2021

THE OQ-45: THE STRUCTURE OF MENTAL HEALTH SYMPTOMS DURING PSYCHOTHERAPY

Lindsay Arader
Saint John's University, Jamaica New York

THE OQ-45: THE STRUCTURE OF MENTAL HEALTH SYMPTOMS DURING PSYCHOTHERAPY

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF ARTS

to the faculty of the

DEPARTMENT OF PSYCHOLOGY

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES

at

ST. JOHN'S UNIVERSITY
New York

by

Lindsay L. Arader

Date Submitted 4/2/2021 Date Approved 4/5/2021

Lindsay L. Arader William F. Chaplin

ABSTRACT

THE OQ-45: THE STRUCTURE OF MENTAL HEALTH SYMPTOMS DURING PSYCHOTHERAPY

Lindsay Arader

The structure of the Outcome Questionnaire-45 (OQ-45; Lambert et al., 2004b) was examined across three time points in a sample of 199 psychotherapy clients at an outpatient community clinic in the greater New York area. Five models—a one-factor model, two-factor (oblique) model, three-factor (oblique) model, two sub-factor bifactor model, and three sub-factor bifactor model—were tested at baseline, 8 weeks, and 16 weeks. The two sub-factor bifactor model fit the data best at baseline, and the three sub-factor bifactor model fit the data best at 8 weeks and 16 weeks. These results demonstrate that the OQ-45 items load onto one general distress factor across time. The extent to which the items additionally load onto the Symptom Distress, Interpersonal Relations, and Social Roles subscales of the OQ-45 changes over time.
Keywords: Outcome Questionnaire-45 structure, bifactor model, confirmatory factor analysis, psychotherapy

TABLE OF CONTENTS

LIST OF TABLES

Introduction

There is a debate among clinicians and researchers in the field of psychology regarding how to best conceptualize the major domains and structure of mental health symptoms. This debate ensues because successfully capturing mental health symptoms relies on accepting a particular theory that informs these structures. For example, in studying depression, a research-practitioner following Beck's theory of depression might find it prudent to measure symptoms related to negative automatic thinking, negative self-schemas, and faulty logic. In contrast, a research-practitioner following psychoanalytic theories of depression might be more likely to measure symptoms Freud prioritized, such as inwardly directed anger and ego-demands. However, to successfully create a comprehensive structure of mental health symptoms, it is critical that we include symptoms from multiple theoretical orientations.

Indeed, strict adherence to one theoretical orientation to inform symptom domain conceptualization is unsustainable. First, there are estimated to be over 500 schools of thought among clinicians (Prochaska & Norcross, 2018). This implies there may be over 500 ways to conceptualize symptom domains in patients. Secondly, a research-practitioner that strictly adheres to only one of these schools of thought may miss symptoms that fall outside their conceptual domain. Thus, rather than relying on theoretical orientation to construct measures of symptom domains, research-practitioners might benefit more from attempting to capture a broad range of symptoms that are not necessarily based on one strict conceptualization.

Separating the symptom measurement from the clinician's theoretical orientation allows us to measure the patient's symptoms *regardless of* the clinician's orientation.

Though the "*why*" of symptoms may be interpreted differently by clinicians of different theoretical orientations, the "*what*" remains the same; a symptom is a symptom, and it will occur objectively no matter the orientation of the clinician. The question remains, then: on what do we base our mental health symptom conceptualization, if not theoretical orientation?

Simply put: multiple sources of information are necessary to broaden our mental health symptom conceptualizations. Goldfried (2019) spoke to this broad approach method, stating:

> ...the development of any treatment package or school should be based more on reliable evidence about human functioning and the change process and less on theory — or the belief on the part of the developer that certain variables are important. Thus, evidence that can be considered as relevant to psychotherapy can come from a variety of different sources, each addressing a different question. (p. 493)

Goldfried's assertion highlights two tasks that are incumbent upon us to undertake as research-practitioners: (1) to use multiple sources in informing our mental health symptom conceptualizations, and (2) to conduct research that empirically supports the domains and structure of our conceptualizations. In the context of developing a broad outcome measure that would be sensitive to change during psychotherapy, Michael Lambert and his colleagues began undertaking this first task in 1996 (Lambert et al., 1996). Addressing the second task—finding evidence for Lambert's conceptualization of mental health symptoms domains and their structure—is the object of this paper.

The Outcome Questionnaire-45 (OQ-45)

Michael Lambert's team developed the Outcome Questionnaire-45 (OQ-45) to fill a gap in objective outcomes measures in psychotherapy. Before it's development, Froyd, Lambert, and Froyd (1996) conducted a literature review and found 1,430 different outcome measures in use across 334 psychotherapeutic outcome studies. Condensing such a vast quantity of measures was prudent for the sake of efficiency and standardization, but posed an obvious challenge: how can one singular outcome measure adequately capture all the symptoms measured in those 1,430 different instruments?

Lambert's team aimed for symptom breadth in their development of the OQ-45, which is named for the 45 outcome symptoms it measures. Test items were selected primarily to "address commonly occurring problems across a wide variety of disorders" (Lambert et al., 2004b; p. 1). The OQ-45 also sought to identify symptoms that are most likely to occur in clinical populations and impact quality of life (Lambert et al., 2004b). The total OQ-45 score reflects overall distress; higher scores indicate more disturbance for the individual (Lambert et al., 2004b).

Psychological literature from the 1980s and 90s informed the OQ-45's conceptualization of mental health symptoms, resulting in three symptom domains within the OQ-45: Symptom Distress (SD), Interpersonal Relations (IR), and Social Roles (SR). The literature behind the SD domain included a 1988 study (Regier et al., 1988) that identified the most common DSM-III-R disorders in the United States, which found that 12% of the total sample (N = 18,571) received an anxiety or affective disorder diagnosis. Substance use disorders followed as the next most common diagnosis. As such, the SD scale was loaded with items to capture symptoms of these commonly diagnosed disorders. Example items include: “I feel blue” and “I have an upset stomach” (Lambert et al., 2004b; p. 12). The IR scale was born of multiple studies that suggest positive relationships are essential to happiness, and interpersonal problems are often related to intrapersonal distress (Andrews & Witney, 1974; Beiser, 1983; Blau, 1977; Diener, 1984; Veit & Ware, 1983; Horowitz, 1979; Horowitz et al., 1988; Zatura, 1983; Horowitz et al., 1991). Example items from the IR scale include: “I am satisfied with my relationships with others” and “I am concerned with family troubles” (Lambert et al., 2004b; p. 12). Finally, the SR domain was informed by the quality of life literature included in the previous domains, as well as research that suggests that mental health symptoms often have an effect on an individual’s work and personal life (Frisch et al., 1992). As such, SR seeks to capture symptoms that may manifest in societal contexts, such as those surrounding work and leisure. Example items from the SR scale include: “I feel that I am not doing well at work/school” and “I find my work/school satisfying” (Lambert et al., 2004b; p. 12). Although these three symptom domains boast breadth, the extent to which they differ from each other has been brought into question.

Subscales within measures are often created to capture independent or partially independent sub-constructs that are validated by low correlations among the subscales. In the case of the OQ-45, researchers have found little discriminant validity among the three subscales. Conversely, significant positive correlations have emerged between these subscales across multiple different samples (Lambert et al., 1996; Lambert et al., 2004b; Umphress et al., 1997). Given these high intercorrelations, researchers have been interested in determining what, exactly, the subscales of the OQ-45 represent, if not three independent factors. One pathway toward understanding the utility of the separate subscales is to examine the theory that justifies keeping them distinct.

