



2017-05-01

The Development of a Short Form of the Clinically Adaptive Multidimensional Outcome Survey

Peter William Sanders
Brigham Young University

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Counseling Psychology Commons](#)

BYU ScholarsArchive Citation

Sanders, Peter William, "The Development of a Short Form of the Clinically Adaptive Multidimensional Outcome Survey" (2017). *All Theses and Dissertations*. 6462.

<https://scholarsarchive.byu.edu/etd/6462>

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

The Development of a Short Form of the Clinically Adaptive Multidimensional Outcome Survey

Peter William Sanders

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

Doctor of Philosophy

P. Scott Richards, Chair
Brent D. Slife
G. E. Kawika Allen
Tim Smith
Richard R Sudweeks

Department of Counseling Psychology and Special Education

Brigham Young University

Copyright © 2017 Peter William Sanders

All Rights Reserved

ABSTRACT

The Development of a Short Form of the Clinically Adaptive Multidimensional Outcome Survey

Peter William Sanders

Department of Counseling Psychology and Special Education, BYU

Doctor of Philosophy

The Evidence-Based Practice (EBP) movement has gained considerable influence in the healthcare industry, including psychotherapy. The American Psychological Association's (APA) official stance on EBP encouraged clinicians to use standardized outcome measures in routine practice in order to establish the efficacy of their interventions. Routine Outcome Measurement (ROM) systems were designed specifically to accomplish this purpose, and have been shown to improve client outcomes and provide valuable aggregate data that contributes to empirical literature. Despite this research and the endorsement of the APA's official EBP stance, these measures have not been widely adopted by clinicians. Several studies have found that clinicians find the measures impractical and lacking in clinical relevance.

In order to accommodate these clinician concerns, while still maintaining the major features of ROM, the Clinically Adaptive Multidimensional Outcome Survey (CAMOS) was developed. The CAMOS employs a unique system that allows clinicians to be able to tailor the measure to the needs of their client, while still maintaining a core of standardized items. The present study attempted to identify a short form of McBride's measurement model, in order to determine which items would form this standardized core. The study found evidence for the validity and reliability of the CAMOS short form. With this evidence, the short form can serve as the basis for the CAMOS's unique tailoring system. It is hoped that the novel features of the CAMOS can help accomplish the APA's goals in relation to EBP.

Keywords: psychotherapy, routine outcome measurement, psychometrics

ACKNOWLEDGEMENTS

I am very grateful for the many people who contributed to this paper. I am especially grateful to my chair Dr. Richards for all of his support in developing my ideas and for his caring and excellent mentoring. I am also very grateful to my committee members, and in particular Dr. Sudweeks for taking several hours with me for consultation with data analysis. I am also very grateful for my colleagues Jason McBride and Justin Zamora for their help throughout this process. I could not have done this without my wonderful wife Amanda Fujiki and all the support she offers me and for her example of hard work. I am also grateful to my parents John and Suely for encouraging me every step of the way with my education.

TABLE OF CONTENTS

TITLE PAGE	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	viii
DESCRIPTION OF DISSERTATION STRUCTURE AND CONTENT	ix
Introduction.....	1
Routine Outcome Monitoring	2
Obstacles to Implementation.....	3
Relevance and practicality.....	3
Implementation approaches.....	5
The Clinically Adaptive Multidimensional Outcome Survey.....	7
The clinically adaptive method.....	7
CAMOS item development.....	9
The Need For a Short Form.....	10
Method	11
Participants	11
Sample 1.....	11

Sample 2.	11
Procedures	12
Sample 1.	12
Sample 2.	13
Data Analysis	13
Removal of rarely endorsed items.	14
Exploratory Factor Analysis.....	14
Confirmatory Factor Analysis (CFA).....	15
Reliability.	18
Full form comparison.	19
Convergent validity.	19
Results.....	21
Removal of Rarely Endorsed Items	21
EFA 1	22
EFA 2	23
CFA on Sample 1	24
Convergent and discriminant validity of factors	24
Internal consistency reliability.....	25
CFA on Sample 2	26
Model fit.	26

Convergent and discriminant validity.....	27
Internal consistency reliability.....	27
Correlation with full form.....	28
Convergent validity.....	29
Discussion.....	34
Limitations.....	39
Future Directions.....	41
Conclusion.....	42
References.....	44
APPENDIX A: Literature Review.....	53
Evidence-Based Practice.....	53
Empirically-Supported Treatments.....	54
Evidence-Based Practice in Psychology.....	55
Routine Outcome Monitoring.....	57
Improving individual treatment.....	58
Using ROMs to evaluate therapist effectiveness.....	67
Lack of Usage of ROM.....	72
Clinician attitudes towards ROM.....	73
Support for clinician attitudes.....	77
Researcher conceptualizations of lack of usage.....	80

A New Implementation Approach 85

References 88

APPENDIX B: Consent Forms..... 99

LIST OF TABLES

Table 1: Sample 1 Confirmatory Factor Analysis	25
Table 2: CFA Reliability and Validity Statistics	26
Table 3: Sample 2 Factor Loadings	29
Table 4: CAMOS Factor Correlations	31
Table 5: Convergent Validity Correlations.....	33

DESCRIPTION OF DISSERTATION STRUCTURE AND CONTENT

This dissertation, *The Development of a Short Form of the Clinically Adaptive Multidimensional Outcome Survey*, is written in a hybrid format. This hybrid format integrates traditional dissertation requirements and journal publication formats. The initial pages of this dissertation are for the purpose of fulfilling requirements for submission to the university. The remainder of the dissertation is written in a format that will allow it to be converted for a journal submission. The review of the literature is included in the Appendix. There are two reference lists contained in this dissertation. The first reference list contains references included in the journal-ready article. The second reference list includes all the references for the review of the literature.

Introduction

In the early 1990's a movement called Evidence-Based Practice (EBP) began to gain considerable influence in the field of applied psychology (Walker & London, 2007). The central tenet of EBP was that clinical work should be directly tied to scientific findings in order to ensure that practitioners were using interventions that had documented efficacy, were of high quality, and were cost-effective (Bower & Gilbody, 2010; Levant & Hasan, 2008; Spring, 2007). In 2005, the American Psychological Association (APA) assembled a presidential task force in order to formulate an official organizational stance on the issue of EBP. The result of this task force's work was a policy known as Evidence-Based Practice in Psychology (EBPP; APA Presidential Task Force on Evidence-Based Practice, 2006). This report differed from previous EBP policies, such as APA Division 12's Empirically-Supported Treatments (EST), in that the identification of specific treatments for specific disorders was de-emphasized in favor of general principles of evidence-based decision-making (Chambless et al., 1998; Levant & Hasan, 2008; Task Force on Promotion and Dissemination of Psychological Procedures, 1995). This policy provided a bottom-up approach to EBP, in which clinicians would be able to creatively tailor their treatment approaches while still practicing EBP if they were incorporating the decision-making principles espoused in the report. (Wampold, Goodheart, & Levant, 2007).

In moving away from creating lists of specific treatments for specific disorders, a new way to ensure accountability would need to be determined (Wampold et al., 2007). The EBPP task force's solution was to use systems that enabled "ongoing monitoring of patient progress and adjustment of treatment as needed" and stated that these systems "are essential to EBPP" (APA Presidential Task Force on Evidence-Based Practice, 2006, p.280). This would provide evidence for the work that the clinicians were doing with the actual clients they were seeing,

instead of relying on treatments validated in another treatment center. As of the time of the EBPP report, several measures already existed that were designed for these specific purposes (Stuart & Lilienfield, 2007). These measures are part of a growing paradigm known as Routine Outcome Monitoring (ROM; Carlier et al., 2012; Trauer, 2010a).

Routine Outcome Monitoring

ROM involves the use of standardized instruments that are designed to be practical, relevant, and rigorous in assessing patient progress throughout the course of therapy in naturalistic settings (Goodman, McKay, & DePhilippis, 2013). In order to reliably assess relevant variables in as few items as possible, ROM systems generally use a global measurement approach (Lambert, Okiishi, Finch, & Johnson, 1998). From a global measurement perspective, optimal items are those that are applicable in most contexts, with items that are applicable only to specific populations, theories, or clients being excluded (Lambert et al., 1996; Ogles, Lambert, & Masters, 1996). A key component of global assessments is the necessity of establishing the psychometric properties of the measure, in order to ensure that the meaning of the items is standardized (Ogles, Lambert, & Fields, 2002; Tarescavage & Ben-Porath, 2014). Several ROM systems have demonstrated strong reliability and validity with relatively few items, which ROM proponents view as evidence that the measure is generally capturing the primary relevant phenomena of interest to clinicians, even if it is not capturing everything (Boswell, Kraus, Miller, & Lambert, 2015; Tarescavage & Ben-Porath, 2014).

This measurement approach provides the basis for several features that allow ROM systems to be suitable for aggregate data analysis as well as to aid in the treatment of individual clients (Carlier et al., 2012). At the aggregate level, ROM can provide a relatively un-intrusive means of gathering quality data that could be used in research or program evaluation to provide

evidence that groups of clients are benefitting from treatment (Trauer, 2010b). At the level of the individual client, ROM systems provide real-time feedback that allows clinicians to track a client's progress across the course of therapy and to compare that client's improvement with either rationally or statistically-derived expected treatment response (ETR) curves (Beutler, 2001; Lueger, et al., 2001). This has been shown to improve treatment outcomes and contain costs, especially for clients who are at risk of treatment failure (Carlier et al., 2012; Goodman et al., 2013; Lambert et al., 2003; Shimokawa, Lambert, & Smart, 2010). With these features, ROM appears to be an ideal solution for facilitating Evidence-Based Practice in routine treatment without the need for top-down prescriptive guidelines (Wampold et al., 2007).

Obstacles to Implementation

Despite the endorsement of the EBPP task force, as well as nearly two decades of empirical research documenting its benefits, ROM has not been widely adopted by clinicians (Bickman et al., 2000; Gilbody, House, & Sheldon, 2002; Goodman et al., 2013; Hatfield & Ogles, 2004; Hatfield & Ogles, 2007). Some studies have found that even among clinicians that do routinely assess outcomes, only a small percentage actually use the real-time feedback provided by the measures (Garland, Kruse, & Aarons, 2003; Trauer, 2010b). For ROM to fulfill its role in achieving the objectives of EBPP, it is essential that it be more widely adopted, and that where adopted, the feedback be used to improve the quality of care. In order to increase the implementation and usage of ROM, it is necessary to understand some of the possible reasons that clinicians have not adopted it as well as discuss some of the ways that ROM developers have attempted to deal with these concerns.

Relevance and practicality. Two of the primary reasons for a lack of clinician usage of ROM identified in the literature are that they are perceived as not relevant or impractical

(Abrahamson, 1999; Trauer, 2010b). This is often surprising to ROM developers given that, as discussed above, they have developed brief assessments with good psychometric properties (Boswell et al., 2015). The use of global items to assess outcomes, however, comes at the expense of assessing aspects of the therapeutic encounter that are context-specific, that psychotherapists and patients may consider important (Evans 2012, Lakeman, 2004). In contrast to ROM developers' assumption that psychometric reliability and validity are sufficient for a measure to be relevant to clinical needs, several studies have found that practitioners often do not view standardized global items as possessing high clinical utility, and often place higher value on population-specific and individualized outcomes (Ashworth et al., 2007; Happell, 2008; Meehan, McCombes, Hatzipetrou, & Catchpoole, 2006). If practitioners place high value on these context-specific outcomes, which are purposely excluded from ROMs, it would come as little surprise that they would not see great value in adopting these systems. Clinicians would essentially be tracking their clients' progress and having their work evaluated based on criteria that they do not view as germane to treatment.

There is some evidence that these clinicians' concerns may not be unfounded, with Ashworth and colleagues (2007) finding that the CORE-OM (a prominent ROM system) did not address 60% of the topics that brought clients into therapy in a relatively large sample of clients. From this perspective, although a broad approach to measurement may allow for brevity, it could be argued that any time spent on a measure that is not assessing what is most important to treatment is "too much" time. If this is the case, then even a brief global measure could be both irrelevant and impractical because it is not providing enough useful information to justify the time and money spent. Additionally, some have expressed concern that a focus on the outcomes

included in ROM may lead practitioners to decrease their attention to other areas of concern that are not included in the instrument (Browne, 2006; Evans, 2012; Lakeman, 2004).

On the other hand, a major limitation of more individualized approaches to outcome measurement (i.e., Target Problems or Goal Attunement Scaling) is that they are more difficult to compare across clients, thus making it difficult, if not impossible, to aggregate data across clients or to create ETR curves (Lambert et al., 1998; Ogles, et al., 1996). If this were the case, the major benefits of ROM would be lost. This puts researchers and clinicians at an impasse: Both want the assessments to be brief, but ROM developers place high priority on standardization, aggregation, and comparability, while many clinicians value assessments that are specific to the populations they work with, their theoretical approach, and the unique circumstances of their clients (Sales & Alves, 2012).

Implementation approaches. In attempting to implement ROM systems, researchers have often conceptualized clinician concerns about relevance as obstacles or barriers that need to be overcome (Meehan, et al., 2006; Walfish, McAlister, O'Donnell, & Lambert 2012;). In order to overcome these barriers, ROM developers have attempted to change what they perceive to be erroneous clinicians' attitudes (Abrahamson, 1999; Hannan et al. 2005; Meehan, et al., 2006). The lack of usage of ROM suggests that clinicians have generally not changed their attitudes, which has led many ROM proponents to lament clinician resistance toward measures that have documented efficacy in improving outcomes (Boswell et al., 2015; Tascavage & Ben-Porath, 2013; Walfish et al., 2012;).

This lack of adoption is unsurprising in light of a growing literature in the field of organizational behavior that has begun to question the validity and utility of conceptualizing organizational change as a process of overcoming barriers or obstacles from "resistant" change

recipients (Ford, Ford, & D'Amelio, 2008; Thomas & Hardy, 2011). In regards to the efficacy of approaches that attempt to minimize “resistance”, Thomas & Hardy (2011) stated: “While change can be imposed, it is more likely to be taken on by members of the organization if they have played a part in the negotiations of new meanings, practices and relationships” (p.323). From this perspective, implementation would be more likely to be successful if ROM developers allowed treatment sites to have input about the measures. In existing ROM systems, however, only minor logistical concerns such as the frequency or format (digital or paper) of administration are negotiable, but tailoring the content of the measure itself to the unique needs of a treatment site is typically not offered as an option (Boswell et al., 2015; Mellor-Clark, Cross, Macdonald, & Skjulsvik, 2014).

The approach to development that undergirds existing ROM systems does not allow for the flexibility necessary for an implementation approach in line with Thomas & Hardy's (2011) views. Clinicians cannot negotiate the outcomes that will be assessed in the measure, because in the current paradigm of ROM development, such adaptations are difficult, if not impossible, to accommodate (Boswell et al., 2015). The primary reason for this lack of flexibility is the assumption that for an item or dimension to be useful, it must be psychometrically validated and that if validated, the measure is sufficiently relevant to generate locally applicable data (Lambert et al., 1998; Slade, 2002). This validation process can be very time, labor, and resource intensive, often involving years of work across thousands of participants (Kraus, Seligman, & Jordan, 2005). If, however, this assumption is invalid and clinicians are correct that the measures lack relevance for their specific contexts, then ROMs offer little of value. What use is an ETR for outcomes that are not relevant to treatment, or how useful are outcome studies based on data that is not reflective of the true goals of therapy? Thus, in the current paradigm, ROM

developers have little option but to view clinician non-adoption as resistance, because their systems were not designed to facilitate negotiations about local meanings of outcome, and the utility of their instruments are threatened by clinician feedback.

The Clinically Adaptive Multidimensional Outcome Survey

In order to both maintain the current benefits of ROM as well as allow clinicians to have a voice in defining the meaning of outcome, a new paradigm for ROM development is necessary. With these considerations in mind, the Clinically Adaptive Multidimensional Outcome Survey (CAMOS) was designed. Contrary to existing ROM systems, the CAMOS does not make the assumption that only standardized or global items are of clinical value, nor that psychometric validation is sufficient for a measure to be of clinical relevance. Thus, in addition to having a core of standardized items similar to existing ROM systems, the CAMOS allows all stakeholders involved in the therapeutic process to have input in defining which concerns and constructs will be assessed in the measure. This flexibility allows the CAMOS to be relevant to specific contexts, not because the developers are assessing every possible outcome, but because the major stakeholders in treatment can adapt it to specific treatment contexts. This system changes the agenda of implementation from one of attempting to change clinician attitudes in order to fit the measure, to making the measure change to fit local needs.

The clinically adaptive method. One major concern with this tailoring is that it may create an excessive time burden for clients and therapists, which could be a significant barrier to adoption of the measure (Duncan, 2012). This has been shown to be the case with systems that have attempted to combine standardized and individualized approaches to treatment (Sales & Alves, 2012). Thus, in order to accomplish the goal of adapting to the contextual needs of

clinicians, while still maintaining the main features of ROM, the tailoring process must be more feasible than existing systems.

The CAMOS's unique approach is called "Clinically Adaptive" assessment and involves making a small subset of items required for all administrations, while allowing other items to be optional. These optional items can then be added or removed from the CAMOS at the discretion of the treatment site or individual therapist. Required items are standardized in the same way as traditional ROMs, and allow for aggregation and comparability, thus maintaining the major features of these systems. The optional items can either be chosen from an existing pool of items (the item pool will be discussed in depth below) or can be written by the treatment site and/or therapist. This tailoring can occur at three levels: (a) items or dimensions can be added, excluded, or created that will display only on the CAMOS survey of all clients at a specific treatment site, (b) items can be added or authored that appear only on the CAMOS survey of clients assigned to a specific therapist (c) items can be added or authored that are displayed only to individual clients. This multi-level tailoring approach differs from idiographic approaches which primarily allow for tailoring at the level of the individual client (Sales & Alves, 2012).

The clinically-adaptive method allows for the creation of a template that incorporates the views of various stakeholders, while still allowing for individualized tailoring. Additionally, the option to add items or dimensions based on an existing pool, instead of relying exclusively on clients or therapists to author new items for each client, allows the tailoring process to be much more efficient. At the site and therapist level, pilot testing has found that this tailoring process only requires a one-time 5-10 minute investment for therapists, substantially reducing time burden relative to idiographic approaches, and allowing for needs at various levels to be accommodated, while still allowing for comparability. At the level of the individual client, items

can be easily added or removed from the survey at any session by simply checking a box in the client's report. As discussed above, if the clinician or treatment site would like to invest more time to author items, the option is available. Additionally, if a clinician is content to use the CAMOS as a traditional ROM, they can use it with no tailoring involved.

