Alternative Models of the Outcome Questionnaire-45

Amber Gayle Thalmayer

February 2014

Summary

The Outcome Questionnaire-45 (OQ) reliably quantifies level of psychological functioning and change during treatment. The three subscales, however, are not well validated. Could alternative scales, based on personality dimensions or other psychological problems scales better explain patterns of response? In Study 1, the intended structure and four alternative models were compared using EFA and CFA in random thirds of a community clinic intake sample (N = 1,822). Oblique and bi-level models were compared. Preferred models were tested for stability in samples from later time points. In Study 2, the models were compared in a non-clinical sample (N = 589). Most bi-level models provided adequate fit per standards previously established for the Outcome Questionnaire-45. The seven-factor model provided better fit than any yet reported for this inventory.

Keywords: Outcome Questionnaire-45, psychological assessment, confirmatory factor analysis

Alternative Models of the Outcome Questionnaire-45

Psychotherapy is sought for a wide range of problems, and trust in its efficacy has led to increasing parity in insurance coverage. But the majority of clients receive services not based on treatment of a specific disorder (Shafran, et al., 2009). How can the efficacy of general counseling be quantified? The current study explores the structure of the Outcome Questionnaire-45 (OQ), an instrument designed to meet the needs of diverse clinics. It considers ways to refine interpretation of OQ scores to maximize utility and validity.

The Outcome Questionnaire-45: History, Validity, Structure

Limited resources and diverse clients and issues put constraints on measurement, but clinics, therapists, and insurers need information about the extent of clients' problems, and the rate of improvement during treatment. The OQ was developed as a brief measure of symptoms across a range of disorders and syndromes, including stress-related illness, for baseline screening and to capture change. Administration takes only a few minutes, and allows for a quantitative assessment of treatment effectiveness.

Items for the OQ were rationally selected to assess common symptoms that affect quality of life, and to align with the *DSM*. This led to three content areas: Symptom Distress (SD), Interpersonal Relations, and Social Role functioning. The administration manual (Lambert et al., 2004) explains that SD, the domain of intrapsychic problems, is largest because affective disorders are the most commonly diagnosed; because recent literature indicates that symptoms of anxiety and depression covary, such items comingle on the scale. But while large comorbidity studies show that anxiety and depression often co-occur and can be conceptualized as both belonging to an "internalizing" domain (e.g. Krueger & Markon, 2006), such studies also report bi-furcation into depression and anxiety subcategories (Krueger & Markon, 2006). Combining these tendencies indicates potential multidimensionality. Further, only because substance abuse was the next most common diagnosis, such items were included on the same scale (Lambert et al., 2004). But the same studies that make the case for internalizing tendencies make the case for a separate domain of externalizing tendencies, including substance abuse (e.g. Krueger & Markon, 2006).

The Interpersonal Relations (IR) scale was designed to capture relationship difficulties, due to their well-established association with well-being and the frequency of interpersonal problems as a focus of therapy. Items were derived from marriage and family therapy literature. The Social Role (SR) scale assesses problems and conflict in work and school settings, the rationale being that psychological problems and role performance affect each other bi-directionally (Lambert et al., 2004).

People with more severe mental health issues consistently receive higher total scores (TS). The correlation of TS, SD, and to a lesser extent IR and SR with clinical severity has been established in English (Umphress, Lambert, Smart, Barlow, & Clouse, 1997; de Beurs et al., 2005) and in translation (von Bergen & de la Parra, 2002; De Jong et al., 2007; Haug et al., 2004; Li & Luo, 2009). The OQ is also sensitive to change: TS reliably diminishes as symptom severity decreases, in outpatient and inpatient samples (von Bergen & de la Parra, 2002; de Beurs et al., 2005; Doerfler, Addis, & Moran, 2002; Haug et al., 2004; Talley & Clack, 2006).

There is a lack of evidence, however, for the discriminant validity of the subscales. The three are highly correlated across samples, such that SD seems indistinguishable from TS (Umphress et al., 1997). IR and SR fail to preferentially correlate with other measures from their domains, and to capture distinct aspects of functioning (Doerfler et al., 2002; Hess, Rohlfing, Hardy, Glidden-Tracey, & Tracey, 2010; Umphress et al., 1997).