The model of psychotherapeutic change put forth by Howard et al. (1993) influenced the separation of symptom domains in the OQ-45. In their 1993 paper, Howard et al. reported that the recovery process in psychotherapy follows a stepwise pattern: subjective well-being changes first, followed by symptom remittance, and finally general life functioning. As such, Howard et al. (1993) recommended outcome measures that assess interpersonal and work/social functioning separately from symptom distress given their differential recovery timelines. The OQ-45 follows this model by creating separate domains for symptoms (SD), interpersonal functioning (IR), and work/social functioning (SR). Though the theoretical justification for separate subscales is compelling, the extent to which these domains actually emerge in a clinical population as measured by the OQ-45 is not currently clear.

The Structure of Mental Health Symptoms Based on the OQ-45

Empirically evaluating this three-domain structure is critical for our understanding of symptom domains in clinical populations and informing clinical practice. A structural analysis of the symptom domains represented in the OQ-45 clarifies the extent to which it is reasonable to think about these three subscales as different, albeit related, domains. In the OQ-45 Scoring and Administration Manual, Lambert asserts that the subscale scores can be used to identify specific problem areas (Lambert et al., 2004b; p. 20). He notes that consulting "scores on each of the subscales provides a dramatic illustration of poor functioning in a particular domain" (Lambert et al., 2004b; p. 22). However, he also acknowledges that "validity data do not provide strong support for the use of the OQ-45 subtest scores" (Lambert et al., 2004b; p. 22). The validity data he references have provided mixed results regarding the structural legitimacy of the three subscales.

Previous research analyzing the structure of the OQ-45 suggests that a three-factor model may not be the best fit for the data. Confirmatory factor analyses have demonstrated that one-, two-, and three-factor solutions fit the data equally well, which brings into question the validity and applicability of the separate subscales (Mueller et al., 1998). In another study, Bludworth et al. (2010) hypothesized that the OQ-45 contains one general factor and multiple unique subscale factors. Their results confirmed this hypothesis: they found that a 4-factor bilevel model, which accounts for a general distress factor, then item loadings on each of the three subscales, fit the data best. This finding suggests that our conceptualization of mental health symptoms in the clinical population as represented by the three domains of the OQ-45 may be slightly more complicated than we realize in terms of factor structure. An additional variable that has not been considered in structural analysis of the OQ-45 is time.

Structural Change Over Time

There has been little research to assess the extent to which the structure of the OQ-45 holds over time. Following the model put forth by Howard et al. (1993) regarding stepwise change for well-being, symptom distress, and life functioning, it is possible that symptom domains may actually become more clear as clients move through psychotherapy. Bludworth et al. (2010) noted that, though they identified a model fit for the data, future research was needed to evaluate these structures over time.

Past research has been mixed regarding the amount of psychotherapy sessions required to generate reliable change among patients. In their seminal 1986 study, Howard et al. determined that 50% of patients improve after eight therapy sessions (1986). Though this statistic has not been extensively replicated, it continues to drive psychotherapy recommendations. For example, the APA cites the eight session recommendation on their webpage titled *Understanding Psychotherapy and How it Works,* implying that eight sessions is the standard minimum psychotherapy length (American Psychological Association, 2012). As such, after baseline, the 8-week timepoint in therapy has been identified as a dataset of interest for this study. Additional research has been mixed regarding the number of sessions required for "recovery." Harnett et al. (2010) asserted that 14 psychotherapy sessions are required for 50% of patients to recover. Alternatively, Lambert (2016) determined that it takes closer to 18 sessions for 50% of patients to "re-enter the ranks of normal functioning" (para. 4). As such, the third timepoint of interest in this study will be 16 weeks. A primary goal of this paper will be to assess the structure of the OQ-45 across the baseline, 8-week, and 16-week timepoints.

The Present Study

In the current study, we will first conduct exploratory factor analyses to replicate previous analyses that did not support three independent factors among OQ-45 items (de Jong et al., 2007). Following this analysis, we will then move to confirmatory analysis to determine if these factors can be identified when we pre-specify them. Using confirmatory factor analyses, we will test five different models to determine which provides the best structural fit to the data, or rather, the best conceptualization of mental health symptoms as represented by the OQ-45. Our models of interest fall into two categories: (1) nested models, including a one-factor model, two correlated factor model, and three correlated factors model; and (2) non-nested models, including a bifactor model with two sub-factors (bifactor [1-2]) and a bifactor model with three sub-factors (bifactor [1-3]). We hypothesize first that the exploratory factor analyses will not yield three independent factors that map on to the subscales of the OQ-45. Second, following confirmatory factor analyses, we hypothesize that we will find evidence for distinctions among the domains of mental health symptoms specified in the OQ-45 and their relation to each other. Specifically, it is hypothesized that our results will replicate those of Bludworth et al. (2010) in finding support for a general distress factor and three separate factors, as represented by the bifactor (1-3) model. Finally, we hypothesize that the structure of the OQ-45 will change over the course of psychotherapy. Specifically, we expect that there will be clearer evidence of the three factor structure in the 8-week and 16-week data compared to the baseline data.

Methods

Participants

Data were collected from adult clients receiving psychotherapy at the St. John's Center for Psychological Services, which is a community mental health training clinic in the greater New York City area. At baseline, a total of 199 participants completed outcome measures preceding therapy between 2006 and 2019. These participants were majority female (54.3%), and ranged in age from 18 to 70 years old with a mean age of 34.1. Demographics were similar in the 8-week and 16-week samples; see Table 1 for complete demographic information of data collected at baseline, 8 weeks, and 16 weeks.

Instrument

Participants completed the OQ-45 on a bi-weekly basis during the course of psychotherapy. The OQ-45 is a 45-item outcome measure which asks participants to rate symptoms on a five-point Likert-type scale from 0 (never) to 4 (almost always) to describe how much each symptom had been experienced over the past week (Lambert et al., 2004b). The OQ-45 has a total score to represent overall psychological distress, as well as three subscales to represent different domains of mental health symptoms (SD, IR, and SR). The SD scale contains 25 items, the IR scale contains 11 items, and the SR scale contains nine items. Subscale scores are determined by summing scores within subscales, and the total score is yielded from summing scores across all items.

Analysis

We first conducted exploratory principal axis factor analysis (EFA) on the 45 items of the OQ-45 in M-Plus statistical software (Muthén & Muthén, 2009). This analysis was conducted on all items available at the three separate time points of interest in this study: baseline, 8 weeks, and 16 weeks. Individuals who responded to less than half of the items were excluded from the analysis. In the final baseline dataset, most individuals had no missing data. Of those with missing item-level data, 92% of the sample was missing less than five items. The highest number of missing item-level data was 19. Missing data descriptive statistics for each timepoint can be found in Table 2. Missing item-level data was addressed by imputation.

We created parcels (testlets) from item-level data for several reasons. Generally, on measures of mental health symptoms, item responses are highly skewed given that most respondents do *not* commonly endorse high levels of mental health symptoms. Thompson and Melancon (1996) suggest that parceling such skewed data results in more normal distributions that meet assumptions for factor analysis. In addition, after initially conducting analyses on item-level data, we observed that parcel-level results were consistently cleaner and more interpretable than item-level results. Finally, aggregating item-level data into parcels is an effective strategy to increase reliability of measurement.