This tailoring is made possible through modern computational developments and would be impossible to perform with standard pen-and-paper surveys. The use of this technology also allows for the survey to retain data obtained about the client, as well as input from the therapist, in order to generate a customized survey that is relevant to the needs of the client and the orientation of the therapist. The use of computers to administer the measure also makes possible automated scoring and the generation of real-time reports, thus saving valuable clinician time (Boswell et al., 2015). Finally, this facilitates more accurate and feasible data analysis for researchers examining the data, while obviating the need for data entry, which can be time consuming and introduces the potential for human error in creating the data set.

CAMOS item development. The full form of the CAMOS item pool was developed based on Richards and Bergin's (2005) multilevel assessment approach as well as recommendations from the ROM literature (Slade, 2002). Richards and Bergin proposed measuring distress in seven dimensions: cognitive, emotional, relational, physical, behavioral, spiritual, and occupational. In addition to these dimensions, a therapy progress dimension was added to evaluate client perceptions of the therapeutic alliance. Items were developed by the CAMOS research team in order to assess these 8 domains of functioning. On each CAMOS item, the client is asked to report the level of distress they experienced over the past week on a 6-point likert scale (Never, Rarely, Sometimes, Frequently, Almost Always, Always). These items were then evaluated by clinicians at multiple treatment sites and were adjusted after several

iterations of feedback to create an initial pool of 72 items. The full item pool was administered at two treatment sites, and McBride (2015) performed exploratory and confirmatory factor analysis on this data. His analysis led to a 45-item 6 factor model that left the majority of the theoretical dimensions intact, but collapsed the behavioral, emotional, and cognitive dimensions into one factor called psychological distress.

The Need For a Short Form

McBride's model was designed primarily to define the items that would be required at the initial intake assessment, making all others optional. Its focus was to identify a wide variety of potential client concerns. This model, however, was not intended to be administered on a session-by-session basis while still allowing for tailoring. Most ROM systems are designed to be brief in order to allow session by session administration, with most being composed of less than 50 items. Given that the full form is composed of 42 items, this leaves little to no room for clinician tailoring, without making the assessment excessively burdensome. The present study aims to create a short form of the CAMOS that will serve as the required items for session-by-session measurement, with all items not included in the short form becoming optional after the first session. Thus, the items identified in the short form will not vary from session to session. This form will need to be sufficiently brief in order to allow for additional site, therapist, and client level adaption. If this short form proves to have adequate psychometric properties, it would provide a standardized core for the CAMOS that is useful for aggregate data analysis which would allow for the development of ETR curves, while still being responsive to the voice of the clinician and client.

Method

Participants

Sample 1. The sample consisted of 304 participants in treatment at a private university counseling center in the western United States with a wide variety of diagnoses. Therapists assigned diagnoses to 215 (71%) clients, of which, 118 (54%) were given at least two diagnoses. The most frequent diagnoses were major depression ($n = 106$), anxiety disorders ($n = 72$), and impulse control disorder ($n = 23$). The participants' ages ranged from 17 to 57 years of age, with a mean of 21.8. There were 186 (61.2%) female participants and 116 (38.2%) male participants. The self-identified racial demographics were as follows: 76% White/Caucasian, 8% Latino/a, 2% Asian, 1.0% Native American, 1% African American, .3% Pacific Islander. Twelve percent of clients did not respond to the racial demographics item. The religious affiliations of the participants were: 98.7% Latter-Day Saint (LDS), .3% other, and 1 participant did not identify him/herself with a religion. Participants were recruited for the study by a secretary inviting them to participate as part of their intake paperwork. This was voluntary for clients entering treatment and participants could leave the study at any point.

Sample 2. Participants were in treatment at an inpatient eating disorder clinic in the western United States. The sample was composed of 211 clients, all of whom were female. Participants' ranged from 14 - 45 years of age with a mean of approximately 22 years. The racial demographics of the sample were composed of approximately 80% White/Caucasian, 10% Latina, 5% African American and 1% Asian and 9% other/multiracial. The religious affiliations of the sample were approximately 40% Latter-Day Saint, 20% not religious, 20% Protestant Christian, 10% Jewish and 10% other.

Procedures

Sample 1. Once the treatment sites had agreed to participate in the study, the developers of the CAMOS went to the sites in order to integrate the administration of the CAMOS into the flow of existing office procedures and to train the staff in the usage of the system. This involved providing the treatment site with Amazon Kindle Fire tablets, setting up links to the survey on each device, training secretaries and medical technicians in how to access the survey, and working out other minor logistical issues. The developers also met with the clinicians to demonstrate how the system would work, as well as to get feedback about the items. The feedback was then incorporated into the CAMOS.

Upon making an appointment for their first session, clients were invited to participate in the study and signed an informed consent document. They were then entered into the CAMOS system by having secretaries fill out a brief survey with the client's identification number and the identification number of the therapist to which they were assigned. Client and therapist names were not used in order to ensure confidentiality. When a client came in for her first visit, the secretary handed him/her the tablet on which he/she would take the survey. In order to ensure that participants did not make errors in entering their ID numbers, a verification system was implemented that would only allow participants to take the survey if their ID number was entered correctly.

After successful login, the client completed the survey. At intake, demographic items were presented to clients as well as the full CAMOS item pool, allowing for a cross-sectional data analysis of the full item pool. Following intake, participants only responded to the items in the areas where they expressed concern on initial screening items. Only intake data were used in the present analysis. The CAMOS took approximately 5-7 minutes to administer at intake and

approximately 3-5 minutes to administer in subsequent sessions. After the client had completed the CAMOS, the clinicians could access a report that provided information about the client's responses to individual items, as well as track their scores over time.

Sample 2. Upon being admitted to the treatment center, clients were required to complete a battery of assessments. With administrative approval, the full initial item pool of the CAMOS was added to the battery of assessments. The measures were administered via computer, and clients were directed to take the measures by staff members. The assessments were not taken at the same time, with the CAMOS being administered separately from other tests. Given that this assessment procedure was already in place prior to the initiation of the current study, little training was needed for the CAMOS to be implemented. Additionally, the authors developed a digital form that allowed staff to see which clients (by ID only, no names were entered into the system), had completed their assessments. The clients took all items of the CAMOS at admission as well just prior to their discharge. Only the data from intake assessments was used in the current study.

Data Analysis

The present analysis imposed more stringent criteria for item retention than were used in McBride's analysis, with the intention of shortening the measure to make it more suitable for repeated measurement in psychotherapeutic settings while still maintaining strong psychometric properties. The data analysis involved six steps: (1) Removal of highly positively skewed items; (2) Exploratory Factor Analysis (EFA) on McBride's full item pool to identify low loading items; (3) Confirmatory Factor Analysis (CFA) of the model developed through EFA on both samples (4) Calculation of internal consistency reliability of each dimension; (5) Correlating the

short form with the corresponding full form of each dimension; (6) Correlating each dimension with existing measures.

Removal of rarely endorsed items. One of the primary purposes of this study was to identify items that would be able to track progress across the course of psychotherapy. In order to do so, the items that were used in the short form would need to leave room for improvement and be relevant to most clients' presenting concerns. If items were consistently being answered at the lower end of the scale (Never or Rarely) at intake, this was interpreted as a floor effect, and these items may in fact detract from the measure's sensitivity to change (Kraus, Seligman, & Jordan, 2005). Thus, items that were endorsed by over 50% of sample 1 as "Never" or "Rarely" ("Almost Always" or "Always" for positively worded items) at intake were removed from future analyses. The rationale for this cutoff was that if 50% of clients indicated that the item was of minimal concern at intake, then by definition, the majority of clients did not find this relevant to their concerns. This analysis was performed by examining the frequencies of each response using IBM SPSS 21 statistical software.

Exploratory Factor Analysis. Once these strongly positively skewed items were removed from the item pool, two Exploratory Factor Analyses (EFA) were performed on the remaining items. The analyses was performed using the Mplus 7.1 structural equation modeling software on sample 1 using a Geomin (oblique) rotation, and Weighted Least Squares extraction (Muthen & Muthen, 2012). The number of factors to retain was determined by Parallel Analysis (PA), which is considered one of the more precise methods for factor retention decisions (Hayton, Allen, & Scarpello, 2004; Horn, 1965). In PA, the researcher compares the eigenvalues generated in the EFA to those of a set of random eigenvalues generated through a Monte Carlo simulation. Factors are retained that have an eigenvalue that is significantly higher than the

random eigenvalue for that factor. In addition to this empirical analysis, a rational examination of the factors suggested by PA was performed. This is essential in order to determine whether the factors identified were substantively meaningful and clinically relevant (Fabrigar, 2012).

The first EFA was intended to reduce the item pool by removing items that did not have substantial loadings onto any factor or that had had high cross-loadings. For this EFA, a substantial loading was defined as .4 or above, and cross loadings were considered unacceptable if the difference between the two highest loadings for an item were less than .2. These criteria helped narrow down the item pool and create more tightly defined constructs by removing items whose variance was not well explained by any of the factors.

Once this analysis was performed and items were removed according to the criteria outlined above, an additional EFA was conducted on the remaining items. The rotation, extraction, and factor retention procedures were identical to the previous EFA. In contrast to the first EFA, for an item to be included in the final model a loading of .5 or higher on one dimension was required. The .2 cross loading difference criterion from the previous EFA was used in this analysis as well. This was done in order to further narrow down the construct to a small core of items. The minimum number of items for each dimension was put at 4, in order maximize reliability and construct coverage, while still being relatively brief (Kenny, 1979; Kline, 2010).

Confirmatory Factor Analysis (CFA). Once a suitable model was generated from the EFA, a CFA was performed on both samples. CFA allows the researcher to specify an *a priori* model that provides a rigorous test of whether the hypothesized model adequately accounts for the major relationships between the items (Kline, 2010). In contrast to an EFA, CFA involves fixing all cross-loadings and indicator error covariances to 0 (unless manually overridden by the

researcher). These stringent constraints provide strong evidence of the validity of the constructs hypothesized in the model. In order to determine if the model implied covariance matrix adequately fits the observed data, various fit statistics have been developed (Kline, 2010; Sun, 2005). Mplus 7.1 (the software used for all CFAs in this study) provides the following fit indices based on the Weighted Least Squares estimator that was used due to the categorical nature of the items: the Root Mean Square Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI).

The interpretation of these fit indices, however, has been a topic of debate among CFA researchers (Marsh, Hau, & Wen, 2004). The most widely used guidelines for interpreting fit indices were developed by Hu and Bentler (1999). In their study based on Monte Carlo simulations, the following guidelines were suggested: RMSEA of .06 or below and CFI/TLI of .95 or above. Hu and Bentler (1999) warned, however, that these should not be used as strict cutoffs, but general guidelines for model fit. In line with these, and other recommendations (i.e., Marsh et al., 2004), the current analysis accepted the model if the fit was near the values suggested by Hu and Bentler (1999), but gave some flexibility if the exact values were not met.

Many studies perform EFA's and CFA's only on a separate sample from that used in EFA, however, van Prooijen and van der Kloot (2001) found that models generated through EFA often were not confirmed in CFA on the same sample. If this were the case, then attempts to cross-validate would likely be unsuccessful, not necessarily because the measure is not generalizable, but possibly because of methodological errors in the EFA. By examining whether the model hypothesized by the EFA results is confirmed in a CFA on the same sample, future cross validation studies are strengthened, and substantive differences can be sorted from

methodological problems. At this step, no model respecification was performed, in order to minimize capitalization on chance.

With this rationale, a CFA was first performed on the same sample (Sample 1) as the EFAs described above. Running a CFA on only the same sample as the EFA, however, is not advised in the literature (van Prooijen & van der Kloot, 2001). If a CFA is not run on a separate sample there is an increased risk that positive findings may not be substantive, but may be due to unique response patterns or characteristics of the sample used in both analyses (Cudeck & Browne, 1983). With these considerations in mind, an additional CFA was performed on Sample 2.

On both CFAs, if the model demonstrated adequate fit, then it would be possible to interpret the parameters generated by the model. The convergent validity of the items on each factor in the model was then tested by examining whether the Average Variance Extracted (AVE) was above .5 (standardized factor loadings were examined to see whether the average squared loading for each factor was above .5 (Hair, Babin, & Anderson, 2010). The AVE was calculated by squaring the values for each of the factor loadings and computing an average for the factor. This provides an estimate of the average amount of the variance of the indicators is explained by the latent factor. The .5 cutoff was chosen because if the AVE value was not above .5, this would indicate that less than 50% of the variance of the items was explained by the factor, and conversely, that over 50% of the variance was attributed to error. Discriminant validity was examined by comparing the AVE of the factors to the Maximum Shared Variance (MSV) between factors (Fornell & Larcker, 1981). This was calculated by squaring the highest correlation between the dimensions. If the AVE value for a factor was greater than its MSV, this would indicate that the within factor variance was higher than the between-factor variance.

Reliability. Internal consistency reliability was computed using two methods: Cronbach's alpha (Cronbach, 1951) and Raykov's rho (Raykov, 1997a). Cronbach's alpha was computed due to its widespread usage, thus allowing for a point of comparison between the CAMOS and existing measures. Alpha, however, has been criticized for its usage with survey data due to its requirement that the survey items are tau equivalent measures of the construct being assessed, an assumption which is not often met in social science research (Raykov, 1997b). If this requirement is not met, then alpha may over or under-estimate the reliability of the scale, with no way of knowing which was occurring. The author did not anticipate that the CAMOS would meet the assumptions of tau-equivalence (few psychological measures do), thus, alpha would be interpreted tentatively and supplemented with Raykov's rho. A value of at least .8 was determined to be an optimal minimum threshold for reliability (Streiner, 2003).

Raykov's rho is a reliability coefficient that can be used with congeneric measures, meaning that the items do not cross load, but are not required to have nearly equivalent factor loadings (Raykov, 1997a). This appeared to be more congruent with of the type of model that would be likely be generated in this study. Additionally, rho assumes unidimensionality of the construct being assessed, thus it is required to perform a CFA on each individual dimensions for both samples (Raykov, 1997a). Raykov's rho uses the parameters of a CFA model in order to generate the reliability index, thus a CFA was performed on each dimension for both samples. Maximum-Likelihood (ML) estimation was used due to there being no way to compute rho based on the parameters generated with Weighted Least Squares estimation. As with the previous CFAs, it was important to establish that each dimension adequately fits the observed covariance matrix, otherwise the parameters used to calculate the coefficient may not have been accurate and could produce misleading findings (Raykov, Rodenberg, & Narayanan, 2015). Fit

was evaluated for each dimension based on the same criteria used for the other CFA models in the current study, except that with ML the Standardized Root Mean Square Residual (SRMR) is also used. Based on Hu and Bentler's (1999) cutoff criteria, an SRMR of $< .08$ would suggest a good fitting model, although, as discussed above, this cutoff should not be held as absolute. Raykov's rho coefficients can be interpreted in the same way as alpha, thus the same cutoff values were used.

Full form comparison. An essential component of determining whether a short form adequately captures the construct of the full form is to examine whether the retained items capture the breadth of the original construct (Smith, McCarthy, & Anderson, 2000). If such an analysis is not undertaken, then relying exclusively only on empirical criteria may narrow the construct to the point that it will not be measuring the same construct as the full form, and may be reliable but not valid (Smith et al., 2000). Thus, the author examined each dimension to determine whether the major aspects of the full form dimension were retained in the shortened version. Following this, each short form dimension was correlated with its respective full form counterpart. If the major facets of the construct of the full form were maintained, and the correlations were sufficiently high this would provide evidence that the short form was assessing the same construct as the full form.

Convergent validity. One essential test to determine if a measure is valid is to examine whether it is correlated with scores on another measure that assesses a similar construct. In developing a short form, this would involve examining whether it correlated with existing measures at a similar level to the full form. Smith and colleagues (2000) have discussed the importance of not assuming that the convergent validity of the full form of a measure is applicable to short form, thus it is important to perform the analyses on both forms. If the

correlations were found to be similar, this would provide evidence that the latent construct being assessed in the short form was not substantially different from the full form. Pearson correlations were performed between the CAMOS dimensions and the following measures.

The Outcome Questionnaire 45 (OQ-45). The OQ-45 is a 45-item instrument intended to measure client progress in therapy along three dimensions: subjective distress, social roles, and symptom distress (Lambert et al., 1996). Lambert et al. (1996) found the OQ-45 to have excellent internal consistency reliability and strong concurrent validity with several measures of distress including the Beck Depression inventory and Symptom Checklist-90. It is one of the most commonly used ROM systems in the United States (Hatfield & Ogles, 2004)

The Minnesota Multiphasic Personality Inventory (MMPI-II). The MMPI-II is a 567-item assessment of personality characteristics and pathology (Butcher & Pope, 1992). The MMPI-II is composed of 10 primary clinical scales, as well as over 100 subscales to provide an in-depth examination of a person's patterns of pathology. Recently, the MMPI-II primary scales were updated, with the resulting measure being called MMPI-2 RF (Ben-Porath & Tellegen, 2008). This update restructured the clinical scales of the MMPI-2 using a subset of the items in the original version in order to improve its psychometric properties, and is intended to supplement the original version. This study will use both the restructured and original clinical scales, as well as selected subscales that are theoretically relevant to the dimensions of the CAMOS.

Multidimensional Self-Esteem Inventory (MSEI). The MSEI (O'Brien & Epstein, 1988) is a 116-item scale designed to measure various aspects of a patients' self-perceptions and self-esteem. The MSEI provides 11 scores, including a global composite score, 8 self-esteem

sub-scales, an identity integration scale and a validity scale. The MSEI has been found to have acceptable internal consistency reliability as well as concurrent and discriminant validity.

Theistic Spiritual Outcome Scale (TSOS). The TSOS is a 17-item scale that assesses theistic spirituality in 3 domains, and also provides a total score based on these domains (Love of God, Love of Self, Love of Others). Richards et al. (2005) found the TSOS to have excellent internal consistency reliability as well as convergent and discriminant validity. It was hypothesized that the TSOS would correlate highly with the spirituality dimension of the CAMOS. In concurrent validity analyses, only the TSOS scores for clients in sample one were used. This was because respondents from sample two who were not religious or spiritual often responded with the lowest possible answer for each item, thus providing a response set that rendered interpretation of the correlations difficult.

