Canadian Psychology / Psychologie canadienne
2012, Vol. 53, No. 2, 105–114

0708-5591/12/$12.00 DOI: 10.1037/a0027833

Errors in Treatment Outcome Monitoring: Implications for Real-World Psychotherapy

Andrew A. McAleavey and Samuel S. Nordberg
The Pennsylvania State University

David Kraus
Behavioral Health Laboratories, Marlborough, Massachusetts

Louis G. Castonguay
The Pennsylvania State University

In the last several years, considerable progress has been made in treatment outcome monitoring (TOM) in psychotherapy. Numerous instruments have been developed to assist practicing psychotherapists in assessing the impact of their services, and several clinical tools have been developed to directly improve the quality of services provided. As a field, we have begun to view these outcome monitoring and feedback systems with increasing confidence as evidence accrues to support their efficacy. In this paper we examine the types of errors that may occur in making inferences in TOM, in particular the determination of whether change in psychological symptoms is occurring, has occurred, or not. We examine this as any other empirical question, using the classic hypothesis-testing framework to describe two types of errors in decision-making. In particular, we discuss the strengths and vulnerabilities of two prominent assessment strategies in TOM (general and multidimensional measurement) in order to minimise inferential errors and maximize the effectiveness of outcome monitoring efforts. Finally, we provide a few examples of new developments that make use of multidimensional measurement to minimise Type II errors to improve outcomes.

Keywords: treatment outcome monitoring, psychotherapy, measurement, the Treatment Outcome Package (TOP)

Although many mental health professionals can, when called upon, identify what they consider their own specific areas of expertise or specialisation, most treat clients with a variety of diagnoses, comorbidity, and personalities. Similarly, any large-scale organisation for mental health care, be it a clinic, hospital, or a managed care company, will necessarily need to provide treatments for a broad range of clinical problems: given enough time or a large enough client population, virtually all diagnoses listed in the major diagnostic systems will be represented in a caseload. In addition, practicing psychotherapists and counselors frequently do not specialise in only one level of severity of clientele. Psychotherapists may have some relatively less severe clients, and they may have other clients whose clinical severity may require longer treatments, more intense intervention, and/or more frequent appointments. The result is tremendous heterogeneity among clients on a given therapist's caseload.

This heterogeneity in patient presentations—defined at a minimum by the type of symptoms present as well as their level of severity—is an inescapable part of everyday psychotherapeutic practice. In addition to having direct impacts on the assessment, case formulation, and treatment planning for each client, this heterogeneity may also have important implications on a set of tasks in which an increasing number of clinicians (as well as large-scale management organisations) are invested: treatment outcome monitoring (TOM).

In the last several years, several instruments have been developed and validated to assess the impact of psychotherapy in day-to-day practice. The implementation of these instruments (and associated clinical tools and interpretive aides) has largely been driven by empirical evidence supporting their clinical usefulness. However, heterogeneity of clients may influence the accuracy of inferences about clients and therapists that can be made from this type of outcome tracking. The goal of this article is to describe some potential inferential errors of two different types of self-report measures used for treatment outcome monitoring—general and multidimensional. As such instruments are likely to be used frequently in day-to-day practice, the article also discusses some clinical implications of these potential errors.

Statistical Decision Making

The most common approach to statistical decision making will be familiar to most readers. In this approach, the general goal is to decide whether a given set of data support or reject a particular null hypothesis. This approach, though not without critics (e.g., Wagenmakers, Wetzels, Borsboom, & van der

Andrew A. McAleavey, Samuel S. Nordberg, and Louis G. Castonguay, Department of Psychology, The Pennsylvania State University; David Kraus, Behavioral Health Laboratories, Marlborough, Massachusetts.

Correspondence concerning this article should be addressed to Andrew A. McAleavey, Department of Psychology, The Pennsylvania State University, University Park, PA 16802. E-mail: aam239@psu.edu

Maas, 2011), has proven to be a remarkably generalisable tool for decision-making. One of its most basic functions is to determine the probability of correct and incorrect inferences about a particular hypothesis, which is a familiar and relatively simple undertaking.[1] Generally, a 2 × 2 table is presented to show all four possible situations in which the null hypothesis is either accepted or rejected and whether this is, with regard to an unknowable "true" population, a correct or incorrect inference. In the simplest and most common instance, the statistical question comes down to a determination of whether a number representing an effect (which could be a mean) is different than 0. In this case, the null hypothesis is that the effect in the population is really 0, and the alternative hypothesis is that the effect is not 0.

As the table shows, in the general case there are two theoretical ways to be correct and two theoretical ways to be incorrect. The two correct decisions are either to accept the null hypothesis when it is in reality true for the population, or to reject the null hypothesis when it is in reality false for the population. When the latter of these is expressed as a probability, it is referred to as statistical *power*. The incorrect decisions in this table have been called Type I and Type II errors, and there has been extensive discussion on the importance of minimising each in any task requiring statistical inference. Type I errors are considered false positives, since the decision made based on the data would suggest that an effect is present, when in reality, no such effect is present in the population. Type II errors, in contrast, are failures of detection, because a true effect is present in the population, but the data do not sufficiently support this inference.

In research methods, the implications and relative costs of Type I and Type II errors are well known and have had dramatic influence on the way psychologists conduct and report research studies. For instance, the chance of making a Type I error in a research study can be determined a priori and at the will of the researchers, by setting the alpha level for the statistical tests. Despite numerous and frequent acknowledgments that this is an imperfect adjustment for the possibility of Type I error, most psychology research is conducted with an alpha level set to .05, which theoretically means that up to one in 20 statistically significant findings reported in the field may represent Type I errors. To counteract this problem, the field of psychology has emphasised the need for repetition and replication of findings over time, which can successively reduce the likelihood that a given significant finding is actually driven by error.

In order to minimise Type II errors and maximise power, researchers have often used the method of increasing sample size. Given more observations, the power of a statistical test will increase, which will reduce the likelihood of Type II error. The research community has grown quite accustomed to altering the way that it evaluates and conducts research on groups of individuals because of these errors, and the recommendations to replicate findings and increase sample size are generally second nature to most psychological researchers at this point.

We suspect that an analysis of potential Type I and Type II errors in TOM will be profitable to the field, and ultimately improve the accuracy of inferences made about individual psychotherapy clients as well as counselors and psychotherapists.