All OQ-45 items were separated into 16 parcels containing two or more items. Parcels were generated within each subscale using a random number generator to cluster subscale items within parcels. More information regarding parcels can be found in Table A1 in the Appendix.

We conducted EFA on the 16 parcels at baseline, 8 weeks, and 16 weeks. Next, we used confirmatory factor analysis (CFA) to evaluate the fit of the bifactor (1-2) model, the bifactor (1-3) model, the one factor model, two correlated factors model, and three correlated factors model to the data using M-Plus statistical software (Muthén & Muthén, 2009). The CFA was conducted for each of the five models on the item-level data and 16 parcels at baseline, 8 weeks, and 16 weeks. Model fit was evaluated using the root mean square error of approximation (RMSEA), the chi-square (X^2), the comparative fit index (CFI), the standardized root mean square residual (SMSR), and Bayesian Information Criteria (BIC) difference. Of all possible indicators of fit, the RMSEA, X^2, and CFI have been shown to be least affected by estimation technique and sample size (Cangur & Ercan, 2015). BIC difference is considered a conservative and appropriate estimate of relative model fit because it penalizes complex models with numerous parameters and allows for comparison across nested and non-nested models (Kass & Raftery, 1995).

The following thresholds were considered in model fit interpretation. First, the X^2 statistic will be considered as a dichotomous indicator of fit. RMSEA fit values can be interpreted as: .05 or lower indicates close fit, .06 - .08 indicates good fit, and .10 and above indicates poor fit (Browne & Cudeck, 1993). CFI fit values between .90 and .95 may be considered marginal, above .95 indicates good fit, and below .90 indicates poor fit (Kenny, 2015). SMSR values below .08 indicate good fit (Hu & Bentler, 1999). The magnitude of difference in BIC between models is described as follows: 0 to 6 is a *positive* difference, 6 to 10 is a *strong* difference, and differences greater than 10 are *decisive* (Kass & Raftery, 1995).

Results

Descriptive statistics for the three datasets (baseline, 8 weeks, and 16 weeks) can be found in Table 1. The median score totals for the OQ-45 and each of the three subscales was within the clinical range at each timepoint. Lambert et al. (2004b) assert that the low end of the clinical range of OQ-45 scores begins at 63 for the total score, 36 for the SD score, 12 for the SR score, and 15 for the IR score. Further, for clinically reliable change to have occurred, the total score must drop by 14 points, the SD score must drop by 10 points, the SR score must drop by seven points, and the IR score must drop by eight points. Within the sample analyzed in the current study, differences between median scores across subscales and timepoints did not indicate clinically reliable change. However, median scores did decrease as patients participated in additional sessions of psychotherapy.

Table 1

Sample Descriptive Statistics

Timepoint	OQ Median (SD)	SD Median (SD)	SR Median (SD)	IR Median (SD)	Median Weeks Since First Appointment (SD)	N	Mean Age	% Female
Baseline	76.0 (22.5)	43.0 (14.5)	13.0 (4.7)	19.0 (6.4)	0	199	34.1	54.3%
8 Weeks	68.0 (23.5)	38.0 (14.6)	12.0 (5.0)	16.72 (6.6)	8.0 (1.1)	192	33.1	54.2%
16 Weeks	64.7 (24.7)	36.0 (15.2)	12.0 (5.0)	16.9 (6.7)	16.0 (1.7)	178	34.2	54.5%

Table 2

Missing Data Descriptive Statistics

Timepoint	% of sample with 0 cases of Missing Data	Minimum Missing Data Cases	Maximum Missing Data Cases	Missing Data Cases Mode
Baseline	53.3%	1	19	1
8 Weeks	77.1%	1	13	1
16 Weeks	79.8%	1	15	1

Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) values from the exploratory analyses can be found in Table 3. These values were high, indicating the variables within the samples are highly correlated with each other. In baseline item-level analyses, the initial principal components analysis resulted in 12 factors with eigenvalues greater than one (see Table 4), however, the first factor accounted for far more variance than any of the others, providing some support for a unidimensional model.

Table 3

KMO Values Across Timepoints

Sample Time Point	Item-Level KMO	Parcel-Level KMO
Baseline	0.82	0.92
8 Weeks	0.88	0.92
16 Weeks	0.89	0.93

Visual inspection of the scree plot for the baseline item-level data provides some support for a three-factor solution given that eigenvalues begin to plateau around 1 after the third factor. The first four eigenvalues were 10.49, 3.43, 2.16, and 1.99 (see Table 4). With baseline parcel-level data, the initial principal components analysis resulted in three eigenvalues greater than 1 (7.16, 1.50, and 1.06), which provides further evidence for a potential unidimensional model. Eigenvalue patterns remained consistent across the item- and parcel-level analyses with both 8- and 16-week data, wherein the first factor accounted for far more variance than any of the subsequent factors. All eigenvalues greater than one for each sample can be found in Table 4.

Principal Axis Factoring analysis followed by an oblique (Oblimin; Delta = 0) rotation of three factors was conducted on item-level and parcel-level data at baseline, 8 weeks, and 16 weeks, which resulted in six total pattern matrices. Inspection of the pattern matrix and interfactor correlation table for item-level data at baseline (see Table 5) revealed a lack of a clear third factor and did not reflect any clear adherence to distinct SD, IR, and SR subscales. Notably, the correlation between factors 1 and 2 was moderate ($r = .39$), while the correlation of factors 1 and 2 with factor 3 were very weak ($r = -.11$, $r = -.08$; see Table 6). Secondly, factor loadings were relatively low across items, and several items did not load above .5 on any factor. Finally, the groupings of items within factors did not follow the purported structure of the OQ-45. For example, the first factor included nine items from the SD scale, nine items from the IR scale, and two items from the SR scale. Thus, exploratory factor analysis using baseline item-level data reveals a shared variance between factors that do not fall out in accordance with the proposed structure of SD, IR, and SR.

Table 4

Eigenvalues Across Datasets and Timepoints

Timepoint/ Dataset	Eigenvalues > 1.0											
Baseline Item-Level	10.49	3.43	2.16	1.99	1.88	1.72	1.42	1.37	1.26	1.11	1.08	1.06
Baseline Parcel-Level	7.16	1.5	1.06									
8-Week Item-Level	13.4	2.95	2.32	1.95	1.85	1.7	1.44	1.28	1.07	1.04	1.01	
8-Week Parcel-Level	8.38	1.43	1.01									
16-Week Item-Level	14.95	3.05	2.12	1.9	1.76	1.68	1.33	1.26	1.11	1.02	1.00	
16-Week Parcel Level	9.12	1.18										