Results

Removal of Rarely Endorsed Items

In order to identify the items that had 50% of respondents answer at “Never” or “Rarely” (or “Always”/”Almost Always for positively worded items), descriptive frequency statistics were calculated in SPSS on sample 1. In examining the frequencies, 18 items were found that 50% or more of the respondents in sample 1 endorsed at “rarely” or “never.” These items were removed from future analyses. With these items removed, the work concerns dimension from McBride’s model was left with 3 items, which was less than the required minimum of 4 items for a dimension. Additionally, in McBride’s model, this dimension showed relatively weak reliability and concurrent validity. With these considerations, the remaining items in the work concerns dimension were dropped.

EFA 1

Prior to beginning the EFA, items that were a part of McBride's method bias factor (due to positive wording) were excluded. This left some positively worded items, but only those that loaded more highly on a substantive dimension than on the method factor. An additional item presented a likely method bias, in that it was nearly identical to another item except for one word. This was removed as it was part of a dimension which had many items, and was not conceptually essential. These exclusions left 35 items to be analyzed in the EFA.

EFA was performed on sample 1 using Mplus 7.1 and the extraction and rotation procedures described in the methods section. Parallel analysis of the first EFA on sample 1 suggested retaining 5 factors. In examining these factors, their groupings were both interpretable and consistent with the factors identified in McBride's analysis. The 5 factor model accounted for 60% of the total variance of the data. Within the common variance, 60% was accounted for by factor 1. This indicates that one factor was accounting for the majority of the variance of the model. These estimates should be interpreted tentatively due to the fact that this EFA used an oblique rotation which allows for correlations between the factors. Thus, the exact amount of variance accounted for by a factor was difficult to ascertain due to the fact that some of the variance was shared between multiple factors. The exclusion criteria described previously were then employed. Three items were removed based on the cross loading criteria (difference of highest loadings $<.2$) and one item was removed due to the loading criterion (loading $<.4$). The factors all had significant correlations with each other, but none were above $.5$, indicating that they were likely distinct, but related, factors.

EFA 2

With the items removed that did not load well or that cross loaded to an unacceptable degree, it would be possible to obtain a more precise model. This EFA used the same rotation, extraction, and retention procedures as the previous analysis. Parallel Analysis suggested retaining 5 factors. These factors accounted for 62% of the total variance. Within the common variance, the percentage of variance explained by the 5 factors were: 57%, 14%, 11%, 9%, and 8%. Three items were removed due to not meeting the requirement of loading at .5 or above, and no items were removed due to the cross-loading exclusion criteria. The removal of these items left 3 dimensions with 4 items, while one factor had 10 items and another had 6. The dimension with 10 items was deemed to be still too lengthy for the short form, but met all empirical criteria for inclusion. Thus, a rational content analysis was performed in order to remove items that appeared to carry the least amount of theoretical importance. Upon examining the items in this dimension, all but three were symptoms of an anxiety or depressive disorder according to the *Diagnostic and Statistical Manual for Mental Disorders* (5th ed.; DSM-5; American Psychiatric Association, 2013). These three items were also redundant or excessively broad to the point of not adding much clinically relevant information, and thus were removed from the model. The model that emerged from this analysis had 25 items divided up into 5 dimensions: Therapy Expectations (4 items), Relationship Concerns (6 items), Psychological distress (7 items), Spiritual concerns (4 items), and Physical Health concerns (4 items). The content of the dimensions closely resembled that of McBride's model and there did not appear to be a substantial loss of content breadth in any factor. Additionally, all dimensions correlated at $>.90$ with their full form counterpart, providing further evidence that the theoretical core of each dimension was maintained. All items loaded on their respective dimension at .5 or higher and

did not have any cross-loadings above .2, making this model an excellent candidate to be tested in CFA.

CFA on Sample 1

As discussed in the methods section, an important step in determining the factor structure of the CAMOS short form was to examine whether the model developed in the EFAs could be confirmed through a CFA on the same sample before moving to cross-validation. Therefore, a CFA was performed on sample 1 with Mplus 7.1 using Weighted Least Squares estimation due to the categorical nature of the items. The model demonstrated good fit, with all indexes being at or near the optimal values proposed by Hu and Bentler (1999). The fit statistics were as follows: Chi-square= 616.7, df=265, p=0.00, RMSEA=.066 (.059-.073), CFI= .954, TLI=.948. These values suggest a good fitting model, and that the parameters estimated in the CFA are interpretable.

Convergent and discriminant validity of factors. Average Variance Extracted (AVE) and Maximum Shared Variance (MSV) were calculated (See table 2) for each dimension. All dimensions had an AVE of at least .5, providing evidence for the convergent validity of the indicators on their respective factor. No dimension had a higher MSV than AVE, indicating that more variance was explained within the factors than between them. The Relationship Concerns dimension, did however, come very close to having an MSV that was higher than the AVE (AVE=.54, MSV=.52) due to its high correlation with the Psychological Distress dimension. This suggests that these dimensions may have substantial overlap. Overall, these findings provide support for the convergent and discriminant validity of the 5 CAMOS factors for sample 1.

Table 1

Sample 1 Confirmatory Factor Analysis

Item	PD	RC	SC	TE	PC
I felt worried, agitated, fearful, or tense	.81				
I felt sad or depressed	.88				
I thought about past personal failures/ mistakes	.79				
I had thoughts or images that I couldn't get out of my head	.63				
I felt powerless or stuck in my problems	.75				
I had difficulty concentrating or remaining focused on a task	.61				
I felt worthless or "not good enough"	.88				
I felt misunderstood by my loved ones and friends		.83			
I felt hurt or disappointed by how others behaved		.85			
I felt concerned about my relationships		.68			
I felt accepted by my friends and loved ones		-.67			
I felt sad about how I acted towards others		.70			
I felt irritated and angry towards others		.65			
I felt a loss of inspiration or spiritual direction			.92		
I felt distant in my relationship with God or my Higher Power			.89		
I felt concerned about my religious or spiritual life			.64		
I felt guilt over mistakes that were inconsistent with my religious beliefs			.74		
I had concerns about beginning therapy				.84	
I felt anxious about beginning therapy				.80	
I felt uncertain about whether I can be fully open with my therapist				.69	
I had doubts about whether my therapist will understand my concerns				.72	
I felt physically well and healthy					.83
I experienced physical pain or discomfort					-.75
I felt light headed, weak, or fatigued					-.81
I had a stomach ache or other gastro-intestinal problems					-.62

Note. PD=Psychological Distress; RD=Relationship Concerns; SD=Spiritual Concerns; TE=Therapy Expectations; PC=Physical Concerns. All loadings were significant.

Internal consistency reliability. Cronbach's Alpha for each dimension was at or above .8 for each dimension (see Table 2). This provides evidence that all of the short form scales

demonstrated good internal consistency reliability. Raykov's rho was also calculated due to the considerations previously discussed. In order to calculate rho, CFAs were performed on each dimension individually. These CFA models were evaluated for model fit in order to determine whether the parameters generated by the model were interpretable, as these parameters would provide the values for the calculation of rho. These estimates were obtained using a syntax template from Raykov (2014) in Mplus 7.1, and used Maximum-Likelihood estimation due to the lack of a procedure for calculating rho using Weighted Least Squares estimation. All dimensions were found to have at least acceptable fit except for Therapy Expectations. With this consideration, this dimension's rho should be interpreted cautiously. Rho coefficients were all at or above .85 (.85-.91), which indicates that most dimensions were underestimated in alpha. These values provide compelling evidence for the internal consistency reliability of the data obtained from this sample.

Table 2

CFA Reliability and Validity Statistics

Dimension	Sample 1					Sample 2				
	α	Rho	AVE	MS V	FF <i>r</i>	α	Rh o	AVE	MSV	FF <i>r</i>
Psychological Distress	.89	.91	.59	.52	.97	.90	.92	.63	.54	.975
Relationship Concerns	.80	.85	.54	.52	.96	.83	.86	.51**	.53**	.97
Spiritual Concerns	.84	.91	.65	.39	.98	.81	.84	.63	.42	.99
Therapy expectations	.80	.85*	.59	.20	.91	.84	.87	.62	.40	.94
Physical Health Concerns	.80	.84	.57	.38	.92	.82	.86	.60	.54	.90

Note. AVE=Average Variance Extracted, MSV=Maximum Shared Variance, FF *r*=Correlation with full form

*Model did not demonstrate adequate fit

**AVE less than MSV

CFA on Sample 2

Model fit. In order to minimize the probability of capitalization on chance, it was critical that the factor structure identified in the previous analyses be replicated on a separate sample. A

CFA was performed on sample 2 in order to provide this data. The procedures and decision-making criteria for this CFA were identical to the previous CFA. Model fit was found to be very good (Chi-square 437.767, RMSEA= .056, CFI= .972, TLI=.969) indicating that the parameters of the model were interpretable. Table 3 shows the factor loadings for each item on sample 2.

Convergent and discriminant validity. All dimensions had an AVE of at least .5, providing evidence for the convergent validity of the factors. In four of the dimensions, the AVE was substantially larger than the MSV. The only exception to this was the Relationship Concerns dimension, in which the MSV of .53 with Psychological Distress was slightly higher than the AVE of .51. These findings support the convergent validity of all five dimensions, but suggest that the relationship concerns and psychological distress dimensions may have substantial overlap.

Internal consistency reliability. Table 2 shows the values for both alpha and rho for this sample. Cronbach's alpha for each dimension was at or above .8 for each dimension. This provided evidence that all of the short form scales demonstrated good internal consistency reliability. Similar to sample 1, a CFA was performed on each dimension in order to generate the parameters that would be used to calculate rho. All dimensions were found to have at least acceptable fit except for Spiritual Distress. This may be due to the fact that many respondents who did not identify as religious or spiritual often responded "never" to every item in the dimension. With this consideration, this dimension's rho should be interpreted cautiously or not at all. All rho coefficients, except for Spiritual Distress (.84), were above .86 (.86-.92), which suggests that most dimensions were underestimated in alpha. These values provide compelling evidence for the internal consistency reliability of the data obtained from this sample.

Correlation with full form. All dimensions correlated with their full form equivalent at $>.9$ (.90-.99). Similar cautions should be applied in interpreting these values as discussed with the sample 1 data. Given that the majority of the items in each dimension were present in both the full and short form, it is important that a conceptual analysis be performed, comparing the construct coverage of the full and short forms (Smith et al., 2000).

The most obvious change from the intake form is the removal of the work distress dimension. As was found in McBride's analysis, this dimension demonstrated relatively low reliability. Outside of this, the most changed dimension was physical health distress. Several items from this dimension were removed due to a very low percentage of clients endorsing them as concerns, with 75-85% of clients endorsing these items at "never" or "rarely." Although this limits the scope of the construct, these items did not seem to be relevant to most clients. The Psychological Distress scale remained largely the same, except for the removal of items assessing general concerns with thoughts, emotions, and behaviors. These items were viewed as being redundant to the more specific items in the scale. Additionally, two items related to stress were removed, as they were viewed as redundant to the item that assessed worry and fear. The Relationship Distress dimension maintained all but two items from the intake form, with the removed items being not often endorsed by most clients. This limited the construct in terms of assessing angry or explosive interpersonal interactions, although this is likely covered by the item assessing irritability and anger toward others. The Therapy Expectations and Spiritual Distress dimensions remained identical to the intake form, except for the removal of 1 item from each, thus the construct coverage is likely to be very similar to the full form.

Table 3

Sample 2 Factor Loadings

Item	PD	RC	SC	TE	PC
I felt worried, agitated, fearful, or tense	.87				
I felt worthless or "not good enough"	.87				
I felt powerless or stuck in my problems	.84				
I felt sad or depressed	.81				
I thought about past personal failures/ mistakes	.77				
I had difficulty concentrating or remaining focused on a task	.74				
I had thoughts or images that I couldn't get out of my head	.59				
I felt misunderstood by my loved ones and friends		.81			
I felt accepted by my friends and loved ones		-.74			
I felt irritated and angry towards others		.72			
I felt concerned about my relationships		.71			
I felt hurt or disappointed by how others behaved		.70			
I felt sad about how I acted towards others		.55			
I felt a loss of inspiration or spiritual direction			.88		
I felt distant in my relationship with God or my Higher Power			.80		
I felt concerned about my religious or spiritual life			.75		
I felt guilt over mistakes that were inconsistent with my religious beliefs			.75		
I felt anxious about beginning therapy				.90	
I had concerns about beginning therapy				.87	
I felt uncertain about whether I can be fully open with my therapist				.70	
I had doubts about whether my therapist will understand my concerns				.67	
I felt physically well and healthy					-.85
I felt light headed, weak, or fatigued					.82
I had a stomach ache or other gastro-intestinal problems					.74
I experienced physical pain or discomfort					.68

Note. PD=Psychological Distress; RD=Relationship Concerns; SD=Spiritual Concerns; TE=Therapy Expectations; PC=Physical Concerns. All loadings were significant.

Convergent validity. The 5 dimensions of the CAMOS short form were compared to several existing measures in sample 2 in order to understand in greater depth what each

dimension was assessing. Table 4 shows the correlations of the CAMOS short form dimensions with measures that were hypothesized to assess similar constructs as well as some that were theorized to be different than the dimension. The following section also provides a conceptual analysis of the constructs assessed in the context of

Psychological Distress. The Psychological Distress (PsD) dimension demonstrated strong correlations with the OQ-45 Total (.812) score. This indicates that PsD accounts for approximately 66% of the variance of the OQ-45 and provides strong evidence for its utility as a valid measure for monitoring outcome in everyday practice. The strong correlations with the MMPI-II RF demoralization scale (.661) indicate that this dimension may be assessing a client's degree of general distress and feelings of inability to cope with stress. PsD also demonstrated moderately high correlations with the low positive emotions scale (.538) which suggests that it may pick up on anhedonic or depressive symptoms in addition to demoralization. The moderately high correlation with the dysfunctional negative emotions scale (.585) also provides evidence that PsD is likely assessing anxiety, anger, or general emotional reactivity.

Table 4

CAMOS factor correlations

	PD	TE	PHD	SD	RD
Sample 1					
Psychological Distress (PD)	-				
Therapy Expectations (TE)	.44	--			
Physical Health Distress (PH)	.62	.32	--		
Spiritual Distress (SD)	.63	.17	.37	--	
Relationship Distress (RD)	.72	.38	.50	.46	--
Sample 2					
Psychological Distress (PD)	-				
Therapy Expectations (TE)	.64	--			
Physical Health Distress (PH)	.73	.55	--		
Spiritual Distress (SD)	.65	.44	.56	--	
Relationship Distress (RD)	.73	.57	.54	.54	--

Additionally, this dimension demonstrated high correlations with the MMPI-II's Anxiety (.670) and Depression (.665) subscales. Overall, these results suggest that PsD is assessing demoralization as well as symptoms of anxiety and depression. This dimension had low correlations with the MMPI-2 RF Antisocial Behavior (.143) and Ideas of Persecution (.202) scales, suggesting that it does not assess paranoid or antisocial symptoms.

Relationship Distress. The Relationship Distress (RD) dimension demonstrated moderate correlations with several measures. Its moderately high correlation with the OQ-45 Interpersonal Relations score (.572) provides evidence that it is assessing general relationship concerns. Additionally, its moderate correlation with the MMPI-2 Social Alienation scale (.498) indicates that RD is tapping into a person's feelings of loneliness and of potentially feeling unloved or misunderstood. This is further corroborated by a large correlation with the MSEI

Lovability (-.600) scale. The moderate correlation with the MMPI-II Family Concerns subscale (.447) provides some evidence that RD may be assessing the degree of perceived social support the client experiences. Given that there was only one large correlation, it may be necessary to examine this further, however, there is evidence that this dimension is in fact assessing relationship distress. Many ROM systems, however, have had difficulty obtaining large correlations of relationship concern dimensions with existing measures of interpersonal problems (Lambert et al., 1996; Kraus et al. 2005; Evans, 2012). Its low correlations with the MMPI-II RF Low Positive Emotions scale (.222) indicate that the dimension is not assessing depressive or anhedonic symptoms.

Physical Health Concerns (PHC). The PHC dimension had a large correlation with the MMPI-2 RF Somatic Complaints (.619) as well as the MMPI-II Health Concerns scale (.616) scales. This suggests that PHC is assessing physical discomfort or illness, but also possibly assessing somatic or conversion symptoms. Additionally, PHC demonstrated a moderate correlation with the MMPI-2 Physical Malfunctioning subscale (.433), which suggests that it may be assessing some of the physical symptoms associated with depression. As expected, this dimension does not assess psychotic or antisocial tendencies, as evidenced by low correlations with the MMPI-II Bizarre Mentation subscale (.118) and MMPI-II RF Antisocial Behavior scale (.101). These results suggest that PHC appears to be assessing distress due to physical health or psychosomatic concerns.

Spiritual Concerns (SC). The SC correlations were performed on sample 1 given that this dimension on sample 2 had a problematic response set for those who reported not being spiritual. The SC dimension had a moderately large correlation with the TSOS Love of God scale (-.56) indicating that it is assessing a client's perceived relationship with God as well as

spiritual purpose in life. The highest correlation was with the TSOS Love of Self scale (-.668), which provides evidence that the SC dimension is also assessing the client's feelings of spiritual worth and congruence with moral values. This dimension had very low correlations with the MMPI-II RF Abberant Experiences (.074) and Antisocial Behavior (.115) scales. Overall, these findings suggest that SC is assessing client's spiritual concerns.

Therapy Expectations (TE). TE had a moderate correlation (.445) with the Negative Treatment Indicators (TRT) subscale of the MMPI-II. The lack of a large correlation is not entirely unexpected, as the TRT subscale has broader aims than was desired for the TE dimension. Thus, aspects such as pessimism about therapist understanding and fear of self-disclosure may not correlate very well with the TE dimension, but aspects such as fatalism and avoidance of responsibility may not be as relevant for the TE dimension. This dimension did not correlate highly with the OQ social roles scale (.238) or MMPI-II RF low positive emotions scale (.288).