Table 1
Null Hypothesis Testing Inference Table

Decision	Population true value: No effect is present	Population true value: Effect is present
Accept the null hypothesis: no effect is present	Correct	Incorrect Beta Type II Error
Reject the null hypothesis: effect is present	Incorrect Alpha Type I Error	Correct 1 − Beta Power

Treatment Outcome Monitoring

In this article, we take a definition of TOM to mean any systematic, repeated assessment of psychological variables during the course of a psychological treatment, which in turn may be modified in response to the results of the monitoring process (e.g., via feedback to the therapist). Most monitoring and feedback systems have focused primarily on levels of psychological functioning and symptoms of psychological disorders. Even within this subgroup of TOM instruments, many different systems have been devised for the purpose of TOM, varying in the types of outcome information collected and eventually fed back to the therapist, but also across the types of settings (e.g., inpatient, outpatient, and training clinics), clients (e.g., adolescents, college students, families, or adults), and languages (large systems have been developed in English, German, and Dutch, e.g.). Despite the diversity of these systems, they all share a basic function: to take incoming quantitative data and provide psychotherapists with a relatively simple interpretation of the current state of the client's problems. It is important to point out here, that it is generally the particular system of TOM that defines which variables are necessary in order to understand the client's current state (by providing the scale that is measured), which may be difficult to define across a heterogeneous population.

Perhaps the most common (and perhaps the most important) interpretation to be made with these systems is answering this question: "Has this particular client experienced significant change in his or her level of symptoms (Howard, Moras, Brill, Martinovich, & Lutz, 1996)?" This question can be construed as a relatively simple hypothesis test. The null hypothesis for this test is that no change has taken place. The alternate hypothesis is that some change (either improvement or deterioration) has occurred. For illustration, let us assume that the magnitude of change—an issue that is often addressed in this literature through tests of clinical significance (e.g., Jacobson & Truax, 1991)—is incorporated into the question, in order to differentiate between statically

[1] For the sake of clarity, we only discuss correct and incorrect inferences here, which are often the primary outcomes of such statistical tests. More specifically, these tests provide the likelihood of the observed data, given a particular (null) hypothesis, and that probability is then used to make an inference.

significant but small change and changes of sufficient magnitude to be clinically meaningful. While the issue of clinical significance has been the subject of much discussion and deserves its own treatment, for this article we will assume that the amount of change necessary to be "meaningful" is a known quantity and simply restate the original question: "Has this particular client experienced meaningful change in his or her level of symptoms?" This is the main question answered by many or most TOM systems.

Heterogeneity of Clinical Problems

Although the straightforward question posed above promises an intuitive answer with logical consequences (e.g., this client has not yet experienced meaningful change, so continue treatment and consider a change of strategy), the heterogeneity of presentations seen in psychotherapy brings a new question to the fore: "Change in what, exactly?" Obviously, clinical problems and psychiatric disorders are multifaceted entities, so when assessing change in psychological functioning, a clinician may be interested in any number of different constructs (see Castonguay & Oltmanns, in press). For instance, if the presenting problem is depression, one could see improvements in sleep and overall energy levels, decreases in suicidal thoughts, improved performance at work, or all three signs as improvements in depression. In addition, the highly comorbid nature of psychiatric conditions suggests that if a client is depressed, there are likely to be additional content areas of concern which may or may not be functionally related to the depression for a given individual. These might include some of the many clinical features that are frequently associated with depression, such as marital, health, and occupational problems. Since these factors can contribute to, exacerbate, and/or maintain depressive symptoms, it is wise for clinicians to assess them prior to, during, and before terminating therapy (LeMoult, Castonguay, Joorman, & McAleavey, in press).

Moreover, when considering multiple clients (and any measure of psychological symptoms, if it can provide generalisable information, must be able to assess multiple clients), one must ask whether change for one person is qualitatively the same as change for another. If one client is depressed, but another has substance abuse, most clinicians would not expect that both clients would experience the same type or level of benefits from therapy. Most clinicians would likely focus on the particular needs and concerns of each client and use appropriate interventions to reduce the most disturbing symptoms, based on their assessment and case formulation. But how can any measure of psychological functioning be applicable to both clients?

There are two major, nonmutually exclusive, solutions to this problem in terms of measurement strategies: unidimensional and multidimensional (Hill & Lambert, 2004). The unidimensional measurement strategy is to measure one construct, while the multidimensional strategy aims to measure more than one (often five or more). Of important note: This is a distinction between assessment strategies, rather than instruments. Though it may seem counterintuitive that measuring a single construct is a solution to the problem of heterogeneity, unidimensional measurement and interpretation is logical and quite popular. Frequently, the construct assessed in a unidimensional measure is not specific to a single group of experiences, as, for instance, the Yale-Brown Obsessive-Compulsive Scale (Goodman et al., 1989a, 1989b) is specific to symptoms of obsessive–compulsive disorder, but is rather a measure of a *general* construct, such as general distress or overall severity, and it is these general measurements that we discuss here. Some examples of general measurements include the Global Severity Index from the Baseline Severity Index (BSI; Derogatis & Melisaratos, 1983) and Symptom Checklist-90 Revised (SCL-90–R; Derogatis, 1994), the total score from the Outcome Questionnaire-45 (OQ-45; Lambert et al., 1996), the clinical score of the clinical outcomes in Routine Evaluation-Outcome Measure (Evans et al., 2002), and the summary score of the Behaviour and Symptom Identification Scale-24 (BASIS-24; Eisen, Normand, Benager, Spiro, & Esch, 2004). These are some of the most common measures in clinical psychology. Many general measures seek to assess what might be called quantitative differences between individuals (and within individuals over time), in that two different scores can be compared on a quantitative scale. However, general measures can only assess qualitative differences through less formal means (e.g., through comparison of responses to specific items); that is, if there are different types of people or distress, a general score may not be able to detect it.