Table 5

Baseline Item-Level Pattern Matrix

OQ Subscale	OQ-45 Item	Factor 1	Factor 2	Factor 3
IR	Item 20	**0.75**	-0.2	0.01
IR	Item 43	**0.75**	0.03	-0.02
SD	Item 13	**0.71**	-0.02	-0.13
SD	Item 31	**0.7**	-0.06	-0.06
SD	Item 24	**0.69**	0.02	-0.1
SR	Item 21	**0.56**	-0.02	-0.22
IR	Item 18	**0.54**	0.15	0.1
IR	Item 37	**0.54**	-0.11	0.23
SD	Item 42	**0.51**	0.33	-0.03
SD	Item 15	**0.51**	0.22	-0.27
IR	Item 1	**0.49**	0.01	-0.1
SR	Item 12	**0.47**	-0.02	-0.05
SD	Item 6	**0.44**	0.29	0.02
IR	Item 17	**0.44**	0.04	0.36
SD	Item 3	**0.42**	0.18	-0.11
SD	Item 23	**0.37**	0.24	-0.11
IR	Item 16	**0.33**	0.07	0.32
SD	Item 8	**0.32**	0.15	-0.2
IR	Item 19	**0.29**	0.13	0.01
SD	Item 29	-0.15	**0.68**	-0.09
SD	Item 10	0.08	**0.59**	-0.17
SR	Item 4	0.07	**0.58**	0.13
SD	Item 2	0.06	**0.56**	-0.05
SD	Item 36	0.11	**0.54**	-0.07
SD	Item 33	0.1	**0.53**	-0.08
SD	Item 40	0.21	**0.51**	-0.15
SD	Item 27	-0.16	**0.51**	-0.07
SD	Item 35	-0.13	**0.5**	0
SD	Item 25	0.04	**0.49**	-0.16
SD	Item 34	-0.1	**0.49**	0.2
SD	Item 9	0.16	**0.46**	-0.12

SR	Item 38	0.01	**0.45**	0.09
SD	Item 41	0.14	**0.45**	0.06
SD	Item 45	0.11	**0.42**	0
SD	Item 22	0.25	**0.4**	0.17
SR	Item 28	0.16	**0.4**	0.13
SR	Item 39	0.03	**0.37**	-0.09
SR	Item 14	-0.12	**0.33**	0.32
IR	Item 30	0.26	**0.32**	-0.08
SD	Item 5	0.27	**0.3**	0.04
SR	Item 44	0.04	0.26	**-0.43**
SD	Item 11	0.14	0.01	**-0.38**
IR	Item 26	0.14	0.16	**-0.32**
IR	Item 7	0.3	0.16	**0.31**
SR	Item 32	0.11	0.12	**-0.31**

Table 6

Baseline Item-Level Factor Correlation Matrix

Factor	1	2	3
1	1.00	.39	-.11
2	.39	1.00	-.08
3	-.11	-.08	1.00

The pattern matrix resulting from the baseline, parcel-level data reveals higher factor loadings within each factor and a more consistent adherence to the SD, IR, SR structure between factors (see Table 7), which may be a result of the parcels controlling for item-level idiosyncrasies. However, the factor correlation matrix for the baseline parceled data (see Table 8) reveals substantial interfactor correlations, suggesting a common influence on each factor. Thus, though our statistical software diligently provided us with the three factors we asked for, the data do not necessarily provide support for a clear distinction between factors.

Table 7

Baseline Parcel-Level Pattern Matrix

Subscale Parcel	Factor 1	Factor 2	Factor 3
SD Parcel 4	**.84**	-.08	.07
SD Parcel 8	**.73**	.10	-.08
SD Parcel 6	**.72**	-.07	.00
SD Parcel 7	**.62**	.16	.14
SD Parcel 2	**.61**	.11	.04
SD Parcel 9	**.61**	.04	.09
SD Parcel 5	**.56**	-.07	.28
IR Parcel 2	-.06	**.73**	.01
IR Parcel 1	.30	**.63**	-.07
IR Parcel 3	.14	**.55**	.11
IR Parcel 4	-.09	**.55**	.07
SD Parcel 3	.39	**.47**	-.05
SR Parcel 2	.05	-.06	**.80**
SR Parcel 1	-.02	.08	**.71**
SD Parcel 1	.21	.14	**.53**
SR Parcel 3	.14	.23	**.25**

Table 8

Baseline Parcel-Level Factor Correlation Matrix

Factor	1	2	3
1	1.00	.52	.65
2	.52	1.00	.43
3	.65	.43	1.00

Given the cleaner structure resulting from parceled data analysis, remaining item-level analyses will be featured in the Appendix and will not be discussed moving forward (see Tables A2 and A3).

Principal Axis Factoring analysis followed by an oblique (Oblimin; Delta = 0) rotation of three factors on 8-week parceled data revealed a structure that adheres more closely to Lambert's three-subscale conceptualization than baseline parceled data (see Table 9). In the 8-week parceled data, the factor loadings suggests that the parcels within each factor adhered exactly to Lambert's proposed SD, IR, and SR structure. The parceled data thus provides clearer support for a three factor solution that is consistent with the domains referenced within the development of the OQ-45.

Notably, the 8-week parceled data EFA results continue to reveal a lack of clear distinction between factors. The factor correlation matrix (see Table 10) reveals that the three factors correlate highly with each other, again suggesting the influence of a commonality between factors.

In the pattern matrix from the 16-week dataset (Table 11), cross loadings were more common across factors than at baseline or 8 weeks. However, none of the parcels loaded primarily on the third factor. The correlation matrix (Table 12) shows a high correlation between factors 1 and 2 (r = .68) and essentially nonexistent correlations between factors 1 and 2 with factor 3 (r = .03, r = .02). Overwhelmingly, the third factor does not emerge in the 16-week dataset.

Table 9

8 Week Parcel-Level Pattern Matrix

Subscale Parcel	Factor 1	Factor 2	Factor 3
SD Parcel 4	**.93**	-.14	.02
SD Parcel 6	**.80**	-.12	-.06
SD Parcel 7	**.75**	.15	.07
SD Parcel 9	**.71**	.14	-.07
SD Parcel 8	**.69**	.18	.09
SD Parcel 5	**.65**	.10	-.18
SD Parcel 1	**.64**	-.01	-.20
SD Parcel 2	**.59**	.21	-.07
SD Parcel 3	**.44**	.35	-.05
IR Parcel 1	.25	**.65**	-.04
IR Parcel 2	.01	**.64**	-.24
IR Parcel 3	.03	**.45**	-.41
IR Parcel 4	.26	**.41**	-.04
SR Parcel 1	.21	.07	**-.72**
SR Parcel 3	-.09	.18	**-.64**
SR Parcel 2	.42	-.19	**-.60**

Table 10

8 Week Parcel-Level Factor Correlation Matrix

Factor	1	2	3
1	1.00	.47	-.54
2	.47	1.00	-.45
3	-.54	-.45	1.00

Table 11

16 Week Parcel-Level Pattern Matrix

Subscale Parcel	Factor 1	Factor 2	Factor 3
SD Parcel 4	**.94**	-.12	-.05
SD Parcel 6	**.86**	-.10	.01
SD Parcel 9	**.80**	.07	.08
SD Parcel 7	**.78**	.08	-.05
SD Parcel 8	**.78**	.04	-.09
SD Parcel 1	**.74**	.05	.15
SD Parcel 5	**.64**	.20	-.17
SR Parcel 2	**.63**	.13	.34
SD Parcel 2	**.50**	.35	-.16
IR Parcel 2	-.16	**.91**	.05
IR Parcel 3	.11	**.68**	-.04
IR Parcel 1	.31	**.55**	-.29

SD Parcel 3	.30	**.51**	.05
SR Parcel 1	.39	**.49**	.37
IR Parcel 4	.30	**.37**	-.29
SR Parcel 3	.27	**.36**	.26

Table 12

16 Week Parcel-Level Factor Correlation Matrix

Factor	1	2	3
1	1.00	.68	.03
2	.68	1.00	.02
3	.03	.02	1.00

Given the cleaner structures that emerged from parceled data in our exploratory analyses, we focused our confirmatory analysis on parceled data. Model fit determined by confirmatory factor analysis on item-level data can be found in the Appendix (Table A4).