Table 5

Convergent Validity Correlations

	PsD	RD	SD	PhD	TE
OQ-45 Total	.812	.543	.487	.662	.476
MMPI-II RF Low Positive Emotions	.538	.412	.362	.466	.288
MMPI II RF Dysfunctional Negative Emotions	.585	.480	.347	.394	.341
MMPI II RF Demoralization	.661	.412	.457	.488	.375
MMPI II Obsessiveness	.543	.413	.377	.404	.335
OQ-45 Interpersonal relations	.621	.572	.449	.482	.341
MMPI-II Social Alienation	.397	.498	.259	.248	.311
MMPI-II Family Discord	.239	.449	.172	.131	.131

MMPI II Family Problems	.298	.447	.160	.238	.213
MSEI Lovability	-.647	-.608	-.397	-.495	-.496
MMPI-II RF Somatic Complaints	.443	.404	.337	.619	.288
MMPI-II Health Concerns	.456	.465	.355	.616	.309
MMPI-II Physical Malfunctioning	.398	.335	.308	.539	.209
TSOS Love of God	-.427	-.344	-.560	-.151	-.350
TSOS Love of Self	-.561	-.391	-.668	-.229	-.295
MMPI-II Negative Treatment Indicators	.520	.375	.329	.392	.445

Note. Expected theoretical correlation bolded. MMPI-II= Minnesota Multiphasic Personality Inventory 2, MMPI-II RF = Minnesota Multiphasic Personality Inventory 2 Restructured Form, OQ-45= Outcome Questionnaire 45.

Discussion

The present study found strong evidence to support the reliability and validity of the five factor 25-item short form of the CAMOS. The short form demonstrated excellent model fit in samples from a college counseling center and inpatient eating disorder clinic. The factor loadings were sufficiently large to suggest strong convergent validity, and the ratio of average squared factor loadings to maximum squared correlations between dimensions suggested that all but one dimension (Relationship Concerns) in one sample did not demonstrate a problematic overlap with another dimension. All but one dimension demonstrated strong correlations (>.60) with a theoretically related measure, with the only scale that did not (Therapy Expectations), demonstrating a moderate correlation with a criterion measure. Every short form dimension demonstrated strong internal consistency reliability with both samples, with all being above .8 on two separate indexes. All short form dimensions correlated with their full form counterpart at greater than .9, suggesting that the CAMOS short form is a suitable alternative to the full form.

With the removal of items that were highly positively skewed, only items that allowed for improvement for most clients were maintained.

With this short form in place, the CAMOS now has a brief standardized core that has strong evidence of reliability and validity. This short form model can serve as required items that are administered to clients on a session by session basis, thus providing a standardized routine outcome monitoring system. The items that were not used in the short form can serve the purpose of optional items that can be added to the CAMOS if the clinician feels that they are important to assess. This potentially resolves the problem of how to include items that are of obvious clinical utility (i.e., items about suicidal or homicidal ideation), but do not fit well into specific factors (Kraus et al., 2005). This approach is similar to hybrid measurement systems such as the Psychological Outcome Profiles (Ashworth et al., 2007), and the Individualized Patient Progress System (IPPS; Sales & Alves, 2012), in that it combines global and individualized assessment approaches, but with the short form model, the CAMOS offers several advantages relative to these systems. One advantage is that the CAMOS allows for a therapist to create a tailored measure for all of their clients with just a one-time five minute investment of their time. Once the items for the template have been selected and authored, the therapist or treatment site could opt not invest any more time, and the measure will be adapted to their needs. An additional benefit is that if client-level tailoring is desired, the therapist can select from the existing item pool, and add items with a clicking of a check box. These features allow the CAMOS to adapt to the needs of individual therapists and clients, while substantially decreasing the time burden to both client and therapist relative to hybrid measures. The development of the short form of the CAMOS is what allows for this tailoring system while still being feasible for session-by-session administration. Thus, the CAMOS's tailoring system offers a promising

alternative to hybrid systems in that it provides a more structured and efficient way of tailoring, while still giving clinicians the ability to author items if they so desire.

The tailoring feature could be a step toward resolving the issue of having an adequate balance of providing a clinically meaningful and psychometrically sound assessment while still keeping the measure brief. In the ROM literature, the emphasis appears to be on treatment sites finding an outcome measure that balances relevance and practicality in the context of their site (Boswell et al., 2015; Tarescavage & Ben Porath, 2014). These approaches, however, do not account for the possibility that individual therapists may have different preferences for the type and amount of information desired, and that these preferences could change relative to the needs of specific clients. Thus, even if a site is able to reach a consensus of which measure most therapists at their site found appropriate, which some, such as Meehan and colleagues (2006), have found to be a difficult task, it would still ignore individual needs. The clinically adaptive features of the CAMOS address this issue by allowing the measure to adapt to specific sites, therapists, and clients.

With this blend of gathering standardized data and easy tailoring options, new ways of implementing ROM systems become available that could help resolve the impasse between ROM developers and clinicians. With this system, clinician concerns about relevance no longer need to be seen as “obstacles” to be overcome, but as valuable information that can be incorporated into the measurement system (Boswell et al., 2015). Additionally, the results of the present study suggest that this accommodation would not come at the expense of comparability. Instead of researchers needing to convince clinicians that their measures are relevant, and minimizing aspects that are not included in the system, they can encourage clinicians to adapt assessment and make it relevant to the populations and individual clients that they serve. This

would be more in line with the recent push in the ROM literature to establish more bottom-up implementation procedures (Mellor-Clark et al., 2014), as well as developments in the organizational behavior literature about the detrimental effects of demonizing “resistance” (Ford et al., 2008; Thomas & Hardy, 2011).

This could change the narrative of ROM implementation efforts from one of heroic researchers attempting to overcome the irrational objections of clinicians in order to bring about scientifically informed practice, to one of collaboration. With existing ROM systems, the possibilities for adaptation are limited to primarily logistical concerns (Boswell et al., 2015). The individual therapist has no say in defining the items and constructs used to define their treatment outcomes, nor in determining the length of the survey. In contrast, the philosophy and features of the CAMOS can encourage and even seek after individual input from sites and clinicians, without threatening the legitimacy of the assessment system. As this information is added by clinicians, it could potentially help to expand the researcher’s views of what kinds of outcomes are valued by clients and therapists, leading to further refinements of the standardized items. In this approach, the researcher is no longer imposing a definition of outcome, but providing a platform that allows for researchers and clinicians and clients to co-create the definition of outcome.

The creation of the CAMOS short form allows for these changes, while still offering the major features that make ROM systems so desirable. The results of the present study provide evidence that the CAMOS can be used to evaluate outcomes in routine practice, and to provide aggregate data that can be used to evaluate the efficacy of treatment for large groups of clients and therapists. The evidence gathered in this study for the reliability and validity of the short form is a first step toward generating expected treatment response curves once more data is

obtained (Lueger et al., 2001). It is hoped that, similar to the studies performed by Lambert and colleagues, the CAMOS can help to decrease the incidence of client deterioration throughout the course of treatment (Shimokawa et al., 2010). In addition to providing the benefits associated with traditional ROM systems, it is hoped that the clinically adaptive features of the CAMOS will allow it to provide additional benefits.

In terms of evaluating the efficacy of treatment for specific sites, the CAMOS could allow for a more nuanced picture of the benefits of treatment. In addition to evaluating the standardized portion of the measure, administrators could qualitatively examine the results of the therapist's unique items, to see if there are benefits to the clients that were not captured within the relatively narrow scope of the global items. This could allow clinicians to have a voice in defining the types of outcomes by which they will be evaluated and facilitate productive dialogue about the clinician's performance. At an individual client level, the CAMOS tailoring features could allow for enhanced responsiveness in treatment (Stiles, 2013). As the measure adapts to the needs of the client, it is possible that the client could perceive the therapist as having a greater understanding of their concerns, due to the measure adapting to their needs. It could also help therapists to see if the types of outcomes and processes that they are viewing as important to treatment, are in fact valued by the client.

Finally, if clinicians have access to data that is directly relevant to their therapeutic approach, and to the specific context of the client, it could allow for assessment that more directly informs treatment. Several studies have documented how few clinicians use the reports associated with ROM systems (Garland et al., 2003; Trauer, 2010b). This may be due to the atheoretical nature of the measures, which provides primarily descriptive information about the client (Persons, 1991). With the CAMOS's approach, the therapist could assess areas directly

relevant to the specific co-created outcomes with individual clients, as well as more theory-specific outcomes and processes. This relevant and easy to tailor information could help to increase the usage of ROM reports by clinicians, and thus increase existing benefits of ROMs as well as potentially offer new advantages.

If the usage rates of ROMs can be increased, as well as the quality of the data received from ROMs improved, this could lead to a more full realization of the goals of Evidence Based Practice in Psychology (EBPP). ROM is viewed as a crucial aspect of the EBPP vision (APA Presidential Task Force on Evidence-Based Practice, 2006). If the CAMOS is successful in increasing clinician ROM usage, and broadening the scope of the outcomes measured, then it could be a vital contributor to EBPP. Current ROM models rely solely on research (with input from a few clinicians) to define the outcome criteria by which therapeutic encounters are evaluated, with the integration of clinical expertise and client culture and values occurring after the measurement process is complete (Boswell et al., 2015). The tailoring features made possible by the development of the short form of the CAMOS allow clinical judgments and client culture to be more than post-hoc factors that are used to contextualize the results of ROM, but integral aspects of the definition of outcome. In our view, this approach is more consistent with the spirit of EBPP, as well as with the growing emphasis on multiculturalism in the field of psychotherapy research (Lee, 2014). Thus, the CAMOS's tailoring features are more than just a gimmick to attempt to get clinicians to start using ROM, but a way to facilitate the integration of the principles of EBPP into assessment as well as treatment.

Limitations

One limitation of this study was that it used the same samples as McBride's analysis. In order to minimize capitalization on chance, the samples were divided up differently for the

present study. In McBride's analysis, the samples were stratified by gender, then combined and split. In the current study, the samples were left separate, with more exploratory analyses performed on the college counseling center sample, and the final CFA being performed on the inpatient eating disorder sample. Although this distinction did provide some differentiation, the fact remains that the same participants were used in both studies. Also, one of the treatment sites was a female-only inpatient eating disorder facility, thus, the overall sample was likely under-representative of males. This limitation is somewhat offset by the finding that the factor structure also held up with the university counseling center sample, which had a substantial number of male participants, but additional replication with more clinical samples may be necessary. Finally, the samples were relatively homogenous in terms of their racial identification, thus additional cross validation with more diverse populations is necessary. Despite these concerns about the sample, the tailoring features of the CAMOS allow for items that were not included in the short form, but may be relevant for specific groups, to be added back into the measure easily. Thus, the items removed from both forms of the CAMOS can still be accessed and used in clinical settings.

Another limitation is the cross-sectional nature of the data. One of the primary purposes of the CAMOS is to be a measure of therapeutic change. This study addressed sensitivity to change in the sense of removing items with little room for improvement, however, additional longitudinal studies will need to examine whether the items on the short form demonstrate change in response to relevant intervention. Such longitudinal analyses, however, were beyond the scope of the present study.

The removal of the Work Distress dimension from the final short form model is a limitation in the theoretical and clinical utility of the short form. There are several possible

reasons for the inability of the Work Distress dimension to hold up well in the current study. The first is that occupational and academic functioning may in fact be two separate constructs. Thus, it may be necessary to develop separate academic and employment concerns dimensions in order to account for this possibility. It is also possible that occupational functioning was not a highly relevant variable for the populations that were selected for this study. Many college students are not employed due to their academic demands, and patients at an inpatient treatment center are often not working. Thus, it could also be the case that attempting to validate this dimension on a demographic that is more likely to be employed could yield more positive findings. The difficulty in validating an occupational functioning dimension, however, is not unique to the CAMOS, as other ROM systems have struggled to create a stable measure of social role functioning (Kraus et al. 2005; Lambert et al., 1996; Tarescavage & Ben-Porath, 2014).

Future Directions

Studies are currently underway to cross validate the factor structure of both forms of the CAMOS in more culturally diverse populations and in different types of treatment settings. Data has already been gathered in order to replicate the factor structure on a non-clinical sample, and to create indexes of clinically significant change (Jacobson & Truax, 1993). Initial results are promising that the factor structure fits with a non-clinical sample, and that the CAMOS dimensions can differentiate between clinical and non-clinical populations. These studies should also test the invariance of the parameters of the CFA model in different samples and with different groups. Additional studies are also underway to examine the stability of the factor structure across repeated administrations, as well as to evaluate the CAMOS's sensitivity to change. These will be essential to the validity of the CAMOS, in that it was designed to be a

measure of change as a result of therapy. These studies will also need to provide support for the Therapy Progress dimension that begins after the first session.

A revision of the Work Distress dimension will be necessary in order to make it both clinically relevant and psychometrically sound. An updated Work Distress and a new School Distress scale are currently in development. Initial findings look promising for improved psychometric performance by separating occupational and academic functioning into two separate scales. Additionally, a new scale of cultural distress, related to perceived microaggressions, is currently in development and being tested in various treatment sites throughout the country. This could be a step forward in helping the CAMOS to potentially become a tool for enhancing multicultural awareness and competence in clinicians.

Finally, the most important consideration of the value of the CAMOS will be evaluating whether it can be both feasible and relevant in clinical practice. Additional research should examine ways to optimize reporting systems so that they provide the most useful information to clinicians in as concise a manner as possible. Also, examining clinician usage rates of various CAMOS features will be highly important to determining whether the philosophy and features of the CAMOS do in fact lead to greater clinician acceptance of a ROM system in clinical practice. Additional studies will also need to examine whether the use of CAMOS reports improves outcomes, and how these results compare to previous findings about the benefits of using ROMs in clinical practice.

Conclusion

This study found support for the reliability and validity of a short form of the CAMOS. This short form provides the psychometric foundation needed to allow for accurate aggregate data analysis on a session-by-session basis, while leaving room for additional clinician tailoring.

Future psychometric studies will need to address the issues of clinical significance and sensitivity to change, but this study provides a strong base from which to proceed with these types of analyses. Overall, these results show that the CAMOS short form is a promising measure that can hopefully be facilitative of more widespread implementation of ROM systems while still empowering the voice of the clinician and client.

References

- Abrahamson, D. (1999). Outcomes, guidelines, and manuals: On leading horses to water. *Clinical Psychology-Science and Practice, 6*, 467-471.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC:
- APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.
- Ashworth, M., Robinson, S., Evans, C., Shepherd, M., Connolly, A., & Rowlands, G. (2007). What does an idiographic measure (PSYCHLOPS) tell us about the spectrum of psychological issues and scores on a nomothetic measure (CORE-OND)? *Primary Care and Community Psychiatry, 12*, 7-16.
- Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Beutler, L. E. (2001). Comparisons among quality assurance systems: From outcome assessment to clinical utility. *Journal of Consulting and Clinical Psychology, 69*, 197-204.
- Bickman, L., Rosof-Williams, J., Salzer, M. S., Summerfelt, W. T., Noser, K., Wilson, S. J., & Karver, M. S. (2000). What information do clinicians value for monitoring adolescent client progress and outcomes? *Professional Psychology: Research and Practice, 31*, 70-74.
- Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*, 6-19.

- Bower, P. & Gilbody, S. (2010). The current view of evidence and evidence-based practice. In M. Barkham, G. E. Hardy & J. Mellor-Clark (Eds.), *Developing and delivering practice-based evidence: A guide for the psychological therapies* (pp. 1-19). West Sussex, UK: John Wiley & Sons.
- Browne, G. (2006). Outcome measures: Do they fit with a recovery model? *International Journal of Mental Health Nursing, 15*, 153-154.
- Butcher, J. N. & Pope, K. S. (1992). The research base, psychometric properties, and clinical uses of the MMPI-2 and MMPI-A. *Canadian Psychology/Psychologie Canadienne, 33*, 61-78.
- Carlier, I. V. E., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van, d. W., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: Evidence and theory. *Journal of Evaluation in Clinical Practice, 18*, 104-110.
- Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., . . . Haaga, D. A. (1998). Update on empirically validated therapies, II. *Clinical Psychologist, 51*, 3-16.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
- Cudeck, R. & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*, 147-167.
- Duncan, B. L. (2012). The partners for change outcome management system (PCOMS): The heart and soul of change project. *Canadian Psychology, 53*, 93-104.
- Evans, C. (2012). Cautionary notes on power steering for psychotherapy. *Canadian Psychology/Psychologie Canadienne, 53*, 131-139.

- Fabrigar, L. R. (2012). *Exploratory Factor Analysis*. New York, NY: Oxford University Press.
- Ford, J. D., Ford, L. W., & D'Amelio, A. (2008). Resistance to change: The rest of the story. *The Academy of Management Review*, *33*, 362-377.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50.
- Garland, A. F., Kruse, M., & Aarons, G. A. (2003). Clinicians and outcome measurement: What's the use? *Journal of Behavioral Health Services & Research*, *30*, 393-405.
- Gilbody, S. M., House, A. O., & Sheldon, T. A. (2002). Psychiatrists in the UK do not use outcome measures. *The British Journal of Psychiatry*, *180*, 101-103.
- Goodman, J. D., McKay, J. R., & DePhilippis, D. (2013). Progress monitoring in mental health and addiction treatment: A means of improving care. *Professional Psychology: Research and Practice*, *44*, 231-246.
- Hair, J., Black, W., Babin, B., and Anderson, R. (2010). *Multivariate data analysis* (7th ed.): Prentice-Hall, Inc. Upper Saddle River, NJ, USA.
- Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*, 155-163.
- Happell, B. (2008). Meaningful information or a bureaucratic exercise? Exploring the value of routine outcome measurement in mental health. *Issues in Mental Health Nursing*, *29*, 1098-1114.
- Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice*, *35*, 485-491.

- Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research, 34*, 283-291.
- Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*, 191-205.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 32*, 179-185.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
- Kenny, D. A. (1979). *Correlation and causality*. New York, NY: Wiley.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling (3rd ed.)* New York, NY: Guilford Press.
- Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61*, 285-314.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.
- Lakeman, R. (2004). Standardized routine outcome measurement: Pot holes in the road to recovery. *International Journal of Mental Health Nursing, 13*, 210-215.

- Lambert, M., Burlingame, G., Umphress, V., Hansen, N., Vermeersch, D., Clouse, G., & Yanchar, S. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, 3, 249-258.
- Lambert, M. J., Okiishi, J. C., Finch, A. E., & Johnson, L. D. (1998). Outcome assessment: From conceptualization to implementation. *Professional Psychology: Research and Practice*, 29, 63-70.
- Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, 10, 288-301.
- Lee, C. C. (Ed.). (2014). *Multicultural issues in counseling: New approaches to diversity*. Alexandria, VA: American Counseling Association.
- Levant, R. F., & Hasan, N. T. (2008). Evidence-based practice in psychology. *Professional Psychology: Research and Practice*, 39, 658-662.
- Lueger, R. J., Hoard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology*, 69, 150-158.
- Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Benler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, 3, 320-341.
- McBride, J. A. (2016). *Initial development and validation of the clinically adaptive multidimensional outcome survey* (Unpublished doctoral dissertation). Brigham Young University, Provo, UT.