Since the widespread implementation and translation of the OQ, studies using confirmatory factor analysis (CFA) have not supported its intended structure (De Jong et al., 2007; Mueller, Lambert, & Burlingame, 1998; Wennberg et al., 2010; cf. Li & Luo, 2009). Marginal fit was observed in an Italian version, however, for a bi-level model that better matches clinical usage of the OQ -- items were allowed to load on their intended subscale and secondarily on a TS scale (Coco et al., 2008). Bludworth, Tracey and Glidden-Tracey (2010) found support for this bi-level model in an American sample. Both observed, however, that items loaded more highly on TS than on intended subscale.

Authors have used a variety of criteria and proposed different solutions to the difficulties of structural fit for the OQ. At the item level, some suggest dropping substance abuse because items detract from fit and are highly skewed (Coco et al., 2008), though removal could reduce utility. Item 14 ‘I work/study too much’ fails to correlate with other items across samples, and scores are often higher in non-clinical samples (De Jong et al., 2007). But no studies to date have explored alternative latent factor models beyond OQ subscales (e.g., collapsing SR and IR, or adding TS).

Personality and the OQ

It is worth considering whether scores on the OQ are driven by individual differences in temperament/personality that are observable from infancy (Rothbart, 2007), have a strong genetic component (Bouchard, 2004), and have been shown to underlie disorders (Clark, 2005). While the OQ was not designed to measure these domains, patterns of responses are likely influenced by these robust attributes.

Most OQ items (particularly SD), may measure Negative Affectivity, (Neuroticism in Big Five inventories), the tendency to experience more or less activation of internalizing negative emotions (Saucier, 2009), related to individual differences in reactivity of the amygdala and limbic system (Whittle, Allen, Lubman, & Yücel, 2006). Personality psychologists observe most disorders, particularly affective ones, to occur more often in those high on this attribute (Clark, 2005).

Disinhibition (which underlies Big Five Conscientiousness) has strong associations with externalizing tendencies (Clark, 2005) assessed by OQ substance abuse, conflict, and impulsivity items. Positive Affectivity (an aspect of Extraversion) may influence scores on reverse-keyed SD items regarding satisfaction and pleasure. Affiliation (part of Agreeableness), the tendency to get along with others, should predict responses to conflict items.

An alternative hypothesis would be that other problem domains underlie OQ scores. Responses might fall into broader patterns than those conceptualized by the OQ intended subscales, perhaps relating to basic domains of internalizing (depression, anxiety, phobias) versus externalizing (substance abuse, conflict) tendencies. Alternatively, patterns of response might be more granular, relating to more specific problems. An initial rational sorting of items by the author, based on specific content, indicated the potential for scales relating to: internalizing negative affectivity, positive affectivity, somatic complaints, family and spouse stress, anger, substance use, and functioning in work and school.

Goals

The OQ was designed to efficiently assess overall functioning and change, and to screen for suicidality and violence. The demand for such a measure is illustrated by its wide adoption and translation. Analyses across samples, however, have failed to establish support for its intended structure. While recent studies suggest a bi-level model, the uncertainty of scale labels and the need to test for consistency over time have been noted (Bludworth et al., 2010).

Here, a rigorous comparison of the OQs intended structure to plausible alternatives, across time, explores how to best use and interpret scores. In the first study, the intended structure is compared to alternatives (three- and four-factor personality models, internalizing/externalizing, and a seven-factor problem model; see Table 1) in separate portions of a large clinical dataset. Refined, preferred models are compared in samples from later therapy sessions. A second study compares models in a non-clinical, student sample.

Study 1

Method

Participants. Two thousand one hundred clients attended at least one session at a couples and family therapy clinic 2006 - 2011 and completed the OQ. Clients were 57% female, 85% white, and had an average age of 34 (SD = 10.5). Half came from households with less than $25K per year income. Half attended individual therapy, and half couple or family sessions.

Materials. OQ items are answered on a 5-point scale from ‘never’ to ‘almost always’. Nine are reverse keyed. The OQ was usually administered before a client’s intake, third, fifth, and tenth sessions, and every ten thereafter.

Scales for alternative models were rationally constructed. For personality models, choices were driven by three- and four- factor models of temperament (Clark, 2005; Rothbart, 2007). Gerard Saucier, an expert on personality structure, was consulted. Although five or six factor models of personality structure (e.g. John & Srivastava, 1999; Saucier, 2009) provide more comprehensive coverage of personality variation, the OQ does not include items relating to all domains (e.g. Openness). Subscale placement for the internalizing externalizing model was determined with reference to Kreuger and Markon (2006). Seven-factor scales were constructed as described above.