Table 13

CFA Results on Parceled Data

Model	Timepoint	df	X^2	CFI	RMSEA	90% CI	SRMR	Sample Size Adjusted BIC
One-factor model	Baseline	104	316.58	.85	.10	[.09 - .11]	.07	12,476.61
	8 Weeks	104	461.56	.82	.13	[.12 - .15]	.07	11,476.53
	16 Weeks	104	425.43	.85	.13	[.12 - .15]	.06	10,432.18
Two-factor model (oblique)	Baseline	103	286.18	.87	.10	[.08 - .11]	.07	12,448.33
	8 Weeks	103	361.69	.87	.11	[.10 - .13]	.06	11,378.76
	16 Weeks	103	369.26	.87	.12	[.11 - .13]	.06	10,421.91
Three-factor model (oblique)	Baseline	101	220.07	.92	.08	[.06 - .09]	.06	12,386.47
	8 Weeks	101	289.76	.90	.10	[.09 - .11]	.06	11,311.01
	16 Weeks	101	298.87	.90	.11	[.09 - .12]	.05	10,311.67
Bifactor (1-2) Model	Baseline	88	167.37	.96	.07	[.05 - .08]	.05	12,361.32
	8 Weeks	88	288.14	.90	.11	[.10 - .12]	.05	11,336.56
	16 Weeks	88	296.94	.90	.12	[.10 - .13]	.05	10,335.93
Bifactor (1-3) Model	Baseline*	NA	NA	NA	NA	NA	NA	NA
	8 Weeks	88	235.42	.93	.09	[.08 - .11]	.05	11,283.84
	16 Weeks	88	254.61	.92	.10	[.09 - .12]	.05	10,293.59

* = Model did not converge.

Table 13 shows goodness-of-fit indices for each model tested at baseline, 8 weeks, and 16 weeks with parceled data. At baseline, the bifactor 1-2 model provided the best fit to the data (CFI = .96, RMSEA = .07, SRMR = .05). The bifactor 1-3 model did not converge at baseline, which is likely because the third factor in the EFA of baseline parceled data had an eigenvalue of 1.064 (see Table 4) suggesting that this third factor accounted for barely any more variance than a single parcel. Because the third factor just barely emerges in the baseline parceled data, once general variance is accounted for in the bifactor 1-3 model, there may not be enough remaining distinguishable variance to generate three separate factors. At 8 weeks, the bifactor 1-3 model provided the best fit to the data (CFI = .93, RMSEA = .09, SRMR = .05). The bifactor 1-2, one-factor, two-factor, and three-factor models showed comparatively more ill-fitting indices of fit. At the 16-week timepoint, the bifactor 1-3 model again provided the best fit to the data (CFI = .92, RMSEA = .10, SRMR = .05). Within each of the five models, contrary to our hypotheses, the CFI, RMSEA, and SRMR values for the 16-week data were either the same or worse than these fit indices for the 8-week data.

Relative fit of both nested and non-nested models was evaluated using change in Bayesian Information Criteria (BIC), described by Kass and Raftery (1995). BIC differences are reported in Tables 14, 15, and 16. Notably, BIC differences between models are all greater than 10, indicating decisive differences between model fit across the board. At baseline, reported in Table 14, the bifactor 1-2 model resulted in a lower BIC than the one-factor, two-factor, or three-factor models, indicating superior fit of the bifactor 1-2 over competing models.

Table 14

BIC Differences Between Models at Baseline

Model	1	2	3	4	5
1. Bifactor 1-2					
2. Bifactor 1-3	NA				
3. One-Factor Model	115.29	NA			
4. Two-Factor Model	87.01	NA	-28.28		
5. Three-Factor Model	25.15	NA	-90.14	-61.86	

Table 15

BIC Differences Between Models at 8 Weeks

Model	1	2	3	4	5
1. Bifactor 1-2					
2. Bifactor 1-3	-52.72				
3. One-Factor Model	139.97	192.69			
4. Two-Factor Model	42.2	94.92	-97.77		
5. Three-Factor Model	-25.55	27.17	-165.52	-67.75	

Table 16

BIC Differences Between Models at 16 Weeks

Model	1	2	3	4	5
1. Bifactor 1-2					
2. Bifactor 1-3	-42.34				
3. One-Factor Model	96.25	138.59			
4. Two-Factor Model	85.98	128.32	-10.27		
5. Three-Factor Model	-24.26	18.08	-120.51	-110.24	

Due to lack of convergence, we cannot explicitly compare the bifactor 1-2 and bifactor 1-3 models at baseline, but recognize that the nonconvergence of the bifactor 1-3 model may in itself be an indicator of the extent to which that model fits the data at baseline. At the 8-week timepoint, the magnitude of BIC differences reported in table 15 reveal that the bifactor 1-3 model fit the data best when compared to the bifactor 1-2, one-, two-, and three-factor models. In the 16-week dataset, the bifactor 1-3 model again emerged decisively victorious per BIC differences between models (see Table 16).

Discussion

The OQ-45 was created on a theoretical basis to capture separate symptom domains labeled Symptom Distress (SD), Interpersonal Relations (IR), and Social Roles (SR). Scoring the OQ-45 yields a total distress score and three subscale scores "to identify specific problem areas" (Lambert et al., 2004b; p. 20). The extent to which these three domains emerge in clinical populations distinctly and over time is not clear. Past research has found some support for a correlated three-factor structure of the OQ-45, but such results have not been extensively replicated or examined over time.

In this study, we found the bifactor 1-2 model fit the data best at baseline, and the bifactor 1-3 model fit the data best at 8 weeks and 16 weeks, indicating shared variance among Lambert's three theoretical domains–SD, IR, and SR–over time. Our model fit results have implications in several areas, including, but not limited to, the conceptualization of mental health symptoms among psychotherapy patients, symptom structure change over time, and clinical interpretation of OQ-45 scores.

Mental Health Symptom Structure

The overarching goal of this study was to identify a latent structure of mental health symptoms within a clinical sample via OQ-45 factor interpretation. Theoretically, we expected the symptom structure to mirror Lambert's SD, IR, and SR domains. This hypothesis reflects the theory behind the creation of the OQ-45 that a typical psychotherapy patient might have difficulties that cluster around distressing symptoms, problems with relationships, and problems with 'roles,' such as being an employee or student.