- Meehan, T., McCombes, S., Hatzipetrou, L., & Catchpoole, R. (2006). Introduction of routine outcome measures: Staff reactions and issues for consideration. *Journal of Psychiatric and Mental Health Nursing, 13*, 581-587.
- Mellor-Clark, J., Cross, S., Macdonald, J., & Skjulsvik, T. (2014). Leading horses to water: Lessons from a decade of helping psychological therapy services use routine outcome measurement to improve practice. *Administration and Policy in Mental Health and Mental Health Services Research, 1-7*.
- Muthén, L.K. and Muthén, B. O. (2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.
- O'Brien, E. J. & Epstein, S. (1988). *MSEI: Multidimensional Self-Esteem Inventory*. Odessa, FL: Psychological Assessment Resources.
- Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcome assessment*. New York, NY: John Wiley & Sons, Inc.
- Ogles, B. M., Lambert, M. J., & Masters, K. (1996). *Assessing outcome in clinical practice*. New York, NY: Allyn and Bacon.
- Persons, J. B. (1991). Psychotherapy outcome studies do not accurately represent current models of psychotherapy: A proposed remedy. *American Psychologist, 46*, 99-106.
- Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184.
- Raykov, T. (1997b). Scale Reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence for fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-354.

- Raykov, T., Rodenberg, C., & Narayanan, A. (2015). Optimal shortening of multiple-component measuring instruments: A latent variable modeling procedure. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 227-235.
- Richards, P. S., & Bergin, A. E. (2005). *A spiritual strategy for counseling and psychotherapy*. Washington, DC: American Psychological Association.
- Richards, P. S., Smith, T. B., Schowalter, M., Richard, M., Berrett, M. E., & Hardman, R. K. (2005). Development and validation of the theistic spiritual outcome survey. *Psychotherapy Research*, 15, 457-469.
- Sales, C., & Alves, P.C. (2012). Individualized patient-progress systems: Why we need to move towards a personalized evaluation of psychological treatments. *Canadian Psychology/Psychologie Canadienne*, 53, 115-121.
- Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, 78, 298-311.
- Slade, M. (2002). What outcomes to measure in routine mental health services, and how to assess them: A systematic review. *Australian & New Zealand Journal of Psychiatry*, 36, 743-753.
- Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, 12, 102-111.
- Spring, B. (2007) Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of Clinical Psychology*, 63, 611-631.

- Streiner D. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 99-103.
- Stuart, R. B., & Lilienfeld, S. O. (2007). The evidence missing from evidence-based practice. *American Psychologist*, 62, 615-616.
- Stiles, W.B. (2013). The variables problem and progress in psychotherapy research. *Psychotherapy*, 50, 33-41.
- Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement & Evaluation in Counseling & Development*, 37, 240-256.
- Tarescavage, A. M., & Ben-Porath, Y. S. (2014). Psychotherapeutic outcomes measures: A critical review for practitioners. *Journal of Clinical Psychology*, 70, 808-830.
- Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist*, 48, 3-23.
- Thomas, R., & Hardy, C. (2011). Reframing resistance to organizational change. *Scandinavian Journal of Management*, 27, 322-331.
- Trauer, T. (2010a). Introduction. In T. Trauer (Ed.), *Outcome measurement in mental health: Theory and Practice* (pp. 1-11). Cambridge, GBR: Cambridge University Press.
- Trauer, T. (2010b). Stakeholder perspectives in outcome measurement. In T. Trauer (Ed.), *Outcome measurement in mental health: Theory and practice* (pp. 196-205). Cambridge, GBR: Cambridge University Press.
- van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, Netherlands: Kluwer Academic.

- van Prooijen, J. & van der Kloot, W.A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement, 61*, 777-792.
- Walfish, S., McAlister, B., O'donnel, P., & Lambert, M. J. (2012). A Investigation Of Self-Assessment Bias in Mental Health Providers. *Psychological Reports, 110*(2), 639-644.
- Walker, B. B., & London, S. (2007). Novel tools and resources for evidence-based practice in psychology. *Journal of Clinical Psychology, 63*, 633-642.
- Wampold, B. E., Goodheart, C., & Levant, R. (2007). Clarification and elaboration on evidence-based practice in psychology. *American Psychologist, 62*, 616-618.
- Yu, C. Y., & Muthen, B. (2002, April). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. In *annual meeting of the American Educational Research Association, New Orleans, LA.*

APPENDIX A: Literature Review

Evidence-Based Practice

In the late 1980's and early 1990's a movement called Evidence-Based Practice (EBP) began to gain considerable influence in the health care professions (Walker & London, 2007). Initially beginning in the field of medicine under the title "Evidence Based Medicine", this movement called for health care practitioners to provide evidence that the treatments they were providing had scientific evidence to back their usage (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). Some of the major motivating factors behind this movement were: the idea that healthcare interventions should be applications of scientific knowledge, a desire to control costs and have evidence that the procedure being paid for (by the patient or insurance company) was necessary, and an ethical imperative to provide the best possible treatment to patients (Bower & Gilbody 2010).

Soon these ideas began to permeate psychotherapy, and questions began to be asked about which types of psychotherapy were most effective for which disorders and whether clinicians were using these (Task Force on Promotion and Dissemination of Psychological Procedures, 1995). Although psychotherapy had research to support its general efficacy as an intervention (Smith & Glass, 1977), there was some concern that the therapies that were used in the research were not being used in clinical practice (Wampold, 2001). Therefore, it was possible that clinicians were doing therapy that did not have scientific support. This was especially problematic because insurance companies would likely not be willing to pay for interventions that had no evidence of their effectiveness. Even if the intervention did work, if there was another that was shown to be superior, then that should be the one used.

Empirically-Supported Treatments. In 1995, with the legitimacy and funding of psychotherapy hanging in the balance, the Division of Clinical Psychology of the American Psychological Association (APA) began an initiative called Empirically-Supported Treatments (EST; Task Force on Promotion and Dissemination of Psychological Procedures, 1995). The goal of this initiative was to meet the demands of EBP by identifying the treatments that had empirical evidence to support their usage with specific diagnoses and to ensure that these treatments were used by clinicians. Therefore, a list of treatments that had empirical support for specific disorders was assembled, and clinicians were urged to use these in order to receive insurance funding and to practice ethically. This list would be updated frequently as new innovations and findings emerged (Chambless et al., 1998). The criteria for a treatment to be seen as empirically supported were:

1. At least two Randomized Controlled Trials (RCT) demonstrating that the treatment is (a) superior to placebo or another treatment, (b) equivalent to an existing EST. Or at least nine Single-N experimental studies showing efficacy of the approach.
2. Studies must use treatment manuals
3. Characteristics of the client samples must be clearly specified
4. Experiments must have been done by at least two different investigators or teams.

With the establishment of the EST paradigm, it appeared that some of the major pressures of EBP had been addressed: Practitioners would primarily use treatments that had firm evidence to back their usage for specific problems, funders and consumers could know that their money was being spent on something that was proven to be effective. Additionally, it would provide clear guidelines for ethical practice, and would bring a rigor used in pharmaceutical trials to psychotherapy research and practice

The EST paradigm was met with substantial backlash from the field. The primary objection centered around the validity of the criteria for inclusion in the Empirically Supported Treatment list (Elliot, 1998). Some of these concerns included the following: No empirical proof existed that RCTs had external validity in treating the specified disorders (Westen, Novotny, & Thompson-Brenner, 2004), the criteria were biased against certain kinds of therapeutic approaches such as humanistic and psychodynamic therapies (Bohart, O'Hara, & Leitner, 1998), methodological criticisms of RCT as a means of establishing the efficacy of treatments (Wampold, 2001), arguments that the criteria were too lenient, and claims that the selection of the criteria represented a political process of creating ideological and methodological monopolies (Smith, 2009).

Evidence-Based Practice in Psychology. This debate continued for nearly a decade before the APA took an official stance on a new policy. The new policy, called Evidence Based Practice in Psychology (EBPP), defined EBP as “the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences.” (American Psychological Association Presidential Task Force on Evidence-Based Practice, 2005, p.1). This approach represented a substantial departure from the EST approach and addressed many of the concerns with the EST paradigm. In the EBPP stance, research still held significant sway in terms of its importance in clinical decision making; however, it was no longer the sole determinant in deciding which treatments should be used. The definition of EBPP added “clinical expertise” and “patient characteristics, culture, and preferences” as essential elements to clinical decision making. The report defined clinical expertise as the ability to use research, training, and previous experience to tailor treatment to the individual. Some of the components of clinical expertise identified in the report are: treatment planning, implementing interventions,

monitoring client progress, adapting to client's culture and preferences, and continually acquiring new skills. Although research is not separate from these tasks, there is considerable clinical skill involved for the therapist to synthesize and apply research and clinical observations to a specific client's situation.

In addition, EBPP incorporated a broader definition of what qualifies as evidence (Wendt & Slife, 2007). In fact, the task force report purposely did not include specific guidelines for what qualifies as evidence for efficacy (Wampold, Goodheart, & Levant, 2007). In the report, the task force listed various types of research that would be considered evidence to support the use of an intervention such as: effectiveness and efficacy research, cost effectiveness, treatment utilization, and more. In addition several research methodologies were outlined, although they were not meant to be an exhaustive list: clinical observation, qualitative research, RCTs, single-n, meta-analysis and more. It is clear to see that EBPP took some of the criticisms of the EST paradigm about the narrowness of the inclusion criteria very seriously and used these to expand the definition of what qualified as evidence.

Although many lauded this new paradigm for its increased flexibility and responsiveness to the criticisms of the EST paradigm, there was still considerable debate about this report (Wampold et al., 2007). One of the major critiques of this report was that it did not provide a concrete definition of what constitutes sufficient evidence to justify the use of a practice (Stuart & Lilienfield, 2007). The EST model had a very clear and concrete definition of what constituted sufficient evidence for the use of a treatment: if the treatment demonstrated in two RCTs that it was superior to placebo, it was a good treatment. In EBPP, there were no specific criteria, but general principles of how the clinician should make decisions. However, with

principles instead of guidelines, it becomes more difficult to keep clinicians accountable, and to provide firm rules about what kinds of interventions should be funded by insurance companies.

In responding to this issue, Wampold and colleagues (2007) pointed out that they intentionally did not give criteria for what evidence and how much of it is needed for a specific treatment to be justified in use. EBPP was not designed to provide prescriptive guidelines for clinicians. They discuss how, instead of only identifying treatments that work in laboratory setting, that clinicians should engage in routine outcome measurement and monitor treatment progress. Whereas EST's solution to the issue of accountability was to identify research supported treatments, and have clinicians use those exclusively, EBPP emphasizes the importance of looking at what is already being done in practice in addition to using research support treatments. If clinicians could provide evidence that their interventions were providing measurable benefit to clients, then they have evidence to support their use of that intervention. Additionally, clinicians would be able to examine in real-time if their interventions were not working, and potentially change them.

Routine Outcome Monitoring

As of the time of the EBPP report, several measures existed that were designed to accomplish the goals outlined in the report (Stuart & Lilienfield, 2007). This paradigm of measures has been called by many names (Routine Outcome Measures, Routine Outcome Monitoring Systems, Progress Monitoring), but for the sake of this paper they will be called Routine Outcome Monitoring (ROM) systems. The literature has focused primarily on two ways that the ROMs could ensure and possibly improve the quality of treatment: (a) Identifying individual clients at risk for treatment failure and (b) examining the efficacy of therapists and

treatment sites in the treatment of large groups of clients (Shimokawa, Lambert, & Smart, 2010; Kraus, Castonguay, Boswell, Nordberg, & Hayes, 2011).

Improving individual treatment. In 1996, Howard and colleagues published an article in *American Psychologist* that laid both the conceptual and practical framework on which most ROM systems are based. They discussed how efficacy research has shown evidence of interventions that are generally effective, but that this does not help clinicians with decision making about the treatment of individual clients. With these considerations in mind, they outlined a system that could potentially use the results of research and make them applicable at the level of the individual client. Their solution rested on their previous work about the dose-response relationship in psychotherapy (Howard, Kopta, Krause, & Orlinsky, 1986). This model essentially posits that there exists a log linear relationship between the number of sessions and outcome, and that treatment involves three phases: remoralization, remediation (symptom reduction), and rehabilitation (improved functioning). These phases are sequential and if the outcomes for each phase are not met by certain sessions, then the likelihood of treatment failure increases (Howard et al., 1993). In sum, the course of a client's treatment response could be predicted.

Expected Treatment Response & Clinically Significant Change. If outcomes could be predicted, then it could be possible to provide what Howard and colleagues called Expected Treatment Response (ETR) curves that would allow clinicians to compare how an individual client is progressing in treatment relative to other clients similar to them. This concept was incorporated into a measure developed by Howard and colleagues called the COMPASS (Howard, Lueger, Maling, & Martinovich, 1993) that assessed areas relating to the dose-response model, and used these and 18 additional case mix variables, later reduced to 7 by Lutz,

Martinovich & Howard (1999), to form individualized ETRs for clients using Hierarchical Linear Modeling procedures. A study by Leon, Kopta, Howard and Lutz (1999) examined the question of whether the predicted ETR values matched observed data. They found that for 75% of the ETRs, the predicted score was similar (the difference was less than measurement error) to the observed scores, and found correlations of .57 between ETR slopes and observed slopes. They also identified several moderating and mediating variables that had differential predictive power in terms of predicting treatment failure. Lueger and colleagues (2001) provide a more comprehensive review of the work of Howard's team in developing the ETR technology, and is recommended for interested readers.

In addition to the ETR system the COMPASS incorporated another concept that is foundational to most ROMs, clinical significance. The concept of clinical significance originated in the work of Jacobson and Truax (1991), which challenged the notion that statistical significance was equivalent to clinically meaningful change. In contrast to depending on statistical significance or effect sizes, the authors propose two alternative methods: Clinical Cutoffs and Reliable Change Indexes. Clinical cutoffs were defined as the score on the outcome measure at which the client's score is more probable in the distribution of a non-clinical sample than that of a clinical sample. This provides stronger evidence of improvement than relying only on statistical significance because it is tied to the 'real world' distinction between those who are in treatment and those who are not. One situation where this procedure could be inadequate is if a client's initial score is in the clinical range by a very narrow margin, and they drop to be narrowly in the normative range by the end of treatment. Using the clinical cutoff criteria, the client would be deemed to have improved despite showing very little change in their symptoms. In order to account for this, Jacobson and Truax developed the Reliable Change Index (RCI),

which provides a minimum value that a person must change in order to ensure that it is not merely the result of measurement error. This value is a function of the reliability of the measure, and suggests change that is above and beyond measurement error. With these two statistics, Jacobson and Truax provided a means for measures to be more relevant to the reality of clinicians.

The COMPASS incorporated both ETR and the concept of clinical significance into the measure, thus providing multiple ways to evaluate the effectiveness of treatment at the level of the individual client (Lueger et al., 2001). Due to the COMPASS being developed specifically for a managed care company (Integra), it has not been as widely disseminated as many of the other measures discussed in this review. Despite this, the ETR system developed by Howard and his team, and the use of clinically significant change in an outcome measure are at the heart of nearly every ROM system (Lambert, 2013).

The Outcome Questionnaire. Soon after the COMPASS, Lambert and colleagues developed a measure called the Outcome Questionnaire 45 that built on Howard's work (OQ-45; Lambert et al., 1996). The OQ-45 bases its definition of outcome on a theory elaborated by Lambert (1983) which posits three primary domains that must be included in an outcome measure: symptom distress (intrapersonal distress), interpersonal functioning, and social role performance (Lambert et al., 1996). The OQ-45 was developed for reasons similar to those used in Howard's work, as a means to provide evidence that a therapist's interventions are effective (Lambert et al., 1996), as well as to aid in predicting and preventing treatment failures (Lambert et al., 2003). Like the COMPASS, the OQ-45 incorporates the concepts of clinical significance as well as ETR (Lambert et al., 1996). The OQ-45's ETR only uses two variables to predict expected treatment response (compared to the 7 used in the COMPASS), initial OQ-45 score and

early treatment response. These criteria were selected based on previous analyses which revealed these to be the only variables out of many that accounted for a significant amount of variance in predicting the outcome of treatment (Lambert, Hansen, & Finch, 2001).

Additionally, they created a signal alarm system that provided clinicians with color-coded notices of whether the client was deviating from the ETR, making the results easily interpretable for clinicians (Lambert et al., 2001).

What sets the OQ-45 apart from all other ROMs is the degree to which it has been studied to determine whether it actually improves clinical outcomes (Goodman, McKay, & DePhilippis, 2013). In 2001, Lambert and colleagues conducted a study to determine whether the signal alarm system helped prevent treatment deterioration. Their study was a randomized controlled trial in which clinicians were divided into two groups, with clients assigned to one of two groups: an experimental group that received feedback based on the signal alarm system, and a control group that did not receive any feedback. They found that the feedback group had 10% more cases of reliable change than the control group, and 17% fewer clients who deteriorated by the end of treatment. These encouraging findings led to two replications with similar results, and these three studies were then meta-analyzed (Lambert et al., 2002; Lambert et al., 2003; Whipple et al., 2003). In the meta-analysis, Lambert and colleagues (2003) compared the experimental and control groups from the previous studies in order to determine whether the feedback improved outcomes across studies. The overall sample size was approximately 2500 clients from a college counseling center. The effect size of the comparison between groups was a modest .09. When comparing the rates of improvement of those who were on track to deteriorate, however, a medium effect size of 0.39 was found (21% improved in control, 35% in feedback group). Based on these results, Lambert and colleagues concluded that the feedback

was helpful for those on track to deteriorate, but not necessarily for those who did not deviate from the trajectory of the ETR. They also found that clients whose therapist received feedback stayed in treatment longer if they were on track to deteriorate, and remained in treatment for less time if they were on track to improve, providing evidence of cost-effectiveness of the OQ-45's feedback.

