0.352	352.36	.000
24	(2)	1.196	0.960	330.79	.000
40	(1)	-0.276	-0.136	307.72	.000
18	(1)	-0.516	-0.815	283.09	.000
24	(1)	0.023	-0.220	261.68	.000
27	(2)	1.105	0.899	247.36	.000
35	(1)	1.558	1.321	233.29	.000
18	(2)	0.363	0.177	216.73	.000
4	(2)	-0.026	-1.301	202.49	.000
30	(2)	1.583	1.405	191.58	.000
40	(2)	0.428	0.549	180.27	.000
44	(1)	1.010	1.181	169.30	.000
6	(2)	0.524	0.376	160.45	.000
42	(2)	0.440	0.296	150.87	.000
15	(1)	-0.087	-0.263	140.36	.000
42	(1)	-0.476	-0.688	129.31	.000
15	(2)	0.761	0.654	119.51	.000
5	(2)	-0.063	-0.199	111.40	.000
34	(2)	1.081	0.928	104.50	.000
10	(2)	0.718	0.657	96.41	.000
35	(2)	1.999	1.822	90.47	.000
4	(1)	-1.108	-1.301	81.80	.001
6	(1)	-0.605	-0.763	73.74	.007
30	(1)	0.648	0.552	67.18	.022

Note. Results shown reflect order in which equality constraints were removed based on largest modification index ($\chi^2 > 3.42$, $p < .05$). Item category thresholds shown are standardized estimates from a model with equality constraints removed for all non-invariant items. Resulting χ^2-difference tests (with one degree of freedom) are shown between models with successively unconstrained item category thresholds. Associated p-values are given.

In sum, a majority of items appears invariant across gender with respect to item factor loadings and item category thresholds in cohort I. In cohort II, more non-invariant item factor loadings and category thresholds were found. Although there are differences between which items show non-invariance between datasets, four items show threshold non-invariance in all samples and three items show invariance across two samples. Items 2 ("I tire quickly"), 17 ("I have an unfulfilling sex life"), 27 ("I have an upset stomach") and 45 ("I have headaches") show non-invariance across all three samples studied for the first threshold (i.e. minimal endorsement). Items 34 ("I have sore muscles") 35 ("I feel afraid of open spaces, of driving, or being on buses, subways, and so forth"), and 40 ("I feel something is wrong with my mind") are non-invariant with respect to its first threshold across two samples. Further, items 4 ("I enjoy my spare time"), 27 and 34 show non-invariance across two samples for the second threshold (i.e. higher anchor endorsement). Thus, items 27 and 34 are non-invariant in two samples across both thresholds; item 45 is non-invariant in all three samples across both thresholds. Examining standardized estimates of these non-invariant thresholds shows that women more readily endorse items 2, 27, 34, 35, and 45 in lower response categories at lower levels of distress than men. The same is true for items 4, 27, 34, and 45 in higher response categories. Items 17 and 40, on the other hand are more readily endorsed in lower response categories at lower levels of distress by men than by women. Of these threshold non-invariant items, only item 34 also displays item factor loading invariance such that the item discriminates more for women than for men. Taken together these findings show significantly different item behavior for men and women, a finding cross-validated across different samples.

Conclusion

Many achievements in clinical measurement are indebted to Classical Test Theory (CTT; Allen & Yen, 2001; Lord & Novick, 1968). As a theory of measurement, it has succeeded in defining the field of clinical psychology measurement. The principle problem for the clinical researcher is unifying a quantifiable method with phenomena resistant to quantification. That is, finding a standard of measurement to capture illusive phenomena, such as worthlessness, hopelessness, phobia, and anguish is a difficult task. Because of the complexity of clinical data, clinical researchers are up against formidable challenges in modeling such data adequately. Traditional CTT methods, although powerful, commit a researcher to assumptions about their data that are not without consequence. I argued that the foundational assumption of parallel measurement in CTT yields an approach to clinical measurement that overemphasizes similarity of item properties and test-takers' responses to such items. However, Item Response Theory (IRT; Embretson & Reise, 2000) defines a measurement model that quantifies both person and item properties such that differences between them can be meaningfully represented. I further argued that the multidimensionality of most psychological constructs warrants an extension of the traditional IRT measurement model into the bifactor IRT model (Gibbons & Hedeker, 1992; Gibbons et al., 2007) to increase applicability. A bifactor IRT approach was shown for a well-known clinical measure, the Outcome Questionnaire-45 (OQ; Lambert et al., 2004). The following sections serve to (a) summarize the main findings; (b) note the implications of these findings given prior research as well as implications for clinical practice; and (c) demarcate limitations of the work presented.