The intake sample (N =1,822) TS mean (69.7, SD = 23.4) was lower than the manual's outpatient mean (83, SD = 22, N = 342; Lambert et al., 2004). Table 2 displays psychometric properties of scales. In addition to Cronbach's alpha and average interitem correlation, the variance of interitem *r* is included as an indicator of unidimensionality (ideally, correlations between items measuring a single construct range .15 to .50 [Clark & Watson, 1995].)

The time-five dataset included 614 responses completed prior to a fifth or sixth session (TS M = 62.11, SD = 23.12). The time-10 set included 361 responses completed prior to a 10th, 11th, or 12th session (TS M = 62.57, SD = 23.36). In all datasets, the items were moderately correlated with one another in the expected directions with the exception of item 14 (uncorrelated with most). Skew and kurtosis were significant for most items, sometimes due to low base rate (most answered "never" to substance, suicide, violence, and phobia items), but only two substance use items had SD < .6 and skew and kurtosis values likely to be problematic per Kline (2011; SI > 3, KI > 10).

Analyses. The structure of the OQ was first assessed using exploratory factor analysis (oblique rotation, listwise deletion) on a random third (N = 624) of the intake data. Factor scores for two-, three-, four-, and seven-factor solutions, plus solutions with increasing numbers of factors until interpretability was lost (up to 10), were compared to scale scores for *a priori* models using Pearson correlation. Empirically derived factors were matched, based on dominant content, to *a priori* factors. Z-score transformed correlations of matched factors were averaged to provide approximate fit between hypothesized and observed models.

Secondly, structural models and a baseline TS model were compared using CFA (Mplus 7) in the second random third of the data (N = 624). Because the data did not meet the assumption of multivariate normality (Mardia's coefficient multivariate skew = 27,673, p < .01;

kurtosis = 49, p < .01), robust maximum likelihood estimation was used and adjusted chi-square values reported (as in Bludworth et al., 2010; Coco et al., 2008). Preferred models were tested in the last random third of intake data (N = 629) and time-five and -10 datasets.

In addition to adjusted chi-square, a mix of indices (SRMR, RMSEA, AIC) evaluating different aspects of fit were examined. One comparative fit index (CFI) tested variance explained compared to a null model and another (TLI$_1$) assessed improvement over the TS model. Hu and Bentler (1999) argue that a good-fitting model should meet several criteria, e.g., CFI > .95, RMSEA < .06, SRMR < .08. Previous OQ studies indicate that such standards are unlikely to be met. Research on the use of CFA with well-validated personality inventories calls into question the likelihood that responses on a measure like the OQ can meet traditional standards, because factors are both meaningfully distinct and interrelated -- unsuited to independent clusters models (Hopwood & Donnellan, 2010) -- and due to item-level analysis (Kline, 2011). Bludworth et al. (2010) adjusted *a priori* standards per Marsh et al. (2005), emphasizing RMSEA and SMSR (cutoffs .08), de-emphasizing incremental fit, and anticipating non-significant chi-square. Here, fit was assessed per standards and in comparison to previous OQ studies.

Results

Exploratory Factor Analysis (EFA). The observed three-factor structure was compared to OQ intended structure and the three-factor personality model. Table 3 presents correlations between observed and hypothesized factors. The first factor was interpretable as SD or Negative Affectivity (NA), and correlated highly with both (.98). The second factor (most reverse-keyed items) was less interpretable as IR than Positive Affectivity (PA; -.88 vs. -.93). The third factor, with content related to arguments, anger, and substance use, was less interpretable as SR than Disinhibition (D; .78 vs. .82). The average absolute correlation between observed factor scores and content-matched OQ scales (.92 after r to z transformation for averaging, then back to r) did not differ significantly from that of the personality model (.94). Five items differed between observed and *a priori* personality: 19 did not load over .25 on any scale; item 1 (get along with others) loaded on D rather than PA; items 3 (lack of interest), 12 (dissatisfaction with work/school) and 18 (lonely), all loaded most highly on a different factor than intended, but with high secondary loadings on *a priori* scale. Item 1 was relocated, and item 14 was added to D to allow for comparison between models. Other original choices were retained to avoid overfit to this portion of the dataset.