Following this meta-analysis, Lambert's team performed several more replications, but added another experimental group that incorporated their newly developed Clinical Support Tools (CST); additionally, two studies added minor variations such as feedback to patients and delays in reporting (Hawkins et al., 2004; Harmon et al., 2007; Slade, Lambert, Harmon, Smart, & Bailey, 2008; Whipple et al., 2003). The CSTs were designed in order to improve the outcomes of potential deteriorators and non-responders, since even with feedback, a substantial percentage of these clients did not improve. The CSTs assessed for three domains when a client was found to be deteriorating: therapeutic alliance, patient motivation, and social support (Lambert et al., 2007). Once the client has taken this assessment, empirically-based decision trees are created that help the therapist to know how to best adapt treatment in order to improve the client's outcomes. The results of these studies supported previous findings about feedback interventions, and showed that the CSTs significantly improved results in terms of how many clients deteriorated, how many reliably improved, and how many sessions deteriorating clients stayed in treatment (Slade, et al., 2008).

In 2010, Lambert's team undertook another meta-analysis in order to examine both the effects of feedback alone, as well as the impact of the CST's (Shimokawa et al., 2010). Their analysis incorporated 6 studies, all of which had a feedback group, and four of which had a CST group, although two studies did not have a control group. All studies were from Lambert's team

and incorporated scores on the OQ-45 as the criterion of outcome. The study performed the analyses in two separate ways, one in which all clients were included (ITT method), and one in which clients were excluded based on whether they were reasonably able to benefit from the feedback intervention (i.e., attended more than one session of therapy) and called this the efficacy method. They divided the groups into 4 major categories: treatment as usual, therapist feedback, therapist and patient feedback, and CST (which included feedback). Each of these groups had two subcategories, on track and not on track to improve. The results of this meta-analysis showed effect sizes ranging from .28 to .44 for the ITT method and between .53 and .70 for the efficacy method. This provided strong evidence for the utility of the signal alarm system and the CST in preventing treatment failure. In examining the difference between the CST and the regular feedback groups, there was a small difference ($g=.16$), indicating that there was likely not much added benefit to the CSTs. The patient feedback was found to not improve outcomes compared to therapist-only feedback. Compared to the previous meta-analysis, the results of this study found similar effect sizes with the ITT method (which was used in the first meta-analysis), but substantially larger effects with the efficacy method. Given that there was no need for the signal alarm system to provide alerts at intake, it would seem that the efficacy method best captures what the OQ-45's capacity to improve outcome. This provides even more promising evidence for the use of ROM in routine practice, although the authors caution that the efficacy method substantially reduced the sample size due to the more stringent inclusion criteria.

There are, however, some limitations to Lambert's research that should be acknowledged. First, all of the research used to show the efficacy of the OQ-45 feedback and CSTs was performed by Lambert and his team. This opens up the possibility of an allegiance effect, which have been found to be common in both psychotherapy outcome studies (Luborsky et al., 1999)

and measure development studies (Blair, Marcus, & Boccaccini, 2008). Second, the OQ-45 was the sole measure of outcome as well as the basis for the data in the feedback in every study. This leaves open the possibility that the positive results of the RCTs may be influenced by a response shift or a variety of alternate explanations that have yet to be ruled out (Evans, 2012). Third, all but one study (Hawkins et al., 2004) in the meta-analysis were performed at the same university counseling center, with the Hawkins et al. study accounting for a very small portion of the overall sample size (n=306 out of n=6,151 total), potentially indicating a lack of generalizability. Fourth, the factorial validity of the OQ-45 has recently come under criticism, as attempts to replicate it have failed to identify discrete social roles and interpersonal relationship dimensions and have found the correlation between the symptom distress and total score to be as high as .98 (Tarescavage & Ben-Porath, 2014). Despite these limitations, the work of Lambert and colleagues has made a significant contribution in providing a foundation for the evidence base for ROMs, helping them to gain increasing prominence in the field of psychotherapy.

Partners for Change Outcome Monitoring System (PCOMS). Another system that has contributed to the evidence base for ROM is the PCOMS, developed by Scott Miller and Barry Duncan (Miller, Duncan, Sorrell, & Brown, 2005). The PCOMS is based on two 4-item measures, the Outcome Rating Scale (ORS; Miller, Duncan, Brown, Sparks, & Claud, 2003) and the Session Rating Scale (SRS; Duncan et al., 2003). The ORS was designed to be a brief alternative to the OQ-45, and as such, attempts to measure the same domains as the OQ-45 and uses an ETR system modeled after Lambert's signal alarm system (Miller, Duncan, Brown, Sorrell, & Chalk, 2006; Miller et al. 2003). The SRS was developed in order to assess therapeutic alliance based on Bordin's (1979)'s definition, as well as a client's theory of change. Both measures are completed by the client and use visual analogue scales, with the ORS being

administered prior to session, and the SRS being administered after each session. Similar to the other ROMs discussed, PCOMS uses ETR and clinically significant change criteria to aid in the interpretation of results.

Given the brevity of these measures, an essential task was to demonstrate that they possess sufficient reliability and validity in order to be of value. Several studies have examined the psychometric properties of the ORS (Bringhurst, Watson, Miller, & Duncan, 2006; Campbell & Hemsley, 2009; Miller et al., 2003) and the SRS (Campbell & Hemsley, 2009; Duncan et al., 2003). The ORS has been shown to have a Cronbach's alpha coefficients between .85 and .95, to be modestly stable over time (although test-retest reliability went as low as .49) and to have modest to high correlations with existing measures (correlations between ORS and OQ-45 ranged from -.52 to -.74 across three studies). The SRS has demonstrated reliability, but has not demonstrated as strong a correlation with existing measures of therapeutic alliance (Campbell & Hemsley, 2009) as the ORS has with measures of distress. To date, there has been no factor analysis performed on either of the PCOMS measures.

The PCOMS group has also performed several studies to examine whether the PCOMS reporting system improves clinical outcomes (Anker, Duncan, & Sparks, 2009; Miller et al., 2006; Reese, Duncan, Bohanske, Owen, & Minami, 2014). The first study (Miller et al., 2006) was performed on a large sample (N=6424) of telephonic therapy clients at an employee assistance program. This study used a quasi-experimental design in which they compared clients treated using the PCOMS system with an archival baseline. They found that the PCOMS improved outcomes substantially, with an effect size of .79 when comparing all clients and 1.06 with clients who began in the clinical range. These promising findings led to three RCTs that employed a similar methodology to Lambert's OQ-45 RCTs, in which the outcomes of the TAU

group were compared with those who received feedback (Anker, Duncan, & Sparks, 2009; Reese, Norsworthy, & Rowland, 2009; Reese, Toland, Slone, & Norsworthy, 2010). A major strength of the PCOMS studies is that they were performed in a variety of treatment settings and examined the utility of the PCOMS system in both couples therapy (Anker, Duncan, & Sparks, 2009; Reese, Toland, Slone, & Norsworthy, 2010) and individual therapy (Miller Duncan Brown, Sorrell, & Chalk, 2006; Reese, Norsworthy, & Rowland, 2009). These studies found an average effect size of .52 for feedback compared to TAU as well as improved retention rates and higher usage rates by clinicians than other measures (Duncan, 2012). These results differ somewhat from the OQ-45 research, where the feedback intervention was shown to be primarily useful for those who were on track to deteriorate. Recently, a study was completed in which data from a large behavioral health care practice used the PCOMS system for all clients for several years and compared these scores to benchmarks established through RCTs on depression as well as the feedback RCTs of the OQ-45 and PCOMS groups (Reese et al., 2014). They found that the outcomes in their study were sufficiently similar to those of the depression RCTs to suggest that the data was likely representative of the effects of treatment. From there, they examined the OQ-45 and PCOMS feedback RCTs and found that their current study demonstrated similar effect sizes to the OQ-45 studies (OQ: $d=.57$, current study: $d=.71$) but substantially less than previous PCOMS studies ($d=1.13$). The effect sizes for the TAU benchmark were .46, indicating that the current study improved outcomes by $d=.25$ relative to TAU, with the results of all feedback research (OQ-45 & PCOMS) suggesting an increase due to feedback as $d=.14$ (total effect size $d=.60$).

This study revealed a trend that the PCOMS system generally outperformed the OQ-45 in improving client outcomes. There are several possible reasons for this. First, using a 4-item

scale potentially narrows what is being measured and may be less stable over time. This seems to be a possibility given that the test-retest reliability was relatively low and the concurrent validity of the ORS with the OQ-45 revealed only modest correlations in two of the three validation studies. Another possibility is that the sample size for the PCOMS RCTs were much smaller (PCOMS= 408, OQ-45=4,268). This smaller sample size might increase the probability of the results being sampling error, and thus not as reliable. Finally, the inclusion of the SRS to assess the therapeutic alliance may have increased the efficacy of the feedback intervention in a similar way to Lambert's CSTs which showed increased benefit compared to feedback alone (Reese et al., 2014). Although there are still questions to be answered, the PCOMS research has strengthened the evidence base created by Lambert and colleagues, while showing that one can achieve similar results to the OQ-45 in much fewer items (Duncan, 2012).

Using ROMs to evaluate therapist effectiveness. One of the major downfalls of the EST movement was that its top-down approach to implementing best practices was not well received. One alternative paradigm to this has been to examine whether specific therapists, instead of treatment packages, are “empirically supported” (Bohart, 2000; Okiishi, Lambert, Nielsen, & Ogles, 2003). This would provide a more bottom-up approach to EBP, as therapists would be measured based on the results of their routine practice, instead of requiring them to adhere exclusively to the results of RCTs performed in other settings. From this perspective, if the therapist could show that their work was providing substantial benefits for clients, then there would be no need to change what that practitioner does, even if they are not using an EST. If, however, the clinician is not performing well, then interventions could be planned to help improve the effectiveness of the therapist (Okiishi et al., 2006). In order to gather this sort of data, it would require practice-friendly measures that are clinically relevant and could be

administered frequently throughout the course of therapy. These requirements make ROM systems an ideal candidate for the collection of such data, as they were designed to be implemented in routine practice.

One of the first studies using ROM to examine the efficacy of therapists in routine practice was performed by Okiishi and colleagues in 2003. This study examined the degree to which 1,779 clients at a university counseling center changed (as measured by the OQ-45) in response to the intervention of the 56 therapists at the center. What they found was that therapists varied substantially in their outcomes, with the most effective therapist achieving an average improvement in OQ scores of 20.77 points in 3.77 sessions (well over the RCI of 14), and the least effective therapist's clients deteriorating by an average of 5.75 points in 8.46 sessions. The authors also found that the therapists' theoretical orientation, type of training, amount of training, and gender did not significantly predict differential client growth trajectories. These results provide evidence that therapists are performing at differential levels of efficacy, and that these differences are not explained by the factors above. Thus, it would seem to be more important to examine the reasons for differential efficacy of therapists, than to continue doing studies attempting to empirically validate treatment packages. These findings also highlight the importance of tracking outcomes in routine practice, as many clients may be getting worse when they see certain therapists. The authors recommend several interventions for improving the outcomes of the treatment center: (a) Show the results of the study to each therapist and potentially explore the reasons for the results, (b) Refer clients to the most effective therapists until their case loads are full and proceed in order of effectiveness and assign other duties and additional training to less effective therapists.

The same group sought to replicate their findings with a larger sample size, in order to ensure the reliability of their results (Okiishi et al., 2006). Their follow-up examined the results of 71 therapists who treated 6,499 clients at the same university counseling center across a period of six years. In this study, more moderate differences were found between therapists, with the top 10% of therapists seeing clients for 7.91 sessions, on average, leading to an improvement of 13.46 OQ-45 points, with the lowest ranking therapist seeing clients for more sessions (10.41) leading to only an improvement of 5 points. This study also examined the rate at which clients reliably improved, recovered (reliably improved and were below the clinical cutoff), did not change, or deteriorated. Clients who saw a top 10% therapist had an average recovery rate of 22.40% (21.54% improved) and a deterioration rate of 5.20%. On the other hand, therapists in the bottom 10% had a recovery rate of 10.61% (17.37% improved) and double the deterioration rate of the top 10% (10.56%). In addition to differences in effectiveness, therapists were also found to vary in efficiency (although the most effective therapists were not always the most efficient) with the most efficient therapists (top 10%) achieving an average improvement of 1.59 OQ-45 points per session, and the least efficient therapists producing .4 points of improvement per session. These results suggest substantial differences in therapist effectiveness and efficiency, providing further evidence of the importance of routinely assessing the outcomes of therapists and encouraging accountability in practice. In addition to the recommendations made in the previous article, Okiishi and colleagues recommended making outcome information available to clients to allow them to choose therapists. They discuss how the way to improve treatment would be to identify what it is that makes effective therapists so effective, and build this into training.

Another approach was taken by the developers of the Treatment Outcome Package (TOP). This group was concerned about the lack of complexity involved in evaluating the quality of therapist outcomes. Their concern came in the context of budding governmental policies in the USA in the early 1990's that began to put pressure on therapists to provide evidence that, on average, those who seek their services receive measurable benefit (Kraus & Castonguay, 2010). The TOP developers' primary concern was that policy-makers would attempt to judge the outcomes of therapists based on narrow definitions that did not incorporate case mix variables into their decisions. Thus, the developers created a measure that broadened the scope of the outcomes measured, as well as incorporated complex case-mix algorithms that could portray a more nuanced picture of clinical outcomes (Kraus, Seligman, & Jordan, 2005). In their initial validation study (Kraus et al., 2005), Kraus and colleagues factor analyzed a large data set from 383 different treatment sites and found that the TOP measured 11 factors. This stands in sharp contrast to other ROM systems which often assess 3 to 5 areas, but generally report only one total score in their studies of therapist efficacy, thus leading to unidimensional assessments of therapist outcomes (Castonguay, Barkham, Lutz, & McAlveavy, 2013).

Using this multidimensional approach, Kraus and colleagues (2011), sought to examine empirically the question of whether therapists demonstrate differential performance across the 11 dimensions (Kraus, Costingway, Boswell, Nordberg, & Hayes, 2011). This study included the TOP data for 6960 patients who were seen by 696 therapists (10 patients per therapist). The authors defined effectiveness as having the average client experience reliable change in a specific dimension. Thus, if a therapists' mean client change was in the positive direction they were labeled "effective" in that dimension, if the average change was negative, then they were labeled "harmful" and if there was no change, they were labeled "unclassifiable." Using these

criteria, the study found that therapists, on average, were effective in 5 TOP domains, with 96% of therapists identified as effective in at least one domain and no therapist was competent in all domains (although the “mania” scale showed very little change). Therapists were then ranked within each dimension in terms of their effectiveness, and it was found that the correlations between rankings in any given domain did not exceed .33. This finding provides support for the idea that competence in one area does not necessarily translate to other areas of functioning. The authors discuss how these findings have many implications such as potentially matching clients with providers who have a demonstrated track record at working effectively with the clients presenting concerns, as well as having therapists use the results to identify areas of strength and weakness that can be targeting for further training. These findings also offer a critical view toward the studies performed by Okiishi and colleagues using the OQ-45. Given the finding that most therapists were effective in at least one domain suggests that a binary categorization of therapists into ‘effective’ and ‘ineffective’ groups based on a single criterion may be misleading.

A substantial limitation of this study, however, is the sample size. Given that there were 11 TOP domains and only 10 clients per therapist, it would not be unlikely that therapists might not see any clients that have problems in specific TOP dimensions. Thus, their results in any given dimension could be based on one client that had concerns in that area. On the other hand, the large size of their overall sample, and the substantial number of therapists involved in the study could account for this problem.

Overall, the research investigating differential therapist efficacy provides strong support for the value of ROM in improving outcomes at the individual client level, as well as facilitating research that can help therapists and treatment sites to gain a clearer sense of their strengths and

weaknesses. It also shows how ROM could be used to examine questions of why therapists may be more effective, which could drastically change therapist training protocols.

Lack of Usage of ROM

Despite the strong body of research supporting the usage of ROM, it has not achieved widespread implementation in routine clinical practice. In order to examine usage rates of outcome measures in routine practice, Phelps and colleagues (1998) gave a 9-item survey to a large sample (n=15,918) of American Psychological Association (APA) members who were identified as direct psychotherapy providers. The authors reported that 24% of Independent practitioners and 40% of medical facilities used outcome measures in their practice, suggesting substantial variability across the types of practices. The authors examined a subset (n=1600) of the sample's answers to a free-response item about whether they use outcome measures in their practice, and if so, the names of the measures. They found that approximately 22% of clinicians used some sort of standardized outcome measure, with the most used measure being the Beck Depression Inventory (n=95, 6% of total). Others reported using unstandardized patient report (8%) or informal clinician report (5%), while 66% (n=1,049) of the sample reported using no outcome measure. Their study suggested that only a minority of practitioners were using standardized assessments in their clinics. Most ROM systems, however, were still relatively new or not yet developed at the time this article was published, but it does provide some context into the lack of value given to standardized assessments by clinicians, especially those not in a medical setting.

In 2004, Hatfield and Ogles performed a survey using similar sampling procedures as Phelps, Eisman and Kohout (1998) in order to examine the usage rates of standardized assessments. Their sample consisted of 874 practitioners from a variety of settings, with

substantial representation from solo private practitioners (46%). They found that 37% of respondents reported using an outcome measure in their practice, although of this group (n=324), 12% endorsed not using a standardized outcome measure. Thus, 32% of the clinicians in their study used a standardized outcome measure in their practice, with the Beck Depression inventory being the most used standardized measure (n=146, 45.3% of standardized measure users). The only ROM system with enough users to be reported in the study was Lambert's OQ-45 with 18 clinicians (2.1% of the total sample) using it in their practice. It is possible that there were some that fit in the "Other" category, but these would represent very small percentages. They also found that cognitive or behaviorally oriented therapists were more likely to use outcome measures than insight-oriented therapists. Thus, although Hatfield and Ogles' study found that approximately one-third of clinicians use some sort of outcome assessment, about 98% of clinicians reported not using ROM systems.

Clinician attitudes towards ROM. Given the lack of usage of ROM systems, there has been an interest in understanding why these measures are not being used. With this in mind, several studies surveyed or interviewed clinicians in order to gain greater insight into their reasons for not using ROM systems. In 1999, Abrahamson examined the implementation process of an outcome measure in routine practice. He reported three major reasons which deterred clinicians from using ROM systems: (a) concerns about performance being measured, (b) logistical considerations (i.e., time burden), and (c) conceptual appropriateness of the measure. The author discussed how logistical details must be worked out meticulously, and that those who are expected to use the measures should have substantial input into what is assessed. These themes are present throughout the literature from the time of Abrahamson's study to more recent studies and will be explored in greater depth as the literature is discussed.

