

**Measures that Make a Difference: A Functional Contextualistic Approach to
Optimizing Psychological Measurement in Clinical Research and Practice**

Joseph Ciarrochi, Robert D. Zettle, Robert Brockman, James Duguid, Philip Parker,

Baljinder Sahdra, and Todd B. Kashdan

Please cite as: Ciarrochi, J., Brockman, R., Duguid, J., Parker, P., Sahdra, B., & Kashdan, T. (2016). Measures that make a difference: Optimizing psychological measurement to promote well-being and reduce suffering. In Zettle, R., Hayes, S., Biglan, T., and Barnes-Holmes, D. (Eds.) *Handbook of Contextual Behavioral Science* (pp. 320-347). John Wiley and Sons, Malden, USA.

The basic researcher and the practitioner seem perpetually at odds. The basic researcher spends years building a beautiful model of the world. The practitioner takes one look at the model and says, “How will that help me to improve the lives of the clients I will work with this week?” The basic researcher often seems stumped for a legitimate answer.

This chapter seeks to help bridge the gap between basic theory and practice, between obscure psychometric concepts and the concrete needs of the therapist sitting in the room with a client. This chapter is comprised of five sections. Because some of the gap we seek to bridge may be attributable to differing philosophical approaches to psychological measurement, section 1 will examine some fundamental differences in how researchers and practitioners typically think about and use clinical assessment. Section 2 focuses on the different ways that a particular measure can be useful to a practitioner, and on various research designs that can be used to assess the utility of clinical assessment. Section 3 then takes a closer look at classic psychometric theory and how it can aid the practical goals of measuring therapeutic change. It will also examine longitudinal research as a means to understanding processes of change and as a compliment to the treatment utility designs reviewed in section 2. Section 4 will seek to organize the bewildering number of clinical process and outcome measures into a simple behavioral framework. Finally, section 5 will discuss promising new directions in contextual behavioral measurement.

Section 1: Philosophical Approaches to Psychological Assessment

The quality of a psychological measure is typically judged by the degree to which it meets the psychometric standards of reliability and validity (American Educational Research Association, 1999). Broadly speaking, reliability is concerned with the degree to which a measure yields consistent assessment data, and is a

necessary, but insufficient condition in determining a measure's validity, or how well it assesses what it purports to (Anastasi & Urbina, 1997). Often overlooked is that psychometric theory is at least implicitly, if not explicitly, based on an elemental realistic perspective of the psychological world that in our view can be antithetical to functional contextualism (Nelson & Hayes, 1986).

Elemental Realism

Elemental realism assumes that one can know the true nature of reality, and objectively discover the elements of which it is composed. Elemental realists view the human psyche as a complex machine, and their purpose is to accurately identify the parts and understand what they do. A key question for elemental realists is “What psychological characteristics are we made up of, how do those characteristics interrelate and influence each other, and how do they link to behavior”? Success is defined by how well a construct is able to predict and help establish meaningful, reliable causal patterns.

The elemental realist assumes that building a good working model of the universe will lead to an ability to change how it works. This is not necessarily true. Consider the example of self-esteem. Research has shown that high self-esteem is likely to precede the development of positive social support networks (Marshal, Parker, Ciarrochi, & Heaven, 2013). Does this suggest that the temporal relationship between self-esteem and the development of social support is causal in nature? Even if it is interpreted in this manner, what is the optimum way to increase self-esteem? Do we seek to undermine negative self-concepts via cognitive disputation? Or is it better to accomplish the same goal via the process of defusion as practiced in acceptance and commitment therapy (Hayes, Strosahl, & Wilson, 2012)? Knowing

that self-esteem predicts social support does not tell us how we might best change self-esteem.

Functional Contextualism

In contrast to elemental realism, functional contextualism assumes we can never know the true nature of reality or the elements that comprise it. Thus, measures are not assumed to reflect hidden “things”. Rather, measures are “behavioral samples” in a particular context. “Context” means whatever comes before the behavior (antecedents) and whatever follows it (consequences). For example, responses on an extraversion scale are not assumed to be caused by an underlying trait. Rather they are viewed as verbal behavior that, like all behavior, is influenced by the context in which it occurs. For example, the same item might be answered differently on an anonymous personality inventory administered during an online research project than when used for employee selection.

Functional contextualists do not view measurement development as its ultimate goal. They view measurement as a tool that can help them improve the human condition. The functional contextualist *will* divide measures into parts and factors, but purely for pragmatic purposes (i.e., does the division help us achieve our goals?). The functional contextualist would make no assumption that this “division” uncovers or reveals something of the “true nature” of character or personality; it is nothing more or less than a useful strategy for achieving a specific goal in a particular context.

The goal of functional analysis is to find ways to predict-and-influence behavior. Prediction in itself is not enough. Typical research in this tradition focuses on manipulating antecedents and consequences and observing how behavior changes as a result. A particular activity is “successful” if it helps to achieve stated goals.