The observed four-factor model was difficulty to interpret – after a large NA scale was a group of PA and Affiliation-like items, perhaps interpretable as "sociability". Next, the three substance abuse items plus 19 (frequent arguments), then three conflict items plus 12. Ten items loaded first on a scale other than hypothesized. In five cases, the secondary loading was on *a priori* scale. In the other five it was not, but no face-valid updates to the personality model were indicated. The average fit of four-factor solution factors to personality scales was still high (.91).

Average fit of two-factor solution factors to internalizing and externalizing scales was weaker (.74). While the first factor encompassed internalizing tendencies, the second did not emphasize externalizing -- no items identified as belonging to the domain loaded most highly on the second observed factor.

The initial seven-factor subscales were difficult to match to the seven-factor EFA solution in terms of primary content. A process of extracting additional factors until they became uninterpretable was used to explore the structure. The fit of models with increasing numbers of factors are reported in Table 4. A ten-factor solution was interpretable as a maximally elaborated model of problems. A large factor of 'Depressive Thinking' (29.5% of variance) was followed by factors interpretable as ‘Relationship Malaise’, ‘Substance Abuse’, ‘Work/School Adjustment’, ‘Family Trouble’, ‘Somatic Depression’, ‘Anxiety’, ‘Positive Emotionality’, ‘Conflict’, and item 14. Many changes implied here matched logical alternatives identified *a priori*. The seven-factor model was revised, named as above, except to avoid scales with fewer than three items, relationship and family troubles ($r = .264$, $p < .001$) were combined and item 14 was added to ‘Work/School Adjustment’ ($r = .036$, ns; in both cases combinations were made on rational rather than empirical grounds.). Somatic Depression items were relocated to Depressive Thinking and Anxiety, based on high secondary loadings. Average correlation between observed factors and analogous revised scales was .88.

Post-hoc parallel analysis (comparing average eigenvalues for datasets of this size and number of variables generated by online utility [Patil, Singh, Mishra, & Donavan, 2007] to unrotated PCA eigenvalues) suggested retaining eight factors. This result supports the seven-factor model, as the eight-factor exploratory model was basically this seven plus item 14 alone.

Confirmatory Factor Analysis (CFA). CFA fit statistics appear in Table 5. All models were run first with factors allowed to correlate, secondly bi-level including TS (as in Coco et al., 2008; Bludworth et al., 2010). All χ^2 values were significant. All SRMR values were under .08. No model had CFI greater than .90, though the bi-level seven-factor model was close. Most bi-level models had RMSEA values under .06, and the seven-factor had values indicating “close fit” (.05 ≥). More factors improved fit across alternatives, and bi-level models fit better than oblique. The three-factor personality model appeared slightly superior to three-factor intended. Comparing the personality models, EFA indicated a slight advantage for three-factors, but CFA evidence was inconclusive. The seven-factor model had a clear advantage over internalizing-externalizing, which was dropped from further analyses.

TS and bi-level intended, personality, and seven-factor models were compared in the last random third of intake data. Results were highly similar: The seven-factor model approached good fit, followed by the four-factor personality model.

The seven-factor model also demonstrated best fit in time-five and -10 samples. In the time-five set, there was no clear advantage for any other model. In time-10, the four-factor personality model had a slight advantage over three-factors, which had a slight advantage over intended structure.

Discussion

The intended structure of the OQ was compared to alternative models of psychological problems and models derived from personality psychology in three random samples of intake session responses, and in samples from later time points. While no models had ‘good fit’ (e.g. Kline, 2011), all provided acceptable fit per *a priori* indices, better than previously reported for the OQ (e.g. Bludworth et al., 2010). Best fit was observed for a seven-factor model of psychological problems.

A limitation is that many clients provided data at all time points. Longitudinal analyses in this sample indicate little average change (< half a point per week; Thalmayer & Baune, 2013). The stability of scores may support the use of the OQ for personality assessment, per the Appendix, but it weakens the significance of the convergence of models across time.

Study 2

Method

Participants and Procedure. Undergraduate students in introductory psychology and linguistics courses (N = 589) completed surveys Fall 2011as part of a half-hour online survey in exchange for course credit. The sample was 64% female and 75% Caucasian (9% Asian, 5% African American, 2% Native American, 10% "other"; average age = 19.5, SD = 2.2)

Materials. Descriptive statistics for the OQ scales are in Table 2. (The 48-item Questionnaire Big Six was also administered – see appendix for correlations and OQ items usable to estimate personality.)