The multidimensional measurement strategy, on the other hand, attempts to assess qualitative differences between individuals by providing multiple, more *specific,* assessments simultaneously, not all of which would be expected to be elevated for all clients. For example, the Treatment Outcome Package (TOP; Kraus, Seligman, & Jordan, 2005) has 12 subscales measuring such problems as depression, panic, quality of life, substance use, work functioning, and sleep functioning. The SCL-90–R, which contains a general measure of severity as mentioned above, also includes specific subscales such as depression, anxiety, hostility, phobic anxiety, and paranoid ideation. A very widely used instrument in college counselling, the Counselling Centre Assessment of Psychological Symptoms-62 (CCAPS-62; Locke et al., 2011), is also multidimensional. In essence, many multidimensional measures can be viewed as a set of measures of separate content areas (although it should be noted that these multiple measures are often correlated). Due to the fact that several multidimensional instruments provide severity scores both for specific subscales as well as some form of general measure (e.g., Boswell, Kraus, Nordberg, & Castonguay, 2009), it may not be productive to discuss *instruments* as either general or multidimensional. In this paper we use the terms *general measures* (e.g., the BSI from the SCL-90–R) and *multidimensional measures* (e.g., the subscales of the CCAPS-62) to refer to their respective assessment strategies, the specific scores or subscales from instruments using these strategies, and the implications of these strategies, not to the instruments (e.g., BDI, TOP, OQ-45) containing these measures. That is, in this paper, we are discussing whether a therapist or administrator is using a general or multidimensional strategy to assess treatment outcome, not what instrument(s) they use to do so, because a single instrument may contain both multidimensional and general measures.

Though general and multidimensional assessment strategies each have strengths (see Hill & Lambert, 2004), it is important to discuss the potential errors (Type I and Type II) of inference that may occur when either is considered in the context of outcome monitoring. We have attempted to list some potential vulnerabilities of each strategy, though this list is not intended to be comprehensive.

Type I Errors

In the context of TOM, Type I errors occur when no meaningful symptom change has occurred, but the inference made is that change has happened. Given that most (but not all) change observed during psychotherapy is in the positive direction (Lambert, Hansen, & Finch, 2001; Lutz et al., 2006), we may assume that most of the time in this situation, the assessment will show that a significant improvement has taken place, when in reality the client has not improved (and may have even deteriorated). In a generic sense, such a faulty inference could have very important ramifications for an individual psychotherapy dyad as well as for larger systems. For one thing, it may provide the therapist with false evidence that the treatment, as provided, is adequate, which would encourage him or her to either begin the termination phase of therapy prematurely or to continue providing services that are unhelpful. For an insurance or health care management system, this could lead to financial rewards to the therapist for a case that is actually likely to lead to neither improved overall health nor reduced cost of care.

In most situations, general and specific measures are likely to be subject to different kinds of Type I errors. For example, a general distress score, like one derived from some unidimensional measures, is especially vulnerable to a particular kind of Type I error related to the heterogeneous content of the items generating a general distress score. Conceivably, small improvements on several content-diverse items or subscales could create an aggregate appearance of overall improvement. While this is actually one of the main goals of such general scores, if these changes are individually less than meaningful to the individual client and/or are functionally unrelated to one another, the accurate inference might be that there is no meaningful change, even if the sum of the small changes appears meaningful. For instance, if a client indicates that he or she is experiencing improvements so slight that they may not be reliably detected in items assessing relationship quality, sleep patterns, sweaty palms, and work performance, but the slight improvements are either artifacts of measurement error or else are so small as to have no discernable bearing on the client at all, summing these small changes could misleadingly suggest that a change has occurred despite a reasonable certainty of the opposite.

When interpreting the results of a general score, therefore, we must be open to the idea that if the changes noted by the client reflect only slight and diverse improvements, the appearance of a significant change may be a result of a Type I error (though obviously, not every instance of this is likely to be an error). When this is not an error, such changes ought to represent true general improvements in symptom severity or overall functioning rather than any specific gain in particular, which is exactly what these measures are designed to assess.

While multidimensional measures are not particularly prone to the sort of Type I errors discussed above, there are at least two other examples of Type I errors that may be more common for multidimensional measures. The first is akin to "experiment-wise" Type I error inflation. Since multidimensional measures provide the opportunity to make inferences about multiple constructs, this is the equivalent of conducting multiple hypothesis tests within one experiment. Researchers have long known that controlling for experiment-wise Type I error is a necessary process in such circumstances, because of the possibility of increasing Type I error rates with multiple tests. Thus, with any multidimensional TOM system, the likelihood of producing a Type I error increases with the number of specific subscales examined, which may be a particular consideration when many different subscales are elevated for a given client. We are not aware of any systematic approach to addressing this concern when using a multidimensional measure, so experiment-wise Type I error rates may be higher for multidimensional measures than for general measures.

Another potential Type I error with a multidimensional measure would occur if temporary and extraneous circumstances obviate a specific measure at the time of administration. For instance, a client with chronic worry related to work may experience a meaningful decrease in anxiety about work functioning while on vacation, but this is more akin to negative reinforcement than treatment success. Therefore, any inference suggesting that reliable change has occurred could lead to misleading conclusions (e.g., that the client has successfully managed symptoms of anxiety), unless it is interpreted strictly in the context of that day and time. While careful clinical or circumstantial assessment will be able to detect such cases, implementing a system that could do so automatically would be difficult.

Type II Errors

Type II errors are errors of failed detection. In a TOM setting, these would primarily include instances in which real and meaningful change has occurred for a patient, but the measurement used suggests that change has not taken place. A particular vulnerability of general scores may be seen when there is real change but multiple components of the general score change in different directions simultaneously for a given client. For instance, if a client begins to report decreased symptoms of anxiety and depression, but items relating to problematic substance use begin to increase significantly at the same time, a general distress score may not be able to separate the improvement from the deterioration. Instead of the more complex but potentially more accurate inference that the client is using alcohol or other drugs to maladaptively self-medicate (see Pihl, & Stewart, in press), the clinician may be led to believe that the client is either not changing, or even steadily improving (which would potentially represent a kind of Type I error as well). Needless to say, this kind of error could be potentially damaging to the treatment's outcome if the therapist does not discover it.

Another potential Type II error that may be more common with a general measure of distress is a simpler signal detection error. When important changes for a particular patient are actually confined to a relatively specific domain of functioning, a general score may not be able to detect it as readily as a domain-specific measure. For instance, people with panic disorder frequently also experience symptoms of generalised anxiety and depression (Teachman, Goldfried, & Clerkin, in press). A panic-focused treatment may have more rapid effects on just the symptoms of panic than on the symptoms of depression and generalised anxiety, and at a minimum, the client and therapist may be interested in tracking just the panic symptoms as a primary outcome. In this case, a general score may not be able to detect small changes in panic symptoms due to fluctuations in other symptoms, even if those symptoms are changing as a direct effect of treatment. And if the panic symptoms are of interest as a specific outcome, a subscale

score is likely to be a more reliable measure than any individual item(s) examined informally (though this is not necessarily a bad strategy; see below).