Truth and Error

The distinction between the two philosophical approaches to psychological assessment is illustrated clearly in the notion of measurement error. According to classic psychometric theory, a score on a measure is comprised of a “true score” plus measurement error (Nunnally, 1967). This formulation follows a clear correspondence-based truth criterion. Differing “true scores” hypothetically exist for respondents and their obtained or observed scores on an inventory would exactly *correspond* to these true scores were it not for error in measurement. By contrast, from a contextualistic perspective, interindividual variability in psychological assessment data is not attributed to differences in true scores and intraindividual variability across settings and time is not assumed to result from measurement error. Scores on a psychological inventory are a measure of behavior and “error” has no place in a contextual approach to understanding behavior (Hayes, Nelson, & Jarrett, 1986). Instead, behavior that varies across individuals, as well as across time and place within individuals, is something to be explained by an analysis of contextual factors that include differing learning histories and current situational variables (e.g., the behavior of the assessor, demand characteristics, recent life events, etc.).

One way to minimize this seeming conflict between elemental realism and contextualism is to use terms other than “true score” and “error” in framing variability in assessment data. Rather, we can talk about sources or types of behavioral variability. Each item in a scale can be seen as being associated with “common variance” plus “unique” or unexplained variance (Kline, 2010). Common variance is what the item shares with every other item on the scale. For example, in an assertiveness scale, the common variance of an item can be thought of as the extent that the item refers to assertive behavior and not to some other type of behavior. In

contrast, unique variance is assumed, from the elemental realist perspective, to consist of both random error and specific variance. The functional contextualist would not assume random error, but could assume specific variance, which refers to the variability in the scale that is due to a type of behavior that is different from the “common” behavior measured by the scale items. For example, a measure of assertiveness may contain the following item: “I often fail to assert myself and I ruminate about my failure.” Let’s assume that this item reflects two “behaviors,” assertion and rumination. If this item was part of an assertiveness scale, then the rumination component would contribute to what might be termed “error,” but it is clearly more usefully conceptualized as “unexplained variance.” Almost all modern measures reflect variance related to the intended target of the measure (common variance) and variance due to specific factors (unique variance) (Marsh et al., 2009).

There is an important reason for the functional contextualist to be interested in ideas like “common variance” and “unique variance.” They suggest that a scale that is assumed to be unidimensional may actually be multidimensional, or indicative of multiple kinds of behavior. Interventions that influence behavior associated with “common variance” may not be the same as those that impact behavior associated with “unique variance.” Using the example from above, we may want to use different types of intervention to increase adaptive assertiveness and to reduce unhelpful rumination. If a client’s response to scale items reflects both of these kinds of behaviors, we practitioners want to be made aware of this, so we don’t merely average scores and assume we are talking about one kind of behavior, assertiveness, rather than two (i.e., both assertiveness and rumination).

Conceptual Flexibility

When people are unaware of philosophical assumptions, they often engage in unnecessary dispute. For example, the elemental realistic scientist may accuse the functional contextualistic therapist of not creating a model where different internal, psychological parts are hypothesized to influence each other. The therapist, in turn, accuses the basic researcher of focusing on obscure theoretical models and failing to provide guidance to practitioners. This is a dispute that can never be won, because worldviews are based on assumptions, not evidence. Acknowledging one's philosophical worldview simply means owning up to one's improvable assumptions.

The philosophical gap that sometimes divides basic researchers and mental health professionals is narrowed if both parties temporarily shift perspectives. The elemental realistic who carefully constructs and tests an information-processing model of anxiety, also seeks to identify interventions that reduce anxiety. The functional contextualistic therapist who wants to alleviate client suffering right now, also has an interest in understanding how different processes interrelate to contribute to psychological pain. The larger philosophical dispute can thus be ended by saying that nobody *is* an elemental realist or a functional contextualist. These are just philosophical glasses we sometimes wear, perhaps without even knowing it.

We have reservations about the exclusive use of reliability and validity in evaluating the quality of psychological assessment, but recognize that an outright rejection of psychometric standards would reflect an inflexible and dogmatic stance that is itself inconsistent with functional contextualism. In fact, we will argue in this chapter that the relevance of psychometric standards in selecting clinically-appropriate measures is contextually-dependent and depends on purpose. Sometimes it may be useful to rely primarily on psychometric standards, and sometimes it may be useful to ignore these standards in favor of others.

In our view, all measures, both within as well as outside of psychology, are simply tools whose usefulness and related critical attributes vary depending on the context in which, and the purpose for which, they are being used. Often the most useful measures may be those that are the most psychometrically sound. However, neither reliability nor validity, as we will argue, may be necessary for a given psychological measure to demonstrate treatment utility.