While many items appeared skewed or kurtotic, none had values likely to be problematic (Kline, 2011). TS mean (M = 60.73, SD = 21.14) was higher than the manual's undergraduate norms (42 to 51; Lambert et al., 2004), and similar to the clinical sample. The clinical sample scored more problematically than the student sample on 31 items. Perhaps in part due to the age difference , the student sample scored more problematically on substance use, work violence, work/school stress, and phobias than the clinical sample.

Analyses. CFA in Mplus 7 (robust maximum likelihood estimation) was used to compare baseline TS and intended structure with preferred alternative models from Study 1.

Results

CFA fit indices for alternative models of OQ structure are reported in the bottom of Table 5. Again, indices suggested best fit for the four-and seven-factor models.

Overall Discussion

The goal of Study 2 was to test models of OQ structure from Study 1 in a non-clinical sample. Fit was highly similar to that of the clinical sample.

As a general measure of psychological functioning and change, OQ TS has criterion validity, and the need for such a measure is illustrated by its wide adoption. Lack of unidimensionality does not preclude it from functioning as a liner combination of psychological problems, outside assumptions of classical test theory for internal consistency (although this could likely be improved by balancing reverse keyed items and removing item 14). The intended subscale structure, however, has not been supported by validity or confirmatory studies. The current study went beyond previous analyses by considering alternative subscale models -- personality dimensions, and more or less elaborated models of problems. Because the OQ is used over time, models were tested in sets of responses from later in therapy.

Fit of models across samples was similar. A four-factor personality model fit better than the intended structure, providing some support for the hypothesis that dimensions of personality attributes underlie scores. Robust individual differences, well-mapped by personality psychologists, likely influence responses on broad psychological inventories.

Best fit (better than any previously reported for the OQ) was observed for a seven-factor model of psychological problems, shaped rationally and using EFA. The domains are easy to interpret and make best use of the length of the OQ. Differential changes on subscales could better inform clinicians. Future work could assess the comparative utility of this scoring system.

References

von Bergen, A., & de la Parra, G. (2002). OQ-45.2, an Outcome Questionnaire for monitoring change in psychotherapy: Adaptation, validation and indications for its application and interpretation. *Terapia Psicológica, 20,* 161-176.

de Beurs, E., den Hollander-Gijsman, M., Buwalda, V., Trijsburg, W., & Zitman, F. (2005). The Outcome Questionnaire (OQ-45): Measuring psychiatric symptoms and interpersonal functioning. *Psycholoog, 40,* 393-400.

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

Bouchard, T. J (2004). Genetic influence on human psychological traits. *Current Directions in Psychological Science, 15*, 148-151.

Clark**,** L. A**.** (2005). Temperament as a unifying basis for personality and psychopathology. *Journal of Abnormal Psychology, 114, 505-521.* doi: 10.1037/0021-843X.114.4.505

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. doi:10.1037/1040-3590.7.3.309

Coco, G. L., Chiappelli, M., Luca, B., Gullo, S., Prestano, C., & Lambert, M. J. (2008) The factorial structure of the Outcome Questionnaire-45: A study with an Italian sample. *Clinical Psychology & Psychotherapy, 15*, 418-423.

Doerfler, L. A., Addis, M. E., & Moran, P. W. (2002). Evaluating mental health outcomes in an inpatient setting: Convergent and divergent validity of the OQ-45 and BASIS-32. *The Journal of Behavioral Health Services & Research, 29*, 394-403.

Haug, S., Puschner, B., Lambert, M. J., & Kordy, H. (2004). Assessment of change in psychotherapy with the German version of the Outcome Questionnaire (OQ-45.2). *Zeitschrift für Differentielle und Diagnostische Psychologie, 25,* 141-151.

Hess, T. R., Rohlfing, J. E., Hardy, A. O., Glidden-Tracey, C., & Tracey, T. J. G. (2010) An examination of the "interpersonalness" of the Outcome Questionnaire. *Assessment,17,* 396-399.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*, 332-346.

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.

John, O. P. & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A. & John, O. P. (Eds.), *Handbook of Personality: Theory and Research.* NY: The Guilford Press.