Garland, Kruse, and Aarons (2003) examined clinician attitudes in a large children's public mental health service system in California. At this time, the state government had mandated the use of an outcome assessment battery that was to be routinely administered, with results being sent to the state government. The authors sought to examine clinician attitudes toward the use of the measures with individual clients. The sample consisted of 50 clinicians with varying years of experience, with the majority being master's level therapists (62%). In order to examine clinician attitudes, the authors developed a semi-structured interview that was administered to clinicians either in a focus group or individually. These interviews were recorded and then coded to identify major themes. They found that 25% of clinicians expressed strong ideological opposition to the idea of a quantitative outcome measure, while another 25% were skeptical about the ability of these measures to assess therapeutic change. "A few respondents" (p.398) endorsed strong support for the use of the outcome measures, while many expressed ambivalence. Additionally, they found that, on average, clinicians rated standardized measures and scales as the least important method of evaluating effectiveness, with functional indicators being rated the highest (school grades, disciplinary actions etc.). Most clinicians reported that they did not feel increased pressure to demonstrate their effectiveness with the use of outcome measures, which offers a contrast to Abrahamson's (1999) fear of evaluation concern. Most attributed this to the perception that the data was not even being used, so the perceived unlikelihood of evaluation may explain why the clinicians in this sample did not experience this fear.

The authors categorized clinician barriers into 3 categories: (a) feasibility concerns, (b) perceived invalidity, and (c) interpretation difficulties. The first two categories are also represented in Abrahamson's (1999) study, but the third was unique to this study. Almost all

respondents (90%) felt that there was a substantial time burden associated with the measures, with most expressing frustration at having to do administer and take the measures. Some also reported having received negative feedback from parents about filling out the measures, with few providing positive feedback. About 55% believed that the measures were not relevant to their clients, or that it did not fit well with their theoretical orientation. Participants stated that the reports were difficult to use and did not provide much useful information. Several clinicians reported that they would have liked to have been involved in the selection of the constructs being measured.

In regards to their use of the scores/reports, 92% indicated that they had never used them in their clinical practice, which was tempered somewhat by over half (60%) of the respondents making comments about how the process of administering the measures was helpful in and of itself. The clinicians talked about how it was helpful to be able to see what parents thought about their children's problem and often gave openings for discussing topics with parents that would not have been explored otherwise. These findings are in line with other studies that have considered the role of ROM feedback in enhancing the communication between therapist and client (Halstead, Yuon, & Armijo, 2013; Reese, Slone, & Miserocchi, 2013).

This study highlights the varying opinions of clinicians regarding ROM, and that the current offerings seem geared toward clinicians with certain philosophies of therapy. Despite these somewhat disappointing findings, the authors found that the overwhelming majority of clinicians in their sample were interested in assessing outcome in some way, even though they found the current measures lacking.

In 2006, Meehan, McCombes, Hatzipetrou, and Catchpoole sought to explore clinician reactions to the introduction of ROM and the utility of outcomes data in their practice. Their

study utilized a focus group methodology, in which 324 practitioners from a variety of settings were interviewed in 34 group discussions. Similar to previous studies, the authors found that the clinicians had mixed opinions about ROM. They reported that “even those in favour of outcome measurement questioned the validity of the measures and felt the selected measures were too brief and broad to be useful” (p. 583).

The participants discussed the difficulties of balancing relevance with practicality, in that if the measures were more comprehensive, they would become too long to be feasible for routine practice. This highlights a common difficulty with ROM systems: how to create a measure that is sufficiently brief to be practical for routine use, while still providing sufficient breadth to be of clinical value (Slade, 2002). Additionally, some participants reported feeling that the measures were overly reductionistic, and offered little insight into the worldview of the client. The authors’ interpretation of these concerns was that they “highlighted a lack of understanding of the measures” (p. 583). When the authors asked which measures would be more appropriate, the clinicians were unable to reach a consensus, as each suggested measure was critiqued. From this, the authors concluded that consensus was very difficult to achieve, and that the clinicians likely could not be pleased. The authors also found that participants reported some trepidation about how the outcome data would be interpreted. Given that this study was performed in Australia, where ROM was required by the government, the data from the outcome measures could be used to evaluate the efficacy of their treatment, despite them feeling that it does not capture the scope of their work. Finally, concerns about time burden were present. There were measures that the therapists needed to complete, or client measures that they needed to enter into the system, and these often required the clinicians to wait in line for a computer to enter the

scores. This study confirms all three of Abrahamson's clinician concerns but provided a more "thick" description of what types of concerns clinicians had about ROM systems.

In a follow-up to their 2004 article, Hatfield and Ogles (2007) performed a study in which they sought to examine reasons why therapists use or do not use outcome measures. The same database as their 2004 study was used, but different analyses were performed. In terms of why clinicians do not use outcome measures, their primary finding was that practical concerns such as time burden and lack of resources was the strongest reason. Additionally, they found that the second most common reason was perceived lack of utility, in that practitioners did not feel it was helpful or potentially distorted the effects of treatment. These reasons are closely aligned with those documented in the studies examined thus far, and suggest that practicality and relevance or clinical utility are the two most prominent reasons that clinicians do not use outcome measures.

Support for clinician attitudes. In the ROM literature, few would disagree that practical/logistical difficulties are a legitimate barrier to implementation, and that work needs to be done to improve this area (Boswell et al., 2015). The concern about relevance, however, has been more controversial. As documented in the previous section, many clinicians report perceiving the measures as not relevant to their work, but ROM developers have expressed disagreement to these claims. (Boswell et al., 2015; Meehan et al., 2006). The next sections will explore this issue from both perspectives.

Some researchers have attempted to examine whether the clinician perception that the measures are not relevant is a valid concern. Ashworth & colleagues (2007) sought to answer the question of whether the CORE-OM, a prominent ROM system, was addressing the majority of concerns that clients brought in to treatment. This was accomplished by comparing the items

generated by clients on an idiographic measure, through thematic analysis, and examining whether these areas were assessed by the CORE-OM. The idiographic instrument used, was the Psychological Outcome Profiles (PSYCHLOPS), which was developed by Ashworth's research team (Ashworth et al., 2004). The sample was composed of 215 clients currently in psychotherapeutic treatment. The authors examined all client-generated items and grouped them into 8 main themes and 61 subthemes. In examining the sub-themes they found that 27 (44%) were not present in the CORE-OM. Additionally, 60% of clients reported at least one presenting concern that did not map onto any CORE-OM item. This suggests that the CORE-OM did not assess at least one major concern for the majority of clients in the sample, and did not account for nearly half of the types of concerns that clients brought to therapy. These results provide support for clinician claims that ROM systems may not be valid to the populations that they serve.

In addition to Ashworth and colleagues' study (2007), some authors have presented theoretical arguments about how ROM systems, as currently constituted, lack relevance. Chris Evans, one of the developers of the CORE-OM, wrote an article in which he advocated exercising caution in the use of ROM systems (Evans, 2012). Evans discusses how ROM systems were generally built using a nomothetic approach to measurement, in which the purpose was to assess constructs at a broad level and to make the measures equally applicable across multiple populations. Additionally, the measures were designed to be administered on a session-by-session basis, thus requiring them to be much more brief than traditional cross-sectional assessments. In doing so, he states that "we cannot escape the fact that we have traded out exploration of certain areas of positive or negative experience, particularly ones relating to interpersonal functioning and risk, in order to get coverage across as many clients as possible"

(p. 133). He describes how the measures generally perform well in measuring areas of intrapersonal concerns (i.e., depression and general well-being), but that they are less able to assess interpersonal concerns in a way that is applicable to most people, because people are involved in different types of relationships (intimate, occupational, nuclear family, etc.). This sounds quite similar to many of the concerns that clinicians brought up about the scope of measurement, and raises the question of whether this trade-off of global applicability for specificity is worth the cost.

His views are corroborated by the relatively weak psychometric properties of scales assessing interpersonal concerns in ROM systems. The OQ-45's interpersonal relations scale was reported to have a Cronbach's alpha of .74, and the TOP's Social Conflict scale was reported to have an alpha of .72. These reliabilities are much lower than what was found for the OQ-45's symptom distress scale (.92) and the TOP'S Depression scale (.93). Additionally, although some moderate and large correlations with relationship-oriented criterion measures were found, in most cases the same criterion measures correlated more highly with intrapersonal scales, suggesting a lack of discriminant validity. These properties are not wholly unacceptable, but are much weaker than those found in the intrapersonal scales of these same measures. Similar patterns were found with scales of occupational functioning in these measures, except that no large correlations were found with a criterion measure (Kraus et al., 2005; Lambert et al., 1996). With these considerations, Evans concludes that the philosophy that guides the development of ROM measures is geared toward "producing useful lowest common denominators that will give precision at low cost... it is not about producing high quality, high precision, adaptable measurement of individual change" (p.133).

Given this approach, Evans worries that a focus on outcomes generated by ROM systems, could lead to therapists to focus on attempting to improve the areas that are assessed by the measures, at the expense of more subtle or difficult topics. Thus, although ROM developers generally advise that the results of the measure be contextualized to the individual circumstances of the clients, it is possible that therapists may try to adapt what they see in treatment to fit the results of the measure. He expresses especial concern about this issue with trainees, in that if the client is within expected treatment response range, the student therapist may not be as attentive to the possible need to address more sensitive topics. Instead, the measure could serve an analgesic function, alleviating the anxiety of the trainee because the score on the measure is improving, while the client may be struggling in areas not assessed by the measure. Overall, Evans's treatise provides a more elaborate explanation of some of the possible therapist concerns regarding the relevance of the measures. Combined with Ashworth and colleagues' work, it shows that at the very least, clinician concerns should be acknowledged, and critical dialogue should be had about the philosophy underlying the development of these systems as well as the areas they assess.

Researcher conceptualizations of lack of usage. In early 2015, as part of a special issue of *Psychotherapy Research* about practice-based evidence, the authors of three of the major ROM systems in the United States came together to discuss what they perceived as the major barriers to ROM implementation (Boswell, Kraus, Miller, & Lambert, 2015). In line with previous research, they identified two major obstacles: practical and philosophical. In terms of practical obstacles, financial burden, time burden, and multiple stakeholders with different needs were identified. They discuss several areas where clinician and client time can be used, and emphasized the importance of creating software that can streamline the process of administering

and scoring the assessments. Additionally, they cite difficulties with making a system that is appealing to the various stakeholders involved in the therapeutic process, and how if stakeholders do not feel their needs are met, the initiative will likely fail.

As the authors begin to discuss “philosophical obstacles,” the tone of the article shifts. In discussing concerns of a practical/logistical nature, the authors emphasized making sure that the measure fits well with the context of local treatment sites, and were careful not to dismiss clinician concerns. In contrast, the “philosophical obstacles” section leads off by discussing how clinician’s concerns about relevance are the result of irrational expectations that the ROM systems should be “perfectly reliable, valid, appropriate and sufficient (i.e., relevant) for each individual client” (p. 6). The almost defensive tone in this section, in which the authors exaggerate the clinician concerns about relevance by suggesting that they want the measures to be perfect, immediately dismisses the concerns about relevance as being the result of erroneous perceptions. The authors then proceed to attempt to educate the clinicians on how the measures should be used, and discuss that these limitations to relevance are inherent to all measures. This way of addressing concerns about the relevance of ROM systems exonerates the properties of the measures from responsibility for the lack of implementation, and places this responsibility on the biases of the clinicians.

Later in the article, the authors briefly mention how they are consistently improving the psychometric properties and predictive algorithms of the measures, and that this addresses the issue of relevance. The authors, however, provide no specific information about how this would be done, or how it would address concerns about relevance. The assumption is that if the psychometric properties and prediction algorithms are improved, then clinicians will find it useful and relevant to their clients. It does not address, however, the concern that the measures

are missing important outcomes for specific clients, or that their scope is too global to be of value to clinicians (Evans, 2012). Up to this point, no ROM systems have had their item pool changed after their initial validation studies, thus it is unlikely that this is what is meant by improving psychometric properties.

Additionally, the authors stated that “therapists’ confidence in their clinical judgment alone stands as a barrier to implementation of monitoring and feedback systems” (p. 8). The question that arises from this statement is why clinicians would rely only on clinical judgment when tools such as ROM systems are available. Walfish, McAlister, O’Donnel, and Lambert (2012) attribute this to therapist self-assessment bias. In this study, the authors examined whether therapists have an overly positive bias toward their abilities. They found that clinicians generally rated themselves as above average (mean rating: 80th percentile), which the authors viewed as a statistical impossibility. Additionally, nearly two thirds of the sample believed that 80% or more of their clients improved by receiving therapy from them, and that only 3.66% of their clients deteriorate. The authors contrasted this with research that suggested much lower rates of improvement and higher rates of deterioration for the average therapist, such as the work by Lambert’s group’s studies about therapist effects (Okiishi, Lambert, Nielsen, & Ogles, 2003; Okiishi et al., 2006). Based on this data, the authors conclude that therapists view their capabilities as greater than they are, and that this suggests a biased self-assessment. The authors then state that “therapist self-assessment bias may be at the root of therapists’ reluctance to take advantage of advances in ‘lab test’ results [ROM]” (Walfish et al., 2012; p.5). Therefore, the reason that clinicians trust their clinical judgment alone is because they have a cognitive bias that leads them to think that they are better than they are, and that they do not need ROM. These results lead to the conclusion that the primary reason for a lack of implementation of ROM is not

a lack of relevance or practicality of the measures, but due to erroneous clinician attitudes that are detrimental to clients.

The rationale just discussed is one of the predominant researcher narratives about the lack of ROM usage, however others exist. Some claim that “most of the ‘restraining factors’ emanate from underlying performance anxiety on the part of clinicians” (Mellor-Clark, Cross, Macdonald, & Skjulsvik, 2014, p.4) or from a lack of understanding of the scope of ROM measures or of measurement theory (Boswell et al., 2015; Meehan et al., 2006). What all of these attributions have in common is that they view the lack of success in implementing ROM systems as being due to problematic characteristics of clinicians, and not the shortcomings of the measure. Thus, clinicians’ concerns about relevance are not really about the validity of the measure to their local context, but about clinician’s attempts to maintain a positive view of their work or to avoid experiencing anxiety related to their performance. From the perspective of ROM proponents, the clear implication of this is that actions need to be performed in order to change clinician attitudes. Additionally, they need to help correct erroneous clinician perceptions through training and education, in order to convince clinicians that ROM aligns with their personal and professional goals (Boswell et al., 2015; de Jong, 2014; Lakeman, 2004; Willis, Dean, & Coombs, 2009).

The fact that researchers attempt to frame the lack of clinician acceptance of ROM as involving some sort of problem with the clinicians’ attitudes is not surprising given a growing body of literature in the management literature that questions traditional conceptions of resistance. Ford, Ford, and D’Amelio (2008) discuss how in accounts of attempts to implement change, that “resistance is portrayed as an unwarranted and detrimental response residing completely ‘over there, in them’ (the change recipients)” (p.362). The authors describe how this

narrative is often self-serving for the change agents, in that it provides an account for the lack of implementation of the change, without implicating the change agent or their product.

Additionally it “shift[s] responsibility for resistance from things under their control (i.e., systemic factors) to the characteristics and attributes of recipients” (p.365). This approach is evident in ROM proponents’ views that clinician concerns about relevance are not to be taken as feedback that needs to be accommodated by the measure, but an attempt to wrongly discredit the measures in order to preserve their self-image. The vast majority of researcher accounts about the lack of usage of ROM suggest, either explicitly or implicitly, dismiss the possibility that clinicians may not use the measures because the measures need to be changed or adapted because they are not relevant or practical for clinical needs (Boswell et al., 2015; Mellor-Clark et al., 2014). Until these views are changed, it is likely that implementation approaches will continue to be focused on correcting clinician cognitive biases, while clinicians focus on the shortcomings of the measures. This situation is unlikely to lead to productive dialogue, as it is possible that both view the other as threatening or at least misguided, and are unwilling to compromise and negotiate.

This leads to a crucial issue, are the clinician concerns about relevance only based on erroneous presuppositions or are they in fact legitimate concerns, or are they some combination of the two? As has been discussed above, both parties appear to have evidence for their position, and will use this against each other, so the answer is not clear cut. If ROM developers simply change clinician attitudes, will the field be losing something of value that could potentially improve the way ROM is done (Evans, 2012; Ford et al., 2008)? On the other hand, if ROM systems as currently constituted are rejected, and ROM developers attitudes changed, would

something of value be lost as well? The answer to these questions could have major implications for how ROMs are developed and implemented.

If the clinicians who claim that the measures are not relevant to their context are correct, then it should follow that ROM systems should be built that focus on adapting the measure itself to the needs of specific treatment sites, therapists, and clients. These types of accommodations, however, are not possible in current ROM systems, because they were built on the assumption that brief, standardized, and broadly applicable assessments are preferable to more individualized measures. It would be neither feasible nor desirable to have the measures adapt to each treatment site, therapist, and client. On the other hand, if the clinicians' skepticism is in fact just the result of self-assessment bias, a lack of understanding of the measures, or based on performance anxiety, then these clinician perspectives are potentially hurting their clients and efforts should be undertaken to help correct practitioner biases. In this case, there would be no need to change existing measures nor the paradigm on which the measures rest and implementation should continue its current course. Unfortunately, these competing views lead to a deadlock between ROM developers and clinicians, which is likely to continue to lead to researchers being frustrated at clinician cognitive biases, and clinicians feeling that researchers are forcing an irrelevant bureaucratic task on them, that could potentially harm their clients.

A New Implementation Approach

Thomas and Hardy (2011) advocate for a different approach to implementation. In this article they discuss how the approach of "demonizing" and thus attempting to overcome 'change recipient' resistance to change has not proven effective (Beer & Nohria, 2000). They argue that "while change can be imposed, it is more likely to be taken on by members of the organization if they have played a part in the negotiations of new meanings, practices, and relationships"

(p.323). In this framework, traditional distinctions of change agents and change recipients are broken down, with both sides at times playing the part of agent and recipient of change. This facilitates a process in which negotiations can occur that will allow for emerging ideas that develop between the participants, and that are acceptable to both. With this conceptualization of change, implementing ROM systems changes from a question of “who resists organizational change, why and when, to a question of how relations of power *and* resistance operate together in producing change and in what ways” (p.326). From this perspective, clinicians’ reluctance to embrace ROM could be seen as “resistance,” but so could ROM developers’ dismissive attitude toward clinician concerns about relevance. Thus, the implementation process does not assume a correct answer from the outset (i.e., Clinicians need to implement ROMs, or ROMs are completely irrelevant to therapy), but focuses on dialogue between the two parties, in order to create shared meaning. It is hoped that this perspective will allow the merits of the current approach to ROM be available for discussion, and for the voices of key stakeholders to be valued.