Similarly, a multidimensional measure would also potentially have some vulnerabilities to Type II errors. For instance, when changes in psychotherapy are, in fact, a function of general distress (irrespective of distress content or type), and these changes are relatively small, a multidimensional measure may not be as sensitive as a unidimensional measure. The multiple dimensions would only be able to detect changes that are reliable across multiple domains. One prime example of this may be the concept of demoralisation, as described by Jerome Frank (1961), Ken Howard (Howard, Lueger, Maling, & Martinovich, 1993), and others. These theories of psychotherapeutic change propose that one common feature of many or all psychotherapy clients when they seek treatment is a generalised loss of hope and belief in their own ability to cope with stress. Howard proposed that the first phase of psychotherapy is primarily a remoralisation phase, in which clients begin to generally feel more capable and hopeful. Unless a multidimensional measure explicitly assesses a concept of general remoralisation or well-being (as some do, e.g., the Behavioural Health Measure-20, Kopta & Lowry, 2002), these broad-based and symptom-independent improvements might not appear as reliable changes across several dimensions.

Type II errors present a serious challenge to the interpretation of both unidimensional and multidimensional measures, particularly in instances where the client may be degrading in meaningful ways though this deterioration is not detected. Prior research has indicated the possibility that TOM systems are most useful in these cases (De Jong, Van Sluis, Nugter, Heiser, & Spinhoven, in press; Shimokawa, Lambert, & Smart, 2010), and the possibility of failed detection of deterioration is potentially vital. While not every TOM system has been empirically shown to be the cause of significant improvements in treatment efficacy, any system that can more reliably predict treatment failure and deterioration is very important, because this has been seen as one of the major positive outcomes of the TOM movement in general (e.g., Lutz, Böhnke, & Köck, 2011).

Overall Comparisons Between Multidimensional and General Measures

Although we have illustrated some of the vulnerabilities of both multidimensional and general strategies to Type I and Type II errors, it remains to be seen if there are clear differences between the two strategies in terms of the types of errors that are likely to occur. General scores, as we have discussed, may be particularly vulnerable to Type II errors: There may be circumstances in which general scores do not capture the real types of changes occurring for clients in psychotherapy. On the other hand, multidimensional measures necessarily perform many more inferences than do general scores, and so are exposed to the risk of additional Type I errors (false alarms). With this in mind, it may be the case that there is a general trade-off between what types of errors are most likely to affect the inferences in these two different measurement strategies. To minimise Type I errors, we may want to rely on the parsimonious general scores, but to minimise Type II errors, clinicians, administrators, and researchers may need to examine more constructs that are subject to change than just a general distress factor.

Implications for Practice

It has been shown with some regularity that repeated use of measures of psychological symptoms can facilitate treatment by examining change in problem areas over time (Lambert, Hansen, et al., 2001; Lambert, Whipple, et al., 2001). However, in any such use of TOM for decision making, Type I and Type II errors can (and necessarily do) interfere with interpretation. In order to accommodate these potential errors, we make the following suggestions for use in clinical practice.

Use both general and specific measures. By using both a general and multidimensional measure, Type I and Type II errors can be mitigated. One challenge to using two different measures is that, given the diverse instruments available, it may be difficult to find measures that adequately assess the constructs that are of particular clinical interest. One solution to this particular challenge could be to use a single instrument that includes specific subscales and a general score (e.g., SCL-90–R, BASIS-24). These are not without their limitations as well—measures designed to have discrete subscales might not necessarily be well-suited to a general score generated from those discrete scales—however, some methods have shown promise in extracting items that load on a "distress" factor, in addition to a discrete subscale.

Another option would be to ask clients to complete both an instrument that has been commonly used in research and practice as a general score (e.g., the OQ-45) and an instrument that has been primarily designed to capture several dimensions of symptoms and functioning (e.g., TOP, CCAPS-62). While this could impose additional time for clients to complete outcome measurement, the combination of such instruments would potentially result in an assessment battery that would include fewer than 100 items, or the measures could be given at alternate sessions, minimising the burden on clients. The use of both types of instruments by a large number of clinicians could lead to very interesting research investigations of the convergent and divergent validity (as well as differential predictive validity) of these outcome measures across a variety of client populations, theoretical orientations, treatment modalities, and therapeutic milieux. Practice research networks, which are designed to facilitate collaboration between clinicians and researchers, may represent the optimal setting for such clinically relevant and scientifically important studies (see Castonguay, 2011). This could also be a promising way for clinicians to develop practice-based evidence of their own effectiveness, given a heterogeneous client caseload.

Use items to examine clinically important nuances of general scores. With general measures, using individual items to better understand the nature of therapeutic change, or lack thereof, can help to reduce the likelihood of both Type I and Type II errors. For example, when some items move up, and others move down, a general score can remain the same. Examining individual items can help a clinician appreciate that there has been change that is not reflected in the general score. Similarly, if a general score changes based on small increments in a large number of items, a clinician can identify this by comparing items over time, and assessing the relative stability of the client's reporting. However, if there are complex patterns, such an informal analysis could easily

begin to take a relatively large amount of time, which many clinicians may not have available.

One challenge to making interpretations based on individual items is that the psychometric properties of individual items (which are generally less reliable than larger scales) are not often known or made explicit. How meaningful is a one-point shift on a five-point Likert scale? If the question is, "I have had thoughts of ending my life," such a change might be extremely meaningful. On a response to the question, "My hands frequently shake uncontrollably," the interpretation may be less clear. When individual items are broken out from an instrument, the scientific rigor that went into the development of that instrument (norms, factor structure, scale building, etc.) is no longer applicable. While item-response theory can provide an empirical basis for interpreting these differences, in clinical practice, reliance on individual items is likely to be an imperfect method. This particular challenge may be at least partially overcome in clinical practice by bringing the item in question to the client's attention, and determining whether the perceived shift (or lack thereof) is consistent with his or her experience.