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

Kline, R. B. (2011). Principles and practice of structural equation modeling. New York: Guilford Press.

Krueger, R.F. & Markon, K. E. (2006). Reinterpreting comorbidity: A model–based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2,* 111-133.

Lambert, M. J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., Shimokawa, K. Christopherson, C., & Burlingame, G. M. (2004). *Administration and scoring manual for* *the OQ-45.2 Outcome Questionnaire.* Salt Lake City, UT: American Professional Credentialing Services, L.L.C.

Li, Y. & Luo, H. (2009). The reliability and validity of the Outcome Questionnaire-Chinese version. *Chinese Mental Health Journal, 23,* 105-107.

Marsh, H. W., Hau, K., & Grayson, D. (2005) Goodness of fit in structural equation models. In Maydeu-Olivares, A., & McArdle, J. J. (Eds) Contemporary psychometrics. Mahwah, N.J: Lawrence Erlbaum.

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*, 248–262.

Patil, V. H., Singh, S. N., Mishra, S. & Donavan, D. T. (2007). "Parallel Analysis Engine to Aid Determining Number of Factors to Retain [Computer software]. Available from http://smishra.faculty.ku.edu/parallelengine.htm.

Rothbart, M.K. (2007). Temperament, development, and personality. *Current Directions in Psychological Science, 16,* 207-212.

Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *Journal of Personality, 77*, 1577-1614.

Shafran, R., Clark, D. M., Ehlers, A., Garety, P. A., Salkovskis, P. M., Fairburn, C. G., Williams, J. M. G., & Wilson, G. T. (2009). Mind the gap: Improving the dissemination of CBT. *Behaviour Research and Therapy, 47,* 902-909. doi: 10.1016/j.bbr.2011.03.031

Talley, J. E., & Clack, R. J. (2006). Use of the Outcome Questionnaire 45.2 with a university population. *Journal of College Student Psychotherapy, 20,* 5-15.

Thalmayer, A. G. & Baune, N. (2013). Client predictors of therapy usage and outcome. Unpublished manuscript.

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 Psychoedacational Assessment, 15,* 40-55.

Wennberg, P., Philips B., & De Jong, K. (2010) The Swedish version of the Outcome Questionnaire (OQ-45): Reliability and factor structure in a substance abuse sample. *Psychology and Psychotherapy: Theory, Research and Practice, 83*, 325-329.

Whittle, S., Allen, N. B., Lubman, D.I., & Yücel, M. (2006). The neurobiological basis of temperament: Towards a better understanding of psychopathology. *Neuroscience & Biobehavioral Reviews, 30*, 511-525.

Table 1

Alternative Models of Outcome Questionnaire-45 Structure

Scales	Items*
Intended	
Symptom Distress	2, 3, 5, 6, 8, 9, 10, 11, 13R, 15, 22, 23, 24R, 25, 27, 29, 31R, 33, 34, 35, 36, 40, 41, 42, 45
Interpersonal Relations	1R, 7, 16, 17, 18, 19, 20R, 26, 30, 37R, 43R
Social Role	4, 12R, 14, 21R, 28, 32, 38, 39, 44
Three factor personality	
Negative Affectivity	2, 4, 5, 6, 8, 9, 10, 15, 16, 18, 22, 23, 25, 27, 28, 29, 33, 34, 35, 36, 38, 40, 41, 42, 45
Positive Affectivity	3, 7, 12R, 13R, 17, 20R, 21R, 24R, 31R, 37R, 43R
Disinhibition	1R, 11, 14, 19, 26, 30, 32, 39, 44
Four-factor personality	
Negative Affectivity	*as above, without 16, 18*
Positive Affectivity	3, 2R, 13R, 21R, 24R, 31R
Disinhibition	11, 14, 26, 32, 39, 44
Affiliation	1R, 7, 16, 17, 18, 19, 20R, 30, 37R, 43R
Internalizing/Externalizing Tendencies**	
Internalizing	2, 3, 4, 5, 6, 7, 8, 9, 10, 12R, 13R, 15, 16, 18, 20R, 21R, 22, 23, 24R, 27, 28, 29, 31R, 33, 34, 35, 36, 37R, 38, 40, 41, 42, 43R, 45,
Externalizing	1R, 11, 19, 26, 30, 32, 39, 44
Seven-factor Psychological Problems Model	
Depressive Thinking	3, 5, 8, 9, 10, 15, 23, 40, 42
Positive Emotionality	12R, 13R, 21R, 24R, 31R, 43R
Anxiety	25, 27, 29, 33, 34, 35, 36, 41, 45
Work/School Adjustment	2, 4, 22, 28, 38
Family/Relationship Stress	7, 16, 17, 18, 19, 20R, 37R
Conflict	1R, 6, 30, 39, 44
Substance Abuse	11, 26, 32

* Table with item text available from author.