In order for such an approach to be realized in ROM implementation, however, there must be room for negotiation from both researchers and clinicians. Unfortunately, existing ROM systems are not designed to accommodate local input, as items are chosen primarily on the basis of psychometric considerations and the opinions of clinicians involved in the initial development of the measures. This is required in order to ensure comparability between clients, to create norms (i.e., clinical cutoffs), as well as generate ETR curves. In order for more productive relationships to develop between ROM developers and clinicians, a new paradigm of ROM development is required that allows for negotiations of the meaning of outcome while still maintaining some degree of standardization. To this point, no such system exists.

References

- Abrahamson, D. J. (1999). Outcomes, guidelines, and manuals: On leading horses to water. *Clinical Psychology-Science and Practice, 6*, 467-471.
- American Psychological Association Presidential Task Force on Evidence-Based Practice. (2005). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.
- Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77*, 693-704.
- Ashworth, M., Robinson, S., Evans, C., Shepherd, M., Connolly, A., & Rowlands, G. (2007). What does an idiographic measure (PSYCHLOPS) tell us about the spectrum of psychological issues and scores on a nomothetic measure (CORE-OND)? *Primary Care and Community Psychiatry, 12*, 7-16.
- Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H., . . . Godfrey, E. (2004). A client-generated psychometric instrument: The development of 'PSYCHLOPS'. *Counselling & Psychotherapy Research, 4*, 27-31.
- Beer, M., & Nohria, N. (2000). Cracking the code of change. *Harvard Business Review*, (May-June), 133-141.
- Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology, 15*, 346-360.
- Bohart, A. C. (2000). Paradigm clash: Empirically supported treatments versus empirically supported psychotherapy practice. *Psychotherapy Research, 10*, 488-493.

- Bohart, A., O'Hara, M., & Leitner, L. (1998). Empirically violated treatments: Disenfranchisement of humanistic and other psychotherapies. *Psychotherapy Research, 8*, 141-157.
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research, and Practice, 16*, 252-260.
- Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy research, 25*, 6-19.
- Bower, P., & Gilbody, S. (2010). The current view of evidence and evidence-based practice. In M. Barkham, G. E. Hardy & J. Mellor-Clark (Eds.), *Developing and delivering practice-based evidence: A guide for the psychological therapies* (1st ed., pp. 1-20). Hoboken, NJ: John Wiley & Sons.
- Bringinghurst, D. L., Watson, C. W., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the Outcome Rating scale: A replication study of a brief clinical measure. *Journal of Brief Therapy, 5*, 23-30.
- Campbell, A., & Hemsley, S. (2009). Outcome rating scale and session rating scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist, 13*, 1-9.
- Castonguay L., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-Oriented Research: Approaches and Applications. In M.J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change (6th Edition; pp. 85-133)*. Hoboken, NJ: John Wiley & Sons.

- Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., ... & Woody, S. R. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist, 51*, 3-16.
- de Jong, K. (2014). Deriving Implementation Strategies for Outcome Monitoring Feedback from Theory, Research and Practice. *Administration and Policy in Mental Health and Mental Health Services Research, 43*, 292-296.
- Duncan, B. L. (2012). The partners for change outcome management system (PCOMS): The heart and soul of change project. *Canadian Psychology, 53*, 93-104.
- Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The session rating scale: Preliminary psychometric properties of a working alliance measure. *Journal of Brief Therapy, 3*, 3-12.
- Elliot, R. (1998). Editor's introduction: A guide to the empirically supported treatments controversy. *Psychotherapy Research, 8*, 115-125.
- Evans, C. (2012). Cautionary notes on power steering for psychotherapy. *Canadian Psychology/Psychologie Canadienne, 53*, 131-139.
- Ford, J. D., Ford, L. W., & D'Amelio, A. (2008). Resistance to change: The rest of the story. *The Academy of Management Review, 33*, 362-377.
- Garland, A. F., Kruse, M., & Aarons, G. A. (2003). Clinicians and outcome measurement: What's the use? *Journal of Behavioral Health Services & Research, 30*, 393-405.
- Goodman, J. D., McKay, J. R., & DePhilippis, D. (2013). Progress monitoring in mental health and addiction treatment: A means of improving care. *Professional Psychology: Research and Practice, 44*, 231-246.

- Halstead, J., Yuon, J. S., & Armijo, I. (2013). Scientific and clinical considerations in progress monitoring: When is a brief measure too brief? *Canadian Psychology, 54*, 83-85.
- Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research, 17*, 379-392.
- Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*, 485-491.
- Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research, 34*, 283-291.
- Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K., & Tuttle, K. C. (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research, 14*, 308-327.
- Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159-164.
- Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy: Causal mediation of outcome. *Journal of Consulting and Clinical Psychology, 61*, 678-685.
- Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*, 1059-1064.

- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.
- Kraus, D., & Castonguay, L. G. (2010). Treatment Outcome Package (TOP): Development and use in naturalistic settings. In Barkham, Hardy, & Mellor-Clark (Eds), *Developing and delivering practice-based evidence: A guide for the psychological therapies* (pp.151-174). Hoboken, NJ: John Wiley & Sons.
- Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21*, 267-276.
- Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61*, 285-314.
- Lakeman, R. (2004). Standardized routine outcome measurement: Pot holes in the road to recovery. *International Journal of Mental Health Nursing, 13*, 210-215.
- Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In M. J. Lambert, E. R. Christensen and S. S. DeJuUo (Eds), *The Assessment of Psychotherapy Outcome*. New York, NY: John Wiley.
- Lambert, M. J. (2013). Outcome in psychotherapy: The past and important advances. *Psychotherapy, 50*, 42-51.
- Lambert, M. J., Bailey, R., Kimball, K., Shimokawa, K., Harmon, S. C., & Slade, K. (2007). *Clinical Support Tool Manual-Brief Version-40*. Salt Lake City, UT: OQ Measures.

- Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the outcome questionnaire. *Clinical Psychology & Psychotherapy*, *3*, 249-258.
- Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*, 159-172.
- Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology*, *10*, 288-301.
- Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., & Nielsen, S. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology & Psychotherapy*, *9*, 91-103.
- Leon, S. C., Kopta, S. M., Howard, K. I., & Lutz, W. (1999). Predicting patients responses to psychotherapy: Are some more predictable than others? *Journal of Consulting and Clinical Psychology*, *67*, 698-704.
- Luborsky, L., Diguier, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., ... & Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, *6*, 95-106.
- Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology*, *69*, 150-158.

- Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient in outpatient psychotherapy. *Journal of Consulting and Clinical Psychology, 67*, 571-577.
- Meehan, T., McCombes, S., Hatzipetrou, L., & Catchpoole, R. (2006). Introduction of routine outcome measures: Staff reactions and issues for consideration. *Journal of Psychiatric and Mental Health Nursing, 13*(5), 581-587.
- Mellor-Clark, J., Cross, S., Macdonald, J., & Skjulsvik, T. (2014). Leading horses to water: Lessons from a decade of helping psychological therapy services use routine outcome measurement to improve practice. *Administration and Policy in Mental Health and Mental Health Services Research, 1-7*.
- Miller, S. D., Duncan, B. L., Brown, J., Sorrell, R., & Chalk, M. B. (2006). Using formal client feedback to improve retention and outcome: Making ongoing, real time assessment feasible. *Journal of Brief Therapy, 5*, 5-22.
- Miller, S. D., Duncan, B. L., Brown, J., Sparks, J. A., & Claud, D. A. (2003). The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy, 2*, 91-100.
- Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The Partners for Change Outcome Management System. *Journal of Clinical Psychology, 61*, 199-208.
- Okiishi, J. C., Lambert, M. J., Eggett, D., Nielsen, L., Dayton, D. D., & Vermeersch, D. A. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology, 62*, 1157-72.

- Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy, 10*, 361-373.
- Phelps, R., Eisman, E. J., & Kohout, J. (1998). Psychological practice and managed care: Results of the CAPP practitioner survey. *Professional Psychology: Research and Practice, 29*, 31-36.
- Reese, R. J., Duncan, B. L., Bohanske, R. T., Owen, J. J., & Minami, T. (2014). Benchmarking outcomes in a public behavioral health setting: Feedback as a quality improvement strategy. *Journal of Consulting and Clinical Psychology, 82*, 731-742.
- Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, Research, Practice, Training, 46*, 418-431.
- Reese, R. J., Slone, N. C., & Miserocchi, K. M. (2013). Using client feedback in psychotherapy from an interpersonal process perspective. *Psychotherapy, 50*, 288-291.
- Reese, R. J., Toland, M. D., Slone, N. C., & Norsworthy, L. A. (2010). Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy: Theory, Research, Practice, Training, 47*, 616.
- Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal, 312*, 71.
- Sales, C. M. D., & Alves, P. C. G. (2012). Individualized patient-progress systems: Why we need to move towards a personalized evaluation of psychological treatments. *Canadian Psychology/Psychologie Canadienne, 53*, 115-121

- Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*, 298-311.
- Slade, M. (2002). What outcomes to measure in routine mental health services, and how to assess them: A systematic review. *Australian & New Zealand Journal of Psychiatry, 36*, 743-753.
- Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy, 15*, 287-303.
- Smith, K. R. (2009). Psychotherapy as applied science or moral praxis: The limitations of empirically supported treatment. *Journal of Theoretical and Philosophical Psychology, 29*, 34.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752.
- Stuart, R. B., & Lilienfeld, S. O. (2007). The evidence missing from evidence-based practice. *American Psychologist, 62*, 615–616.
- Tarescavage, A. M., & Ben-Porath, Y. S. (2014). Psychotherapeutic outcomes measures: A critical review for practitioners. *Journal of clinical psychology, 70*, 808-830.
- Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist, 48*, 3-23.

- Thomas, R., & Hardy, C. (2011). Reframing resistance to organizational change. *Scandinavian Journal of Management*, 27, 322-331.
- van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, Netherlands: Kluwer Academic.
- Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports*, 110, 639-644.
- Walker, B. B., & London, S. (2007). Novel tools and resources for evidence-based practice in psychology. *Journal of Clinical Psychology*, 63, 633-642.
- Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings*. New York, NY: Routledge.
- Wampold, B. E., Goodheart, C., & Levant, R. (2007). Clarification and elaboration on evidence-based practice in psychology. *American Psychologist*, 62, 616-618.
- Wendt, D. C., Jr., & Slife, B. D. (2007). Is evidence-based practice diverse enough? Philosophy of science considerations. *American Psychologist*, 62, 613-614.
- Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, 130, 631.
- Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology*, 50, 59-68.

Willis, A., Deane, F. P., & Coombs, T. (2009). Improving clinicians' attitudes toward providing feedback on routine outcome assessments. *International Journal of Mental Health Nursing, 18*, 211-215.

APPENDIX B: Consent Forms

Consent to be a Research Subject (Client)

Introduction

This research study is being conducted by P. Scott Richards, PhD, at Brigham Young University, and Dr. Randy K. Hardman at the BYU-Idaho Counseling Center, to investigate the effectiveness of theistic spiritually oriented psychotherapy approaches. You were invited to participate because your counselor is experienced in providing spiritually oriented psychotherapy and has agreed to participate in this study.

Procedures

If you agree to participate in this research study, today you will be asked to complete a one-time intake questionnaire which should only take you 2 or 3 minutes to complete. Today you will also be asked to fill out two outcome questionnaires consisting of a total of 65 questions, which will take 7 – 8 minutes. These are designed to help your therapist and the researchers better understand how you have been doing emotionally and spiritually this past week. You will be asked to fill out shorter forms of these two outcome questionnaires (which should take no more than 5 – 6 minutes) before each of your regularly scheduled therapy sessions so that your therapist and the researchers can monitor your progress during treatment. These questionnaires assess the following:

- your perceptions of your spirituality, including your closeness to God, love for other people, and feelings of moral congruence and self-acceptance
- your perceptions of your physical health, relationships, behavior, thinking, emotions, work/school, and therapy progress

Long-Term Follow-Up


If you decide to participate in the study, you will also be invited to complete two brief, confidential follow-up assessment questionnaires at three, six, and twelve months after your treatment is completed. If you decide to participate in the follow-up assessments, the BYU-Idaho Counseling Center will send you an email on those occasions with a link to an assessment website where you can complete the confidential follow-up assessments in the convenience of your home. If you prefer, you may return to the BYU-Idaho Counseling Center in order to complete them.

Risks/Discomforts

The risks of involved in this study are minimal. However, assessing the on-going effects of psychotherapy can be emotionally threatening because it may more clearly reveal possible negative effects of treatment and/or problems in the therapeutic relationship. You should feel free to openly discuss with your therapist how you feel about the assessment procedures throughout the course of treatment. If the assessment procedures are too threatening, or if they seem to be getting in the way of treatment progress, instead of facilitating it, you will have the option of withdrawing from the study at any time, without jeopardizing your right to continue receiving treatment.

Benefits

There are no direct benefits for participating in this study. This study will add to the current research concerning the effectiveness of spiritually oriented psychotherapies. Ultimately, this study and others like it may increase the likelihood that religious and spiritually minded people will have the option of receiving mental health services from practitioners who are competent at providing spiritually sensitive and effective treatment approaches.

	Institutional Review Board	
	5-Nov-12 Approved	14-May-13 Expires

Confidentiality

You will be assigned an ID number. Only your ID number will be used in the online website to track the assessment and outcome data. Your name and other identifying information will NOT be recorded in the research data account. Therefore, there is no risk of violations of confidentiality in this study because the researchers will have no way of linking your ID number to your name. The researchers will not have access to psychotherapists' private case notes or to other client identifying information. Never will any data be shared in any form that will allow other members of the research team, or people outside of the research team, to link treatment outcomes with specific clients or psychotherapists.

Participation

Participation in this research study is voluntary. You have the right to withdraw at any time or refuse to participate entirely without affecting your treatment, current relationship with your therapist or other benefits to which you are entitled.

Questions about the Research

If you have questions regarding this study you may contact Dr. P. Scott Richards at 340 MCKB, Department of Counseling Psychology, Brigham Young University, Provo, Utah 84602 (scott_richards@byu.edu) or Dr. Randy K. Hardman, BYU-Idaho Counseling Center, Rexburg, Idaho 83460 (phone: 208.496.9370) (hardmanr@byui.edu) for further information.

Questions about Your Rights as a Research Participant

If you have questions regarding your rights as a research participant contact IRB Administrator at (801) 422-1461; A-285 ASB, Brigham Young University, Provo, UT 84602; irb@byu.edu.

Statement of Consent

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Name (Printed): _____ Signature _____ Date: _____

I also give my consent to allow the BYU-Idaho Counseling Center to contact me by email three, six, and twelve months after I complete treatment in order to invite me to complete the confidential follow-up assessment measures.

Name (Printed): _____ Signature _____ Date: _____



Consent to be a Research Subject (Therapist)

Introduction

This research study is being conducted by P. Scott Richards, PhD, at Brigham Young University to investigate the effectiveness of theistic spiritually oriented psychotherapy approaches. You were invited to participate because you are experienced in providing spiritually oriented psychotherapy.

Procedures

If you agree to participate in this research study, you will be asked to complete a brief Therapist Session Checklist after each therapy session you provide to clients who are also participating in this research study. The Therapist Session Checklist will be completed on a handheld device (e.g., Kindle Fire or iPad) and should only take you 1 or 2 minutes to complete. The Therapist Session Checklist asks you to briefly record what topics or issues were discussed in the therapy session, including what relationships and emotions were discussed. It also asks you to indicate what, if any, religious or spiritual interventions you used or recommended during the therapy session.

Risks/Discomforts


The risks of involved in this study are minimal. However, assessing the on-going effects of psychotherapy can be emotionally threatening because it may more clearly reveal possible negative effects of treatment and/or problems in the therapeutic relationship. There is a slight possibility that some clients may feel so threatening by the on-going assessments that they may decide to terminate treatment prematurely. You should feel free to openly discuss with your client how they feel about the assessment procedures throughout the course of treatment. If the assessment procedures are too threatening, or if they seem to be getting in the way of treatment progress, instead of facilitating it, your clients have the option of withdrawing from the study at any time, without jeopardizing their right to continue receiving treatment. You also have the right to withdraw from this treatment outcome study at any time for any reason.

Benefits

The potential benefit of participating in the present study is that it may actually enhance the effectiveness of the treatment you provide clients. More broadly, this study will add to the current research concerning the effectiveness of spiritually oriented psychotherapies. Ultimately, this study and others like it may increase the likelihood that religious and spiritually minded people will have the option of receiving mental health services from practitioners who are competent at providing spiritually sensitive and effective treatment approaches.

Confidentiality

You will be assigned an ID number. Only your ID number will be used in the online website to track the assessment and outcome data. Your name and other identifying information will NOT be recorded in the research data account. Only the principal researcher will have the ability to link your ID number to your name. The principal researcher will NOT have the ability to link client ID numbers with client names. The researchers will also not have access to your private case notes or to other client identifying information. Never will any data be shared in any form that will allow other members of the research team, or people outside of the research team, to link treatment outcomes with specific clients or psychotherapists.

	Institutional Review Board	
	5-Nov-12 Approved	14-May-13 Expires

Participation

Participation in this research study is voluntary. You have the right to withdraw at any time or refuse to participate entirely without jeopardy to your current relationship with the researcher, your standing at BYU-ID Counseling Center, and to your employment.

Questions about the Research

If you have questions regarding this study you may contact Dr. P. Scott Richards at 340 MCKB, Department of Counseling Psychology, Brigham Young University, Provo, Utah 84602 (scott_richards@byu.edu) for further information.


Questions about Your Rights as a Research Participant

If you have questions regarding your rights as a research participant contact IRB Administrator at (801) 422-1461; A-285 ASB, Brigham Young University, Provo, UT 84602; irb@byu.edu.

Statement of Consent

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Name (Printed): _____ Signature _____ Date: _____

	Institutional Review Board	
	5-Nov-12 Approved	14-May-13 Expires