Consider multidimensional measures holistically. Just as it is possible to understand the specific components of a general score by examining items, it may be possible to get a holistic assessment of general function from several specific scales. A general score can provide a good overview of the client's functioning, and using multiple subscales to assess a client at a given time is likely to provide a detailed assessment that would be nearly impossible to perform otherwise. Who could claim to be able to assess alcohol use, violence, work functioning, suicidality, sleep, health, depression, and anxiety at every session, while at the same time being able to fully and tactfully address technical, relational, and personal (for the client and therapist) issues that can facilitate or interfere with the therapeutic process? However, it is important not to lose the forest for the trees: As noted above, if small and general changes occur during treatment, these might not appear as meaningful shifts on any domain-specific subscales (a Type II error). Thus, when using a multidimensional measure in the absence of a validated general score, clinicians should consider all of the specific scores in concert. If 6 of 8 subscales show slight improvement, say, this may be informally understood as a potential, small, general change. Obviously, there are drawbacks to this approach, including the potential for seeing a nonsignificant change as significant, but it may help clinicians better understand their clients' overall functioning.

Consult the client. Perhaps the most important suggestion is to acknowledge the inherent error in all such quantitative measures and use an additional source to ground the results: the client. Different clients may interpret items differently (e.g., Nesselroade, Gerstorf, Hardy, & Ram, 2007), and respond in ways that have meaning to them but, perhaps, not to the clinician. By working with a client to better understand the meaning of change (or lack thereof) on any given self-report measure, a clinician can adjust his or her interpretations of scores to better reflect the client's experience. Building a shared consensus can foster the therapeutic alliance as well as better integrate the measure into treatment. Additionally, the clinician can use his or her judgment to determine whether change on an individual item is meaningful to a client, thus mitigating some of the difficulty mentioned above. Obviously, such informal and clinical means come with their own drawbacks, but in conjunction with a quantitative measure, these dangers of clinical judgment may be ameliorated.

Implications for Clinics and Administrators

At an administrative level, using brief measures of symptoms to assess effectiveness and efficiency of psychotherapy and psychotherapists should be possible; however, such use should be tempered by a clear understanding of the potential Type I and Type II errors. There are at least two broad administrative domains in which repeated measurements could be useful: assessment of the effectiveness of a particular treatment course for a client, and assessment of the effectiveness and efficiency of a particular therapist. Both domains are important and can lead to significant improvements in outcomes, but are also rife with potential pitfalls related to false positives and false negatives. For example, different clinicians may have very different ways of incorporating measures into their practices. Some may make the measure a focal point of each session, discussing scores and interpreting patient's responses. Others may rarely mention the measure during practice. An administrator should expect differences in response based on these two different approaches. For example, it is conceivable that clients who consistently discuss their scores and are made aware of their responses will report changes in their symptoms more precisely, as they will be closely attuned to past responses. In such cases, smaller changes may be more meaningful than in cases where clients are less mindful of their responses. As shown by De Jong et al. (in press), there are significant differences between therapists in terms of both whether they use feedback when it is provided, and how effective it may be: For some therapists, feedback is highly useful, and for others, it may not be effective even when they use it clinically. More studies are needed to better understand the sources of these differences.

It should go without saying that great care must be applied to the process of assessing a therapist's effectiveness based on brief measures of psychological symptoms. For example, in the treatment of PTSD, a large amount of evidence supports the theory that evoking distress and exposing clients to that distress in and between sessions is a vitally important component of treatment (Foa & Meadows, 1997; Nemeroff, 2006). This kind of therapy could result in significant but hopefully temporary deterioration as reflected in a general measure, as the client reports increased distress and discomfort, or a nonsignificant total score change, as some real improvements are wiped out by worsening distress. Without careful understanding of the context, an administrator could determine that the treatment is going poorly, or is ineffective. Multidimensional measures have the potential to provide some of that context, but only if they measure the appropriate domains separately. For instance, it may be the case that in an exposure-based treatment, a client would report increased symptoms of anxiety due to the exposures, but this may be associated with a decrease in other problems such as depression (as they client feels less helpless to address his or her problems) or increases in relationship functioning or quality of life. Thus, if the measure is able to capture this, it may be very helpful to interpreting outcomes.

Novel Applications of a Multidimensional Instrument (the TOP) in TOM

In our opinion, multidimensional measurements offer a potential benefit to TOM processes: The ability to detect certain Type II errors that may be occurring with general scores. Using the TOP (Kraus, Seligman, & Jordan, 2005), there have been several attempts to apply new techniques in TOM to help solve problems associated particularly with Type II errors. We will describe two such explorations here: One focused on predicting and (potentially) preventing damaging psychiatric crises, and the other focused on accurately and fairly evaluating the effectiveness of therapists.

Predicting psychiatric hospitalisation. The first question facing researchers and clinicians in considering a multidimensional strategy ought to be one of incremental validity: Do the multiple measurements add value above and beyond a total distress score? If this is not true, then claims of reducing error are suspect. As described below, three independent analyses by managed behavioural health organisations and insurance companies have demonstrated the predictive validity of the TOP's multidimensional structure. In each case, multidimensional TOP data was merged with claims data so that current and future events could be analysed over time, including overall medical expenditures and behavioural hospitalisations. Such modelling is especially important for justifying the utility of well-delivered behavioural health care in controlling health care costs.

The opportunity for behavioural health interventions to deliver tremendous savings has been known. Individuals with chronic medical problems like diabetes, heart disease, and cancer make up approximately 20% of the population, yet they consume 80% of health care dollars (Berk & Monheit, 2001). In addition, this population is highly comorbid (greater than 40%) with behavioural health disorders (Wells, Golding, & Burnam, 1988). Therefore, this comorbid group represents approximately 8% of the population. Compared with their noncomorbid chronic medical patients, this behaviourally disordered group is estimated to have health care costs that are 300–500% higher (Sheehan, 2002). As a consequence, this small segment of the population consumes approximately 54% of health care dollars. Health care providers and systems would be perhaps better served to have tools that can identify these patients and offer alternative treatments in advance of their most pressing needs, even if this means that a portion of the patients identified for these risks would not have actually needed it (i.e., in many of these situations, Type I errors may be more costly than Type II errors).