** 3 items excluded from this model.

Table 2

Descriptive Statistics for OQ Scales in Clinical and Student Samples

				Mean	variance
Scale (number items)	Mean	SD	α	interitem r	interitem r
Total Score (45)	69.79	23.37	.94	.247	.019
	60.73	*21.14*	*.93*	*.240*	*.020*
Intended subscales					
Symptom Distress (25)	1.60	.62	.93	.335	.016
	1.33	*.51*	*.91*	*.293*	*.016*
Interpersonal Relations (11)	1.64	.53	.80	.260	.024
	1.33	*.57*	*.76*	*.241*	*.023*
Social Role (9)	1.34	.58	.66	.203	.023
	1.38	*.45*	*.64*	*.173*	*.028*
Personality subscales					
Negative Affectivity (24)	1.72	.64	.92	.337	.009
	1.50	*.52*	*.90*	*.28*	*.014*
Positive Affectivity (6)	1.58	.74	.84	.463	.006
	2.80	*.64*	*.83*	*.46*	*.015*
Affiliation (10)	1.74	.62	.81	.299	.020
	2.76	*.55*	*.76*	*.25*	*.023*
Disinhibition (5)	.62	.43	.54	.195	.023
	.65	*.62*	*.77*	*.42*	*.012*
OQ seven-factor problems scales					
Depressive Thinking (9)	1.66	.76	.89	.472	.005
	1.28	*.64*	*.87*	*.424*	*.010*
Positive Affectivity (6)	1.56	.75	.84	.478	.008
	1.19	*.66*	*.85*	*.481*	*.014*
Anxiety (9)	1.50	.68	.81	.328	.006
	1.34	*.59*	*.80*	*.306*	*.008*
Work/School (5)	1.94	.65	.68	.262	.040
	2.02	*.53*	*.58*	*.188*	*.033*
Relationship Stress (7)	2.08	.74	.79	.353	.014
	1.45	*.66*	*.74*	*.291*	*.024*
Conflict (5)	1.08	.55	.71	.324	.007
	1.04	*.54*	*.51*	*.231*	*.038*
Substance abuse (3)	.24	.47	.65	.387	.004
	.59	*.72*	*.73*	*.509*	*.003*

Note. Student sample (N=511-588) results italicized. Clinical sample, N = 1630 – 1810. Subscale means divided by number of items for average score on 0-4 scale.

Table 3

Correlations Between EFA Factors and Intended and Hypothesized OQ Scales

Factor	Subscales						
			Two factors				
	INT	EXT					
F1 of 2	**93***	66*					
F2 of 2	-74*	**-23***					
			Three factors				
	SD	IR	SR	NA_3	PA_3	D_3	
F1 of 3	**98***	61*	67*	**98***	67*	47*	
F2 of 3	-68*	**-88***	-39*	-66*	**-93***	-25*	
F3 of 3	54*	43*	**78***	51*	51*	**82***	
			Four factors				
	NA_4	AF_4	D_4	PA_4			
F1 of 4	**99***	57*	34*	67*			
F2 of 4	-63*	**-88***	-10*	-79*			
F3 of 4	30*	32*	**86***	11			
F4 of 4	-46*	-25*	-38*	**-66***			
			Seven factors				
	DT	FRS	SA	PA	WSA	Con	Anx
F1 of 7	**94***	51*	19*	64*	52*	53*	85*
F2 of 7	-54*	**-95***	-12*	-66*	-34*	-25*	-37*
F3 of 7	35*	28*	**95***	31*	24*	38*	30*
F4 of 7	-62*	-32*	-15*	**-82***	-39*	-74*	-36*
F5 of 7	52*	-20*	.09	45*	**88***	45*	49*
F6 of 7	-13*	04	11*	-25*	17*	**57***	11
F7 of 7	56*	47*	11	43*	54*	45*	**76***

Note. N = 530. Decimal points removed for readability. Expected matches bolded. INT = internalizing, EXT = externalizing; SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role. NA = Negative Affectivity, PA = Positive Affectivity, D = Disinhibition; AF = Affiliation. DT = Depressive Thinking, FRS = Family/Relationship Stress, SA =Substance Abuse, WSA = Work/School Adjustment, Con = Conflict, Anx = Anxiety.