Summary of Findings

The application of a bifactor IRT method to the OQ was guided by three research aims. First, in evaluating the dimensionality of the OQ a bifactor model was contrasted with a unidimensional IRT model. Second, item behavior was evaluated in terms of difficulty and discrimination parameters, and subscale utility over and above the primary scale was assessed. Third, measurement invariance, or sensitivity of item parameters to one or more relevant variables, was examined across gender.

Dimensionality. Empirical investigation of the utility of the bifactor IRT structure reveals superiority of a bifactor model over a unidimensional IRT model. The full bifactor IRT structure also show superiority to a constrained, tau-equivalent, version of the bifactor model.

Item behavior. Item behavior was evaluated by examining item factor loadings or discrimination and item thresholds or category difficulty. Items appear to discriminate well on the primary scale. Highest discrimination is obtained for items within the life satisfaction and happiness content domains.

A principle advantage of the bifactor structure is that item factor loadings or discrimination on the OQ's subscales can be assessed over and above the primary scale. Results indicate that items on the OQ's subscales maintain some discriminating ability over and above the primary scale. However, reliability estimates for the subscales, controlling for the primary scale, suggest that reliability of information obtained from subscales beyond that of the primary scale may be limited. Specific data regarding item parameters provided here may be used to guide clinical decisions based on specific constraints dictated by setting, however, clinical use of the subscales should likely proceed with caution.

Item thresholds or category difficulties pertain to an additive composite of all latent variables in the model and are not separated between primary scale

and subscale like item factor loadings. Items with high thresholds, reflecting that endorsement requires a relatively high level of the composite latent trait, are items whose content taps more infrequent or severe symptoms of psychological distress, including suicidal ideation, violent impulses, and agoraphobia. Items with low thresholds, reflecting that endorsement requires a relatively low level of the composite latent trait, are items whose content taps life/relationship satisfaction, and also symptoms of mild psychological distress including feeling stressed, mild negative mood changes, and difficulty concentrating.

Measurement invariance. Measurement invariance or differential item functioning analyses indicate where universally applied measurement models may need to deviate with respect to a relevant variable. Findings were reported for measurement invariance with respect to gender. Results suggest that equal item-factor loadings and threshold parameters hold across gender for a majority OQ items. However, for a subset of items systematic differences exist between men and women. Items that tap somatic content, including fatigue, gastrointestinal problems, sore muscles, and headaches, as well as agoraphobia are more readily endorsed by women than men at lower levels of psychological distress. Items about sex life fulfillment and disturbing thoughts are more readily endorsed by men than by women at lower levels of distress. Of these items, the item that addresses muscle soreness also appears more discriminating for women than for men.

Implications of Findings

The described application of a bifactor IRT method to the OQ has bearing on clinical measurement research, including but not limited to research on the OQ. Similarly, the presented method has implications for clinical practice, both

for use of the OQ, and more generally for clinical measurement in practice. These implications are presented in turn.

Implications for research. With the bifactor IRT method used in the present study, the utility and viability of multidimensional IRT models for clinical measurement is solidified. As argued, measurement development has long since made use of methods embedded in CTT. Given the emerging clinical research using IRT and bifactor methods, the present study included, future clinical measurement development would do well to be informed about and make use of these methods. Specifically, the information IRT methods provide about item parameters and latent trait estimates yields a much richer measurement development framework than CTT methods can provide. Further, accounting for the multidimensionality of clinical instruments is desirable. That is, specifying a measurement model that is correct in the sense that it represents the multidimensional structure of a measure allows researchers to evaluate the utility of the dimensions contained in the measure. From the above it is clear that the bifactor method allows for such an evaluation.