In line with previous research, we found bifactor models tended to fit the data best at all three timepoints (Bludworth et al., 2010). Notably, the bifactor model fit does not fully support Lambert's proposed three-domain structure. Rather, these three domains are best characterized by their high intercorrelations. This shared variance among the three domains supports the hypothesis of general psychopathology, conceptualized as *p*, which has been described as analogous to *g* in intelligence testing (Caspi et al., 2014). In intelligence testing, bifactor models often provide the best fit to the data by accounting for shared variance across domains before identifying any remaining clusters of unique variance (Cucina & Byle, 2017). Our results are also reminiscent of the findings that informed the creation of the Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008). The MMPI-2-RF contains 338 items, reduced from the MMPI-2, which required participants to score 567 items across 10 clinical scales. In the reduction process, researchers determined that a latent factor coined "demoralization" was "responsible for the excessive intercorrelations between the Clinical Scales" (Ben-Porath, 2012; p. 45). In the MMPI-2-RF, demoralization exists as its own subscale and fewer items make up the remaining subscales in attempt to ensure each subscale measures a relatively distinct construct. Within the OQ-45, a demoralization-type factor may explain the shared variance across the SD, IR, and SR, domains. We might ask ourselves: if we remove this common construct from the OQ-45, will we paint a clearer picture of the typical psychotherapy patient? Conversely, might this removal put us at risk of missing the full story of a psychotherapy patient? Psychometrically, it might be satisfying for our analyses to yield clean, distinct factors. In reality, it might behoove us to recognize that human beings' experiences in the world are neither clean nor distinct.

Symptom Structure Change Over Time

This is the first known study to examine OQ-45 model fit over time. The structural analysis of the OQ-45 in the current study sought to examine how closely Lambert's three-domain structure reflects an average clinical population moving through psychotherapy. Our results indicate that psychotherapy patients at 8 weeks most closely follow Lambert's theoretical construct, whereas data from patients at baseline and 16-weeks appear to deviate from his three-domain construct.

At baseline, the EFA resulted in three highly correlated factors that did not perfectly map onto Lambert's proposed SD, IR, and SR, factors. The best model fit per the CFA at baseline was the bifactor 1-2 model, providing little support for three distinct SD, IR, and SR factors at baseline. Rather, once shared variance has been explained, two factors representing SD and an IR+SR combination remain. This means that, in the current study, baseline patients tended to score more consistently within, rather than between, SD and IR+SR.

At the 8-week timepoint, the EFA suggested that three factors, though highly correlated, emerged in perfect consistency with Lambert's proposed factors. In the CFA, the bifactor 1-3 model provided the best fit to the data, which is consistent with the concept of shared variance among symptoms due to general pathology, yet sufficient intra-factor variance for SD, IR, and SR to emerge as relatively distinct. In other words, 8-week patients tended to have more consistent scores within each of the three domains.

The EFA at 16 weeks was the least consistent with Lambert's three symptom domains. At 16 weeks, the EFA resulted in two factors, in which SD clustered together, and IR and SR were much more highly correlated with each other than at any other time point. However, the bifactor 1-3 model still provided the best fit to the data at 16 weeks over all other models, providing some evidence for the emergence of symptom domains representing SD, IR, and SR. Similar to 8-week psychotherapy patients, 16-week psychotherapy patients were more likely to have consistent scores within SD, IR, and SR than they were at baseline.

Our model results suggest that the constructs the OQ-45 measures relate differently to each other over the course of psychotherapy. Though there may be several competing explanations for this result, one possible avenue is the phase model of change (Howard et al., 1993). The phase model of change theorizes that psychotherapy patients improve in three phases—remoralization, remediation, and rehabilitation—with the successful completion of each phase allowing ascendance into the next. According to Howard et al., (1993), remoralization involves building trust, facilitating emotional vulnerability, defining internal problems (as opposed to external), and the strengthening of therapeutic alliance. Next, the remediation phase is simply focused on resolving patients' symptoms or life problems. Finally, the rehabilitation phase serves to help the individual take on major-life roles and unlearn "troublesome, maladaptive, longstanding patterns" (Howard et al., 1993; p. 680). Though it is not clear that SD, IR, and SR map perfectly into the three phases, it is possible that their differential relations over time occur as a result of stepwise psychotherapeutic change consistent with the phase model.

Clinical Interpretation of the OQ-45

The results of this study indicate that cautious interpretation of the SD, IR, and SR subscales may become more warranted as the patient moves through psychotherapy. At baseline, the bifactor 1-2 model provided the best fit to the data, suggesting that Lambert's three constructs are not necessarily present. At 8 weeks and 16 weeks, the bifactor 1-3 model fit was superior to the bifactor 1-2, one-, two-, and three-factor model fit, which indicates that the SD, IR, and SR domains do genuinely emerge in the data once shared variance has been explained. The 8-week and 16-week results provide some support for Lambert's recommendation to use the subscale scores to "identify specific problem areas" (2004b; p. 20). However, the baseline results suggest that interpretation of the three subscales at baseline should be cautious at the very least. Ultimately, subscale scores may be best used to provide appropriate treatment direction when interpreted in the context of information gathered from other sources.

Limitations and Future Directions

The extent to which our results can be extrapolated to generally describe mental health symptoms is an obvious limitation to this study. In the current study, we have only evaluated symptom domains that were measured by the OQ-45, and lack any sufficient understanding of what we did *not* capture.

Although symptom distress, interpersonal relations, and social roles certainly cover a broad range of human functioning, there are several domains relevant to improvement in therapy that are not included in the OQ-45. For example, self-efficacy is a well-established predictor of wellbeing across pathologies, and is notably absent in the OQ-45 (Bandura, 1986). Furthermore, the OQ-45 may lack culturally sensitive symptom domains. For example, Hwang et al. (2006) explain that effective CBT with Chinese Americans requires the therapist to understand cultural nuances in expressing distress. In Chinese cultures, it may be more common for individuals to express distress somatically as opposed to through emotional language. As such, an outcome measure like the OQ-45 may not adequately capture true symptoms in individuals from non-Western cultures. In addition, the OQ-45 may flag distress in areas that are not actually relevant to the client's true symptoms. For example, a client with challenges associated with gender identity or sexual orientation may very well experience symptom distress or difficulty with interpersonal relations or social roles. However, those domains alone would be insufficient to guide treatment planning given that they might fail to wholly capture the internal experience of an LGBTQ+ individual navigating their own identity and self-concept. Ultimately, the OQ-45 may be most useful when administered concurrently with supplemental measures.

Furthermore, we recognize the ethical challenges inherent to appropriately identifying mental health symptoms. Broadly, it is not clear that capturing symptoms is of use to the patient, despite scientists' best efforts to identify objective symptom descriptors. Some might even argue that symptom descriptions enable the unnecessary medicalization of perceived disturbing behaviors, which ultimately achieves social control rather than symptom remittance (Szasz, 2011). The most obvious example to this point is that of same-sex sexual orientation, which, until recently, was conceptualized as a mental illness requiring treatment (Pillard, 2009). This is not to say that symptom measurement tools like the OQ-45 are definitively detrimental, but rather to acknowledge that, at best, symptom descriptions in general are nebulous, and at worst, iatrogenic. Despite this dilemma, symptom measurement tools will indeed continue to dominate the field of psychology. Given this trajectory, the least we can do as scientist-practitioners is question the validity of our tools. As such, we hope to again emphasize that the OQ-45 is likely insufficient to capture the whole story of a patient moving through psychotherapy.

Secondly, our analyses were slightly limited by sample size. With a larger sample size, more robust invariance testing would have been possible. Additionally, in future research, a larger sample size might allow for exploratory analyses among demographic variables that contribute to OQ-45 results.

Overall, this study has contributed new information about differential change over time in psychotherapy as represented by the three subscales of the OQ-45. Future research should seek to explore the mechanisms that might further explain structural differences in clinical data across these timepoints.