* p < .01.

Table 4

EFA Fit Statistics in First Random Third

Factors	df	χ^2	CFI	RMSEA	SRMR	AIC
2	901	3,350	.786	.066	.055	69,494
3	858	2,789	.832	.060	.048	69,018
4	816	2,395	.862	.056	.041	68,708
7	696	1,465	.933	.042*	.027	68,018
8	658	1,274	.946	.039*	.024	67,903
9	621	1,117	.957	.036*	.022	67,820
10	585	995	.964	.034*	.021	67,771

Note. N = 624

* Value not significantly higher (p < .05) than .05.

Table 5

Summary of CFA Fit Indices for Alternative Models for OQ structure

Model	df	adj. χ^2	CFI	TLI$_1$	RMSEA	SRMR	AIC
Second Random Third (N = 569)							
Baseline: TS	945	3,790	.711	-	.069	.068	70,177
Intended (oblique)	942	3,426	.748	.124	.065	.069	69,772
Intended bi-level	897	2,569	.830	.382	.055	.053	68,872
Three-factor Personality (oblique)	942	3,151	.776	.219	.061	.062	69,457
Three-factor Personality bi-level	897	2,565	.831	.392	.055	.051	68,847
Four-factor Personality (oblique)	939	3,100	.781	.236	.061	.064	69,392
Four-factor Personality bi-level	894	2,548	.832	.379	.054	.052	68,813
Int./Externalizing (oblique)	944	3,566	.734	.076	.067	.066	69,912
Int./Externalizing bi-level	899	2,742	.813	.312	.057	.052	69,048
Seven-factor (oblique)	924	2,559	.834	.412	.053*	.059	68,814
Seven-factor bi-level	880	2,148	.871	.521	.048*	.046	68,398
Third Random Third (N = 629)							
Baseline: TS	945	3,938	.697	-	.071	.071	72,074
Intended bi-level	897	2,618	.818	.395	.056	.052	70,621
Three-factor Personality bi-level	897	2,692	.809	.369	.058	.056	70,637
Four-factor Personality bi-level	894	2,480	.840	.440	.053*	.049	70,446
Seven-factor bi-level	880	2,129	.874	.552	.048*	.046	70,065
Time Five (N = 681)							
Baseline: TS	945	4,248	.728	-	.072	.067	70,813
Intended bi-level	897	2,855	.839	.376	.057	.053	69,289
Three-factor Personality bi-level	897	2,930	.833	.352	.058	.052	69,355
Four-factor Personality bi-level	894	2,867	.838	.370	.057	.050	69,251
Seven-factor bi-level	880	2,269	.886	.549	.048*	.044	68,640
Time 10 (N = 392)							
Baseline: TS	945	3,001	.709	-	.074	.073	40,736
Intended bi-level	897	2,118	.827	.376	.059	.059	39,788
Three-factor Personality bi-level	897	2,082	.832	.394	.058	.057	39,744
Four-factor Personality bi-level	894	2,016	.841	.424	.057	.054	39,660
Seven-factor Problems bi-level	880	1,699	.884	.573	.049*	.049	39,333
Student Sample (N = 589)							
Baseline: TS	945	4,071	.636	-	.075	.080	64,302
Intended bi-level	897	3,241	.727	.211	.067	.065	63,090
Three-factor Personality bi-level	897	2,686	.792	.398	.058	.061	62,705
Four-factor Personality bi-level	894	2,555	.807	.439	.056	.059	62,555
Seven-factor bi-level	880	2,372	.826	.488	.054*	.055	62,301

Note. All adjusted χ^2 values p < .01. CFI = comparative fit index; TLI_1 = Tucker-Lewis index comparing model to TS; RMSEA = root mean square error of approximation; SRMSR = standardized root mean square residual. AIC = Akaikes information criteria.

* Value not significantly higher (p < .05) than .05.