Implications for practice. The reported findings have several implications for the OQ over and above findings previously reported using IRT methods (see Doucette & Wolf, 2009; Pastor & Beretvas, 2006). First, the OQ's subscales appear to contribute limited information over and above the primary scale. The SD appears most reliable whereas the IR and SR subscales are much less reliable, bordering on unacceptable by common standards. Use of these subscales as independent sources of information is therefore discouraged. Second, the OQ's primary scale appears to discriminate most on items in the life satisfaction and happiness domains. Thus, the OQ's primary scale is most sensitive as a measure of these domains. Note that these findings correspond to findings concerning the 10-item screening version of the 45-item OQ, the

OQ-10.2 (Lambert et al., 1998). In a factor analytic study, item sensitivity was detected for items in the life satisfaction and happiness content domains (Seelert, Hill, Rigdon, & Schwenzfeier, 1999). The SD subscale appears to discriminate most on items in the anxiety content domain compared to the depression content domain. In deriving symptom- or diagnosis-specific information from the SD subscale this may be take into account. Third, although most of the OQ's items appear invariant between men and women, several exceptions exits. At this point further research is needed to determine the clinical significance of these findings.

Limitations

The presented method has several important limitations, some generally applicable to a bifactor IRT approach, some particular to the specific implementation of the present study. Four such limitations are described. First, the present study used single time-point data to derive item characteristics. Specifically, patient intake data was used. Previous research suggests that a subset of OQ item characteristics may be non-invariant across time (Pastor & Beretvas, 2006). Item characteristics reported here may thus not be comparable across time-points. This is particularly relevant to the OQ since its primary use is as a longitudinal measure of patient change. Second, evaluation of model fit in the present study was done using χ^2-difference test comparison of nested models for both full-information and limited-information estimation procedures. Adequacy of configural invariance models was determined by using absolute and comparative fit indices (i.e. RMSEA, CLI, and TFI). Such methods are common and well-documented in the confirmatory and item-factor analysis literature (Brown, 2006). IRT models, however, are usually evaluated at a local rather than global level. Item and person fit are two common indices of IRT model fit by which adequacy of the model is judged based on congruence

between item and person behavior given a specified IRT model (Embretson & Reise, 2000). No such methods have been developed for the bifactor IRT case. Third, a primary advantage of the bifactor approach is that item-factor loadings can be separated between primary scales and subscales of a measure. This is an advantage on a measurement level, as it allows for a multidimensional structure to be assessed in a single measurement model. It is also an advantage on an interpretative level since relationships between measured and latent variables can be meaningfully interpreted for primary as well as subscale domains. Unfortunately, no such advantage is available for item category thresholds. These thresholds are not defined separately between primary scales and subscales. This limitation is particularly salient for a bifactor IRT implementation because in traditional IRT methods strong interpretative weight is given to both item-factor loadings as well as item thresholds (or rather, their IRT analogues item discrimination and difficulty, respectively). Finally, the use of a full-information estimation procedure in deriving OQ item characteristics is a distinct advantage over a limited-information procedure. Estimates obtained are suboptimal for the former compared to the latter, however, in certain situations a limitation-information approach may be reasonable, or even the only one feasible. Computational complexity increases exponentially for complex models with many latent variables. In these cases a limited-information estimator may be preferable, or the only computationally feasible option. In the present study a full-information estimator was used to derive item characteristics. A limited-information estimator was used to conduct measurement invariance testing. Using a limited-information approach was computationally attractive and addressing the complexities involved with a full-information procedure were beyond the scope of this study. However, using a limited-information approach sacrifices in parameter precision that would have been obtained with a

full-information approach. The following section describes areas where future research may improve upon the present study and the limitations noted.

Future Work

In line with the preceding limitations, four specific recommendations for future research are made. These recommendations apply to the specific topic of the present study, a bifactor IRT model of the OQ, but also apply more generally to bifactor IRT and IRT methods of clinical measures. First, future (bifactor) IRT approaches may use different time-points to obtain item characteristics. Ideally, longitudinal measurement invariance is studied to provide a more precise quantification of (non-)invariance of item characteristics over time. Applying such methods to the bifactor case would be especially interesting as no such research has been done at this time. In addition to studying measurement invariance over time other relevant variables may be studied as well. In addition to gender, which was examined in the present study, clinical diagnosis, level of psychological distress, and others may warrant attention. Second, it is not possible to obtain standard IRT item or person fit statistics for a bifactor IRT model as implemented in the present study. However, indicator bivariate residuals may be used to obtain information analogous to item fit statistics. Bivariate residuals provide an indication of how well a proposed IRT model structure accounts for covariance between indicators. Many large bivariate residuals would suggest poor item fit. Third, although the absence of item threshold separation between primary and subscale domains was presented as a limitation, it is also an inherent property of the bifactor model. Where possible future researchers may compare thresholds derived from a bifactor model to thresholds derived from a unidimensional model. Such a comparison may provide insight into similarities and differences between bifactor and

unidimensional approaches given particular data. Finally, in addition to studying item characteristics using a full-information method, measurement invariance may also be studied using full-information methods. Given sufficient time and computational power, models with and without individual item-factor loading and item threshold constraints may be fit. Loglikelihood-difference testing may be used to compare such models. Because model comparisons will have to be made for every equality constraint this will likely be a computationally expensive and time-consuming procedure.