To assist in helping to lower the cost of care while simultaneously improving its quality, Behavioral Health Laboratories (BHL; Marlborough, MA)—the company that processes TOP data—built a hospitalisation prediction algorithm for Blue Cross and Blue Shield of Massachusetts. This hospitalisation alert algorithm was designed to identify health plan members at high risk for behavioural hospitalisations and then to provide meaningful feedback to those involved in treatment. Behavioural health hospitalisations are low-incident events and therefore difficult to predict, with only approximately 0.35% of the population hospitalised annually. They are also potentially disruptive and stigmatising for individuals and their significant others and extremely expensive for insurers. Accordingly, offering health care providers with ways to predict and potentially prevent hospitalisations (by offering additional, more frequent, intense and/or targeted outpatient services, e.g.) may reduce costs for third party payers and distress for clients, by refocusing services—in essence, similar to providing clinicians with feedback on not-on-track cases. However, hospitalisation is so rare and difficult to predict that a relatively simple signal like a sudden increase in symptom severity may not be sensitive enough to produce accurate alerts (a Type II error).

The initial version of the algorithm relied on patients' scores on TOP subscales, such as violence, suicidality, depression, psychosis, and substance abuse. Scores were computed for all Blue Cross members for whom a TOP was administered, with a score ranging from 0–5 (0 representing *no or little risk* and 5 representing *maximum risk* for hospitalisation), and delivered to clinicians and Blue Cross case managers. In introducing the program, the health plan attempted to increase adoption rates by eliminating other paper work (i.e., prior authorisation forms which place a significant burden on clinicians and divulge sensitive and confidential client information). Initial analysis of the first 24,000 patients revealed a lognormal distribution of risk scores, with 14.9% of the assessed population being assigned nonzero scores. Blue Cross analysis of future claims data demonstrated that all hospitalisations came from patients with nonzero scores, suggesting a test sensitivity approaching 100%. Specificity of the initial algorithm was 18% for all nonzero alert scores (a rate 51.4 times higher than the baseline hospitalisation prevalence rate). The higher scores (i.e., 4s and 5s) were not associated with a higher hospitalisation rate, but were associated with hospitalisations that occurred with fewer days between TOP administration and the hospitalisation event.

The second generation of the algorithm included other TOP-collected data (such as comorbid medical conditions and number of previous psychiatric hospitalisations) and improved the specificity to 40%. As reported by Blue Cross: "Alerts 1–5 are identifying patients at risk for inpatient at 4–5 times the rate of former methods" (Kelly, O'Donnell, Pelletier, & Simmons, 2008). The program calculated claims savings of $3.8 million, primarily through an estimated 20% savings on behavioural hospitalisation (Fitzgibbons, 2006). This algorithm required multiple variables to achieve such results, suggesting that TOM may benefit from using multiple specific pieces of information about individuals rather than a general or total score in exclusivity.

A similar analysis by Neighborhood Health Plan—an insurance company specialising in low-income Medicaid and Medicare recipients—documented similar specificity rates for the hospital alert algorithms (Neighborhood Health Plan, 2008). Finally, a subsidiary of Value Options—a United States managed behavioural health organisation—used TOP data on approximately 60,000 members to build internal models to predict high-cost health care utilisation. The initial results documented the predictive validity of TOP subscales and even individual TOP items (Stelk & Berger, 2009).

These predictive analytics highlight the importance of multiple subscales. Although TOP domains are highly correlated with each other (Kraus, Seligman, & Jordan, 2005), each TOP subscale accounts for unique variance and improves the specificity of the predictions in hospitalisation and treatment cost. The TOP total score did not contribute to the final model in these analyses, suggesting that for predictive analytics and the ability of outcome

systems to help identify and prevent high-cost and high-distress events, multidimensional approaches may be required.

Evaluating therapist effectiveness using multiple dimensions. Another demonstration of the importance of a multidimensional approach can be seen in evaluating or ranking therapists based on their effectiveness. The general outcome approaches are limited to rating clinicians' effectiveness on a single dimension and result in global ratings like "above average," "average," and "below average." With increasing public demand for health care outcome transparency, the near future may include public ratings of clinicians using quantitative tools and rating systems. If these single-dimensional ratings are equally distributed, each clinician will only have a 33% chance of drawing strong public interest. However, data suggests that therapist competency and effectiveness is not a global construct. With its multidimensional assessment of outcome and change, the TOP appears to detect specific areas of expertise, thereby reducing the risk of Type II errors that unidimensional measures might be particularly at risk to commit toward therapists.

In a recent study involving a large number of therapists (N = 696) and clients (N = 6,960), Kraus, Castonguay, Boswell, Nordberg, and Hayes (2011) found that no one therapist showed superior results with all clinical problems measured by the TOP. It is important to note, however, that most therapists appear to have superior outcomes with respect to particular problems (e.g., depression, anxiety, suicide, substance abuse) but not with others. Using the same sample, Nordberg et al., (2010) found that therapist effectiveness in treating one set of symptoms did not predict effectiveness in treating another. Specifically, a clinician's effectiveness in treating depression correlated at r = .18 with his or her effectiveness at treating substance abuse. Further complicating the difficulty in accurately "rating" clinicians, Nordberg et al. (2010) demonstrated that for clients with both depression and substance abuse, it may be essential to look at a clinicians' skill in treating both dimensions, not just one.

The use of a multidimensional assessment tool like the TOP may reduce, interestingly, a particular fear experienced by many clinicians: That outcome instruments could lead health management companies to harm their livelihood. The study by Kraus et al., (2011), found that 96% of clinicians were very good at treating at least one set of symptoms and that the median clinician had at least 5 of the 12 TOP dimensions in which he or she was reliably effective. In addition, the same study also suggested that very few clinicians showed a lack of effectiveness at treating the wide range of symptoms and dimensions of functioning measured by the TOP. If replicated, such a model of assessment of therapist's expertise is likely to be more helpful for client referral (see Youn, Kraus, & Castonguay, in press), and potentially more rewarding for clinicians than a single-dimensional model, where 67% of clinicians will "fail" to meet the grade. This suggests that a general rating system of clinicians would be grossly misleading and quite possibly counterproductive.