Appendix

Table A1

Parcels

Parcel	OQ-45 Item	Item Content
SD Parcel 1	22	I have difficulty concentrating
	41	I have trouble falling asleep or staying asleep
	5	I blame myself for things
SD Parcel 2	31	I am satisfied with my life
	33	I feel that something bad is going to happen
	35	I feel afraid of open spaces, of driving, or of being on buses, subways, and so forth
SD Parcel 3	13	I am a happy person
	24	I like myself
	8	I have thoughts of ending my life
SD Parcel 4	10	I feel fearful
	40	I feel something is wrong with my mind
	45	I have headaches
SD Parcel 5	23	I feel hopeless about the future
	25	Disturbing thoughts come into my mind that I cannot get rid of
	29	My heart pounds too much
SD Parcel 6	11	After heavy drinking, I need a drink the next morning to get going (if you do not drink, mark "never")
	2	I tire quickly
	9	I feel weak

SD Parcel 7	27	I have an upset stomach
	42	I feel blue
	6	I feel irritated
SD Parcel 8	15	I feel worthless
	34	I have sore muscles
SD Parcel 9	3	I feel no interest in things
	36	I feel nervous
IR Parcel 1	18	I feel lonely
	30	I have trouble getting along with friends and close acquaintances
	7	I feel unhappy in my marriage/significant relationship
IR Parcel 2	1	I get along well with others
	20	I feel loved and wanted
	37	I feel my love relationships are full and complete
IR Parcel 3	19	I have frequent arguments
	26	I feel annoyed by people who criticize my drinking or drug use (if not applicable, mark "never")
	43	I am satisfied with my relationships with others
IR Parcel 4	16	I am concerned about family troubles
	17	I have an unfulfilling sex life
SR Parcel 1	21	I enjoy my spare time
	38	I feel that I am not doing well at work/school
	39	I have too many disagreements at work/school
SR Parcel 2	28	I am not working/studying as well as I used to
	32	I have trouble at work/school because of drinking or

		drug use (if not applicable, mark "never")
	4	I feel stressed at work/school
SR Parcel 3	12	I find my work/school satisfying
	14	I work/study too much
	44	I feel angry enough at work/school to do something I might regret

Table A2

8 Week Item-Level Pattern Matrix

Subscale	OQ-45 Item	Factor 1	Factor 2	Factor 3
SD	Item 10	**0.81**	0.09	-0.11
SD	Item 36	**0.72**	0.02	0.00
SD	Item 33	**0.70**	-0.03	-0.03
SD	Item 45	**0.67**	0.26	0.20
SD	Item 9	**0.67**	-0.14	-0.06
SD	Item 5	**0.64**	-0.09	-0.13
SD	Item 42	**0.64**	-0.22	-0.04
SD	Item 27	**0.62**	0.15	0.23
SD	Item 2	**0.61**	-0.11	0.06
SD	Item 15	**0.60**	-0.38	-0.24
SD	Item 6	**0.58**	-0.21	0.12
SD	Item 40	**0.57**	-0.25	-0.18
SD	Item 29	**0.51**	0.09	0.30

SR	Item 4	**0.51**	-0.12	0.03
SD	Item 23	**0.50**	-0.45	-0.18
SR	Item 28	**0.50**	-0.11	0.20
SD	Item 25	**0.49**	-0.05	-0.15
SD	Item 3	**0.48**	-0.42	0.03
SD	Item 34	**0.46**	0.06	0.35
SD	Item 41	**0.45**	0.00	0.14
IR	Item 7	**0.41**	-0.06	0.13
SD	Item 35	**0.40**	0.03	-0.05
IR	Item 16	**0.39**	-0.13	0.05
SD	Item 22	**0.37**	-0.37	0.05
SR	Item 44	**0.37**	-0.06	0.01
SR	Item 38	**0.34**	-0.29	0.07
SR	Item 14	**0.30**	0.17	0.19
IR	Item 17	**0.28**	-0.20	0.12
SD	Item 8	**0.20**	-0.19	-0.14
SD	Item 31	0.07	**-0.83**	-0.02
IR	Item 20	-0.03	**-0.78**	0.20
SD	Item 13	0.20	**-0.72**	-0.06
IR	Item 43	0.04	**-0.66**	0.29
SD	Item 24	0.32	**-0.61**	-0.11
SR	Item 21	0.31	**-0.59**	-0.02

SR	Item 12	0.01	**-0.59**	0.07
IR	Item 37	0.05	**-0.56**	0.09
IR	Item 1	-0.13	**-0.49**	0.35
IR	Item 18	0.38	**-0.43**	0.09
IR	Item 30	0.28	-0.15	**0.52**
IR	Item 26	0.03	-0.08	**0.41**
SR	Item 39	0.35	-0.05	**0.39**
IR	Item 19	0.33	-0.12	**0.38**
SR	Item 32	0.10	-0.06	**0.37**
SD	Item 11	-0.03	-0.06	**0.11**

Table A3

16 Week Item-Level Pattern Matrix

Subscale	OQ-45 Item	Factor 1	Factor 2	Factor 3
SD	Item 10	**0.81**	0.09	-0.11
SD	Item 36	**0.72**	0.02	0.00
SD	Item 33	**0.70**	-0.03	-0.03
SD	Item 45	**0.67**	0.26	0.20
SD	Item 9	**0.67**	-0.14	-0.06
SD	Item 5	**0.64**	-0.09	-0.13
SD	Item 42	**0.64**	-0.22	-0.04
SD	Item 27	**0.62**	0.15	0.23

SD	Item 2	**0.61**	-0.11	0.06
SD	Item 15	**0.60**	-0.38	-0.24
SD	Item 6	**0.58**	-0.21	0.12
SD	Item 40	**0.57**	-0.25	-0.18
SD	Item 29	**0.51**	0.09	0.30
SR	Item 4	**0.51**	-0.12	0.03
SD	Item 23	**0.50**	-0.45	-0.18
SR	Item 28	**0.50**	-0.11	0.20
SD	Item 25	**0.49**	-0.05	-0.15
SD	Item 3	**0.48**	-0.42	0.03
SD	Item 34	**0.46**	0.06	0.35
SD	Item 41	**0.45**	0.00	0.14
IR	Item 7	**0.41**	-0.06	0.13
SD	Item 35	**0.40**	0.03	-0.05
IR	Item 16	**0.39**	-0.13	0.05
SD	Item 22	**0.37**	-0.37	0.05
SR	Item 44	**0.37**	-0.06	0.01
SR	Item 38	**0.34**	-0.29	0.07
SR	Item 14	**0.30**	0.17	0.19
IR	Item 17	**0.28**	-0.20	0.12
SD	Item 8	**0.20**	-0.19	-0.14
SD	Item 31	0.07	**-0.83**	-0.02
IR	Item 20	-0.03	**-0.78**	0.20

SD	Item 13	0.20	**-0.72**	-0.06
IR	Item 43	0.04	**-0.66**	0.29
SD	Item 24	0.32	**-0.61**	-0.11
SR	Item 21	0.31	**-0.59**	-0.02
SR	Item 12	0.01	**-0.59**	0.07
IR	Item 37	0.05	**-0.56**	0.09
IR	Item 1	-0.13	**-0.49**	0.35
IR	Item 18	0.38	**-0.43**	0.09
IR	Item 30	0.28	-0.15	**0.52**
IR	Item 26	0.03	-0.08	**0.41**
SR	Item 39	0.35	-0.05	**0.39**
IR	Item 19	0.33	-0.12	**0.38**
SR	Item 32	0.10	-0.06	**0.37**
SD	Item 11	-0.03	-0.06	**0.11**