In conclusion, the future of research as presented in this study depends on the development of specialized methods software. The complexity of the measurement models and software needed to represent such models is such that they are not among widely used or accessible methods in clinical research and practice. Further dissemination of these methods in both research and practice therefore depends on those conducting such research to make their methods and findings more broadly available to clinical researchers and practitioners alike. It is hoped that methods and findings presented here contribute to such efforts.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

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

Andrich, D. (1978b). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581–594.

Beck, A. T., Ward, C. H., Mendelson, M, Mock, J, & Erbaugh, J. (1961). An inventory for measuring depression. *Archives for General Psychology*, *4*, 53–63.

Bludworth, J. L., Tracey, T. J. G., & Glidden-Tracey, C. (2010). The bilevel structure of the Outcome Questionnaire-45. *Psychological Assessment*, *22*, 350–355. doi:10.1037/a0019187

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280. doi:10.1177/014662168801200305

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Cai, L., du Toit, S. H., & Thissen, D. (n.d.). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Chen, F. F., West, S., & Sousa, K. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*, 189–225. doi:10.1207/s15327906mbr4102_5

Doucette, A., & Wolf, A. W. (2009). Questioning the measurement precision of psychotherapy research. *Psychotherapy Research*, *19*, 374–389. doi:10.1080/10503300902894422

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ...Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19. doi:10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. (n.d.). POLYBIF [Computer software]. Chicago: Center for Health Statistics, University of Illinois. Retrieved from http://www.healthstats.org/bifactor.html

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423 –436.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. London: Routledge.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement*, *68*, 695–709. doi:10.1177/0013164407313366

Lambert, M. J., & Garfield, S. L. (2004). Overview, trends, and future issues. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th edition, pp. 805–821). New York: Wiley.

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., & Al., E. (2004). *Administration and scoring manual for the OQ-45.2*. American Professional Credentialing Services.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, *3*, 249–258.

Lambert, M. J., Finch, A. M., Okiishi, J., Burlingame, G. M., McKelvey, C., & Reisinger, C. W. (1998). *Administration and scoring manual for the OQ-10.2*. American Professional Credentialing Services.

Lambert, M. J., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2003). *OQ-30.1 [Outcome Questionnaire for Adults]*. Salt Lake City, UT: OQ Measures LLC.

Lehman, A. F. (1988). A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning*, *11*, 51–62.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. MA: Addison-Wesley.

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

Muthen, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Retrieved from `http://www.statmodel.com/examples/webnote.shtml`

Muthen, L. K., & Muthen, B. O. (2010). *Mplus user's guide.* (Sixth edition). Los Angeles, CA: Muthen & Muthen. Retrieved from `http://www.statmodel.com/ugexcerpts.shtml`

Nunnally, J. C. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.

O'Neill, H. F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *Journal of Educational Research*, *18*, 234–245.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100–120. doi:10.1177/0146621605279761

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.r-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M. (2009). *Multidimensional item response theory*. Dordrecht: Springer.

Reise, S. P., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31. doi:10.1007/s11136-007-9183-7

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph*, *17*.

Samejima, F. (1996). The graded response model. In W. J. V. D. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233–247). London: Kluwer Academic Publishers.

Seelert, K. R., Hill, R. D., Rigdon, M. A., & Schwenzfeier, E. (1999). Measuring patient distress in primary care. *Family medicine*, *31*, 483–487. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10425529

Spielberger, C. D. (1983). Manual for the State Trait Anxiety Inventory STAI (Form Y). Palo Alto, CA.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577. doi:10.1007/BF02295596

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. doi:10.1177/109442810031002

Wiessman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self report. *Archives of General Psychiatry*, *33*, 1111–1115.

Yung, Y.-F., Thissen, D., & Mcleod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.