Conclusion

It is clear that there are challenges to interpreting measures used in TOM, whether they are general or multidimensional measures. By understanding that all measures will make some errors (which is a simple necessity of making inferences), we may be able to begin to discuss which errors are especially likely and/or damaging. This is an important issue, as there are likely to be real differences between general and multidimensional measures in terms of what mistakes get made: With a general score, we believe that a Type II error is more likely, as change may be more complex than a general measure can detect; conversely, a multidimensional score may be prone to Type I errors to a greater degree than would a general measure of distress. Though the case is much more complex than a simple trade-off between the two errors, using the strengths of each can likely lead to improved TOM tools.

When working with individual clients, we suggest that clinicians approach the use of measures as one would any tool: as a facilitating agent, not a replacement for experience or training. In this context, daily practice may be augmented by the judicious use of general or multidimensional measures, with the understanding that the former may miss important domain-specific change, and the latter may falsely identify the same. Our suggestion, given the limitations on time inherent in regular practice, is that a multidimensional measure, given at regular intervals, should be the most facilitative of practice. This conclusion derives from a belief that it is better to observe (and clinically confirm) a false alarm, than it is to miss potentially meaningful change. Ideally, a measure with both multidimensional scores and a general score would be best-suited for a practice-setting, with the important caveat that checking in with each client regarding his or her responses will help calibrate a clinician's understanding of what a client is communicating.

With regard to use in administration, the same time constraints hold. We suggest that administrators also approach the use of brief instruments from the perspective of using them to augment, not replace, experience and wisdom. The greatest challenge to use of brief instruments in an administrative context is the abstraction of these instruments from the person of the client. While a clinician may check in with clients to calibrate their understanding of responses, administrators are not typically able to do so. Thus, given the heterogeneity in client populations, administrators should pay extra attention to the errors inherent in TOM measures. Finding an informed balance between false alarms and failures of detection is difficult at this stage, given the variety of cases in any given setting, and the relative lack of knowledge about how well brief self-report measures capture heterogeneity. However, if important and meaningful structural decisions are to be based on these actuarial instruments, knowing what types of errors are likely to be produced is vital. The more common general score-based methods are conservative, in that they may best minimise false alarms (Type I errors) at the cost of increased failure to detect heterogeneity in change (Type II errors). In some cases this may be adequate, but if, for instance, an administrator bases therapist evaluations (or pay) on a general score, that administrator may not optimally identify therapeutic success and failure, particularly if that therapist evidences change in specific domains. Likewise, if an administrator is trying to direct clients with particular problems to therapists well suited to address them, a general score may not be sufficient. Thus, we suggest that both general measures and multidimensional measures of psychological symptoms are likely to have appropriate applications. The latter would add the ability to examine domain-specific effectiveness, as mentioned in the TOP examples above.

Administrators must appreciate that change ought not be expected to appear similar across cases. As is the case in research, while aggregating across large amounts of data points may help detect general trends in the data, this may obscure meaningful differences between data points. That is, although it is definitely ideal to have a large sample of clients and therapists, the results of this sample may be misleading if some of the clients and therapists are not like the others. Methods for differentiating between "types" of clients and therapists—and subsequently modifying expectations for change—could be instrumental in improving the specificity of these analyses, and multidimensional measures offer one avenue for this. Until brief self-report measures can better predict courses of change for individual clients, including problems and negative effects of treatment, we suggest that administrators approach these measures with care. These predictions are likely to improve if careful attention is paid to the types of errors that may result. Given the likelihood of errors in inference, analyses based only on the brief instruments used in TOM should not be the sole criterion in determining the success or failure of a course of psychotherapy, or the effectiveness of a particular treatment or therapist.

Brief measures of psychological symptoms are here to stay. When used for the purpose of TOM, they have the potential to augment psychotherapy and administration in mental health from an important perspective, based on empirically sound principles and careful hypothesis testing. Their promise as useful tools for research is already well-established. As the field develops, remaining mindful of the limitations of these measures will remain vital for clinicians, administrators, and researchers alike. We hope this discussion and the examples provided will offer some indication of these limitations, and of the work that remains for future research.

Résumé

Au cours des dernières années, des progrès importants ont été réalisés dans le suivi des résultats des traitements en psychothérapie. De nombreux instruments ont été créés pour aider les psychothérapeutes à évaluer l'effet de leurs services, et plusieurs outils cliniques ont été créés pour améliorer directement la qualité des services offerts. Au sein du domaine, on commence à percevoir ces systèmes de suivi des résultats et de rétroaction avec une confiance accrue à mesure que les preuves de leur efficacité se multiplient. L'article examine les types d'erreurs qui peuvent survenir lorsqu'on tire des conclusions au moyen des résultats des suivis, en particulier pour déterminer s'il y a eu ou non des changements dans les sympt]]&#x12D973;Ce sujet est examiné comme le serait toute autre question empirique, en ayant recours à la traditionnelle vérification d'une hypothèse pour décrire deux types d'erreur dans la prise de décisions. Sont examinées en particulier les forces et la vulnérabilité de deux stratégies d'évaluation courantes en ce qui a trait au suivi des résultats des traitements (mesure générale et mesure multidimensionnelle) en vue d'éliminer les erreurs d'inférence et de maximiser l'efficacité des démarches de surveillance des résultats. Ensuite sont présentés quelques exemples de nouveaux usages des mesures multidimensionnelles pour réduire les erreurs de type II en vue d'améliorer les résultats.

Mots-clés : surveillance des résultats thérapeutiques, psychothérapie, mesure, Treatment Outcome Package (TOP).

References

Berk, M. L., & Monheit, A. C. (2001). The concentration of health care expenditures, revisited. *Health Affairs, 20,* 9–18. doi:10.1377/hlthaff.20.2.9

Boswell, F., Kraus, D. R., Nordberg, S., S., & Castonguay, L. G., (2009, June). The Treatment Outcome Package (TOP): An investigation of its validity. Poster presented at the 39th annual meeting of the Society for Psychotherapy Research, Santiago, Chile.

Castonguay, L. G. (2011). Psychotherapy, psychopathology, research and practice: Pathways of connections and integration. *Psychotherapy Research, 21,* 125–140. doi:10.1080/10503307.2011.563250

Castonguay, L. G., & Oltmanns, T. G. (in press). Psychopathology research and clinical interventions: Broad conclusions and general recommendations. To appear in L. G. Castonguay & T. Oltmanns (Eds.), *Psychopathology: Bridging the gap between basic empirical findings and clinical practice.* New York, NY: Guilford Press.