Table A4

CFA Results on Item Level Data

Model	Timepoint	df	X^2	CFI	RMSEA	90% CI	SRMR	Sample Size Adjusted BIC
One-factor model	Baseline	945	2,390.46	.54	.09	[.08 - .09]	.09	24,629.30
	8 Weeks	945	2,547.08	.60	0.09	[.09 - .10]	.09	21,452.44
	16 Weeks	945	2,554.93	.62	.10	[.09 - .10]	.09	19,189.80
Two-factor model (oblique)	Baseline*	NA	NA	NA	NA	NA	NA	NA
	8 Weeks	944	2,544.65	.60	.09	[.09 - .10]	.09	21,452.10
	16 Weeks	944	2,541.33	.63	.10	[.09 - .10]	.09	19,178.21
Three-factor model (oblique)	Baseline	942	2,296.99	.57	.09	[.08 - .09]	.09	24,542.19
	8 Weeks	942	2,540.40	.61	.09	[.09 - .10]	.09	21,452.03
	16 Weeks	942	2,537.67	.63	.10	[.09 - .10]	.09	19,178.58
Bifactor 1-2 Model	Baseline*	NA	NA	NA	NA	NA	NA	NA
	8 Weeks	900	2,045.99	0.72	.08	[.08 - .09]	.08	21,045.38
	16 Weeks	900	2,057.56	.73	.09	[.08 - .09]	.09	18,783.10
Bifactor 1-3 Model	Baseline	900	1,904.96	.68	.08	[.07 - .08]	.08	24,239.21
	8 Weeks*	NA	NA	NA	NA	NA	NA	NA
	16 Weeks*	NA	NA	NA	NA	NA	NA	NA

*NA = Model did not converge.

References

American Psychological Association. (2012, November 1). *Understanding psychotherapy and how it works*. American Psychological Association.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, *4*(3), 359-373.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of marketing research*, *19*(4), 425-430.

Ben-Porath, Y. S. (2012). *Interpreting the mmpi-2-rf*. U of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota multiphasic personality inventory-2 restructured form*. Minneapolis, MN: University of Minnesota Press.

Bludworth, J. L., Tracey, T. J. G., & Glidden-Tracey, C. (2010). The bilevel structure of the Outcome Questionnaire–45. Psychological Assessment, 22(2), 350–355. doi:10.1037/a0019187

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.

Boring, E. G. (1964). Cognitive dissonance: Its use in science. *Science*, *145*(3633), 680-685.

Cangur, S., & Ercan, I. (2015). Comparison of model fit indices used in structural equation modeling under multivariate normality. *Journal of Modern Applied Statistical Methods*, *14*(1), 14.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... & Moffitt, T. E. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders?. *Clinical psychological science*, *2*(2), 119-137.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, *5*(3), 27.

de Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *14*(4), 288-301.

Froyd, J. E., Lambert, M. J., & Froyd, J. D. (1996). A review of practices of psychotherapy outcome measurement. *Journal of Mental Health*, *5*(1), 11-16.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*(5), 639-662.

Goldfried, M. R. (1982). *Converging themes in psychotherapy: Trends in psychodynamic, humanistic, and behavioral practice*. Springer Publishing Company.

Goldfried, M. (2019). Obtaining Consensus in Psychotherapy: What Holds Us Back*?* *American Psychologists, 74(4)*, 484 – 496.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. Clinical Psychology: science and practice, 9(3), 329-343.

Harnett, P., O'Donovan, A., & Lambert, M. J. (2010). The dose response relationship in psychotherapy: Implications for social policy. *Clinical Psychologist*, *14*(2), 39-44.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American psychologist*, *41*(2), 159.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of consulting and clinical psychology*, *61*(4), 678.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American psychologist*, *51*(10), 1059.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Hwang, W. C., Wood, J. J., Lin, K. M., & Cheung, F. (2006). Cognitive-behavioral therapy with Chinese Americans: Research, theory, and clinical practice. *Cognitive and Behavioral Practice*, *13*(4), 293-303.

Kadera, S. W., Lambert, M. J., & Andrews, A. A. (1996). How much therapy is really enough?: A session-by-session analysis of the psychotherapy dose-effect relationship. *The Journal of psychotherapy practice and research*, *5*(2), 132.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Kenny, D. A. (2015). Measuring model fit.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community mental health journal*, *4*(6), 443-453.

Lambert, M. J. (2013). Bergin and Garfield's handbook of psychotherapy and behavior change. John Wiley & Sons.

Lambert, M. J. (2016, February). Top 10 things learned after two decades of tracking client treatment progress. [Web article]. Retrieved from http://www.societyforpsychotherapy.org/top-10-things-learned-after-two-decades-of-tracking-client-treatment-progress

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: An International Journal of Theory and Practice*, *3*(4), 249-258.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004a). The Outcome Questionnaire. In M. E. Maruish (Ed.), The use of psychological testing for treatment planning and outcome assessment: Vol. 3: Instruments for adults (3rd ed., pp. 191-234). Mahwah, NJ: Lawrence Erlbaum.

Lambert, M. J., Morton, J. J., Hatfield, D. R., Harmon, C., Hamilton, S., Reid, R. C., & Burlingame, G. M. (2004b). Administration and scoring manual for the OQ-45.2

(Outcome Questionnaire). *Orem, UT: American Professional Credentialing Services*.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment*, *70*(2), 248-262.

Muthén, L. K., & Muthén, B. O. (2009). Mplus. *Statistical analysis with latent variables. User's guide*, *7*.

Owen, J., Adelson, J., Budge, S., Wampold, B., Kopta, M., Minami, T., & Miller, S. (2015). Trajectories of change in psychotherapy. *Journal of Clinical Psychology*, *71*(9), 817-827.

Pillard, R. (2009). From disorder to dystonia: DSM-II and DSM-III. *Journal of Gay & Lesbian Mental Health*, *13*(2), 82-86.

Prochaska, J. O., & Norcross, J. C. (2018). *Systems of psychotherapy: A transtheoretical analysis*. Oxford University Press.

Regier, D. A., Boyd, J. H., Burke, J. D., Rae, D. S., Myers, J. K., Kramer, M., ... & Locke, B. Z. (1988). One-month prevalence of mental disorders in the United States: Based on five epidemiologic catchment area sites. *Archives of General Psychiatry*, *45*(11), 977-986.

Umphress, V. J., Lambert, M. J., Smart, D. W., Barlow, S. H., & Clouse, G. (1997). Concurrent and construct validity of the Outcome Questionnaire. *Journal of Psychoeducational Assessment*, *15*(1), 40-55.

Szasz, T. (2011). The myth of mental illness: 50 years later. *The Psychiatrist*, *35*(5), 179-182.

Wells, M. G., Burlingame, G. M., Lambert, M. J., Hoag, M. J., & Hope, C. A. (1996). Conceptualization and measurement of patient change during psychotherapy: Development of the Outcome Questionnaire and Youth Outcome Questionnaire. *Psychotherapy: Theory, Research, Practice, Training*, *33*(2), 275.

Vita

Name	*Lindsay Arader*
Baccalaureate Degree	*Bachelor of Arts, Hamilton College, Clinton, NY* *Major: Psychology & Hispanic Studies*
Date Graduated	*May, 2016*