De Jong, K., van Sluis, P., Nugter, M. A., Heiser, W. J., & Spinhoven, P. (in press). Understanding the differential impact of outcome monitoring: Therapist variables that moderate feedback effects in a randomised clinical trial. *Psychotherapy Research.*

Derogatis, L. R. (1994). *SCL-90-R: Administration scoring and procedures manual.* Minneapolis, MN: National Computer Systems.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13,* 595–605. doi:10.1017/S0033291700048017

Eisen, S. V., Normand, S. L. T., Belanger, A., Spiro, A., & Esch, D. (2004). The revised Behavior and Symptom Identification Scale (BASIS-24): Reliability and validity. *Medical Care, 42,* 1230–1241. doi:10.1097/00005650-200412000-00010

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 180,* 51–60. doi:10.1192/bjp.180.1.51

Fitzgibbons, T. (2006). *NM3 Outcomes ROI.* Unpublished Blue Cross analysis.

Foa, E. B., & Meadows, E. A. (1997). Psychosocial treatments for posttraumatic stress disorder: A critical review. *Annual Review of Psychology, 48,* 449–480. doi:10.1146/annurev.psych.48.1.449

Frank, J. (1961). Persuasion and healing: A comparative study of psychotherapy. Baltimore, MD: The Johns Hopkins University Press.

Goodman, W. K., Price, L. H., Ramussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., & Charney, D. S. (1989a). The Yale–Brown Obsessive Compulsive Scale: II. Validity. *Archives of General Psychiatry, 46,* 1012–1016. doi:10.1001/archpsyc.1989.01810110054008

Goodman, W. K., Price, L. H., Ramussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., . . . Charney, D. S. (1989b). The Yale–Brown Obsessive–Compulsive Scale: I. Development, use, and reliability. *Archives of General Psychiatry, 46,* 1006–1011. doi:10.1001/archpsyc.1989.01810110048007

Hill, C. E., & Lambert, M. J. (2004). Assessing psychotherapy outcomes and processes. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of psychotherapy and behavior change* (5th ed., pp. 84–135). New York, NY: John Wiley.

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. doi:10.1037/0022-006X.61.4.678

Howard, K. I., Moras, K., Brill, B. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy. Efficacy, effectiveness and patient progress. *American Psychologist, 51,* 1059–1064. doi:10.1037/0003-066X.51.10.1059

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. doi:10.1037/0022-006X.59.1.12

Kelly, K., O'Donnell, R., Pelletier, A., & Simmons J. (2008). *Behavioral Health Outcomes Program update*. Unpublished Blue Cross report.

Kopta, S. M., & Lowry, J. L. (2002). Psychometric evaluation of the Behavioral Health Questionnaire-20: A brief instrument for assessing global mental health and the three phases of psychotherapy outcome. *Psychotherapy Research, 12,* 413–426. doi:10.1093/ptr/12.4.413

Kraus, D. R., Castonguay, L. G., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21,* 267–276. doi:10.1080/10503307.2011.563249

Kraus, D. R., Seligman, D., & 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. doi:10.1002/jclp.20084

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. doi:10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

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. doi:10.1037/0022-006X.69.2.159

Lambert, M. J., Whipple, J. I., Smart, D. W., Vermeersch, D. A., Niesen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11,* 49–68. doi:10.1080/713663852

LeMoult, J., Castonguay, L. G., Joorman, J., & McAleavey, A. A. (in press). Depression: Basic research and clinical implications. To appear in L. G. Castonguay & T. Oltmanns (Eds.), *Psychopathology: Bridging the gap between basic empirical findings and clinical practice.* New York, NY: Guilford Press.

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., . . . Hayes, J. A. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58,* 97–109. doi:10.1037/a0021282

Lutz, W., Böhnke, J. R., & Köck, K. (2011). Lending an ear to feedback systems: Evaluation of recovery and non-response in psychotherapy in a German outpatient setting. *Community Mental Health Journal, 47,* 311–317. doi:10.1007/s10597-010-9307-3

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure, and duration—What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy, 13,* 223–232. doi:10.1002/cpp.496

Neighborhood Health Plan. (2008). *Performance objective 5B, alert score analysis*. Internal report.

Nemeroff, C. B. C. (2006). Posttraumatic stress disorder: A state-of-the-science review. *Journal of Psychiatric Research, 40,* 1–21. doi:10.1016/j.jpsychires.2005.07.005

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement: Interdisciplinary Research and Perspectives, 5,* 217–235.

Nordberg, S., Boswell, J., Kraus, D., Castonguay, L. G., Hayes, J., & Wampold, B. (2010, June). Therapist effectiveness treating depression with and without co-morbid substance abuse. Paper presented at the 41st annual meeting of the Society for Psychotherapy Research, Asilomar, CA.

Pihl, R. O., & Stewart, S. H. (in press). Substance use disorders: Nature, etiology, and treatment. To appear in L. G. Castonguay & T. Oltmanns (Eds.), *Psychopathology: Bridging the gap between basic empirical findings and clinical practice.* New York, NY: Guilford Press.

Sheehan, D. V. (2002). Establishing the real cost of depression. *Managed Care, 11,* 7–10.

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. doi:10.1037/a0019247

Stelk, W., & M. Berger (2009, June). Predictive modeling: Using TO clinical domain items to identify adult Medicaid recipients at risk for high utilisation of behavioral health services in a managed care provider network. Paper presented at the 40th annual meeting of the Society for Psychotherapy Research, Santiago, Chile, 2009.

Teachman, B. A., Goldfried, M. G., & Clerkin, E. M. (in press). Panic and phobia. To appear in L. G. Castonguay & T. Oltmanns (Eds.), *Psychopathology: Bridging the gap between basic empirical findings and clinical practice.* New York, NY: Guilford Press.

Wagenmakers, E-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

Wells, K. B., Golding, J. M., & Burnam, M. A. (1988). Psychiatric disorders in a sample of the general population with and without chronic medical conditions. *The American Journal of Psychiatry, 145,* 976–981.

Youn, S. J., Kraus, D., & Castonguay, L. G. (in press). *The Treatment Outcome Package: Facilitating practice and clinically relevant research. Psychotherapy.*

Received January 10, 2012
Revision received February 24, 2012
Accepted February 27, 2012 ■