Section 2: Treatment Utility of Clinical Assessment

Treatment utility refers to “the degree to which assessment is shown to contribute to beneficial treatment outcome” (Hayes, Nelson, & Jarrett, 1987, p. 963), where “assessment” can be broadly defined. Although designs in treatment utility research can be either correlational or experimental in nature, here we selectively emphasize those that evaluate the impact of manipulating assessment practices on treatment outcome. Interested readers are encouraged to consult Hayes et al. (1986, 1987) and Nelson-Gray (2003) for more comprehensive discussions of treatment utility research designs.

While reliability and validity are indispensable psychometric standards, treatment utility is in our view the sine qua non of a functional contextualistic approach to evaluating, selecting, and utilizing clinically-relevant measures. Asking whether certain assessment practices enhance desired outcomes is not unique to psychotherapy. Recent developments within medical research and practice provide clear illustrations of treatment validity that might be generalized to psychological research. We thus present treatment utility examples from both medical and psychological domains.

Manipulated Assessment Design

Mammograms are performed to identify cancerous tumors that would otherwise go undetected. If such tumors can be identified early and successfully treated, survival rates should improve. However, a recent Canadian experiment utilizing what is commonly known as a manipulated assessment design (Nelson-Gray, 2003) found no evidence for the treatment utility of mammography among women aged 40-59 (Miller et al., 2014). The mortality rates from breast cancer among women from this age range who were randomly assigned to receive routine annual mammography screenings were no better than those for women in the control condition. It is important to note that these findings do not call into question the reliability and validity of mammograms, nor do they suggest that mammography is ineffective in reducing deaths from breast cancer among women of younger or older ages. Further research is obviously necessary to determine if age functions as a moderating variable, such that mammograms display treatment utility with one age range of women, but not others.

Somewhat similar to the use of mammography, it is common clinical practice for mental health professionals to screen for psychiatric conditions such as psychotic disorders, substance abuse, and/or suicidality by administering assessment batteries to all new clients. While doing so certainly has the appearance of “good clinical practice,” we are aware of no empirical support for such screenings. For example, for all we know, clinics that at intake administer the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, 2006) and/or the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) provide treatment that is no more efficacious than facilities that routinely do not. Note that our skepticism here is not about the psychometric properties of the MCMI-III or the MMPI-2, but centers around their treatment utility;

i.e., does their use lead to better treatment outcomes? Those who argue for the use of psychiatric screenings have the burden of providing evidence that such practices are efficacious. Otherwise, both clients and therapists are wasting considerable time, expense, and effort.

Manipulated Use of Assessment Design

Continuing with our medical example, if it is determined that treatment for breast cancer is advisable, the oncologist is then faced with a treatment selection question. That is, which treatment choice or combination of options (e.g., radiation, chemotherapy, mastectomy, etc.) is likely to be most effective for each given patient? Recent oncological research increasingly suggests that genomic testing of tumors may make it possible to personalize cancer treatment by matching differing therapeutic options to differing types of tumors (Andre et al, 2014). Evaluating the benefit of assessment on treatment selection has been referred to as a manipulated use of assessment design (Nelson-Gray, 2003). However, we prefer the more descriptive term “manipulated match” for this design, as it highlights the independent variable.

Practicing clinicians face difficult practical questions. How does one intervene with clients that have specific presenting problems, such as certain personality disorders, for which there are no empirically-supported treatments like those recognized by Division 12 of the American Psychological Association (aka Society of Clinical Psychology)? Another question arises, similar to that presented to the oncologist, when the practitioner has the “luxury” of selecting among several empirically-supported therapeutic options for a presenting problem. Are certain kinds of interventions a more optimal “match” for particular types of clients? For example, is there any psychological assessment that can guide practitioners in most effectively matching cognitive therapy, interpersonal therapy, behavioral activation, and

acceptance and commitment therapy to their different clients? Are some clients better candidates for one of these therapies than the rest, whereas others would respond optimally to another choice, and if so, what are the psychological variables that distinguish one subgroup from another?

A similar question can be asked concerning the ingredients of therapy. Are there some components of an intervention that should be highlighted with some clients, but not others? For example, values clarification might be ideally suited to those who have unclear direction in their lives, whereas emotional awareness training may suit those who are alexithymic, or have trouble identifying and describing feelings. In short, is there an assessment that can be conducted that is similar to determining the DNA sequencing of cancerous cells?

Treatment Determination and Guidance

Specific questions about treatment determination and treatment guidance are centrally related to the broader issue of treatment utility. Before further discussion, it may be useful to first clarify what we mean by each of these terms. “Treatment determination” refers to the process of ascertaining what types of presenting problems and issues are to be addressed in therapy. Once a determination has been made to target a particular concern, ‘treatment guidance’ refers to the role of assessment in adjusting treatment of it on a client-by-client basis.

Treatment determination. Often times, one or more presenting problems may be readily apparent based on a client’s stated reason for seeking treatment, the referral source, the context surrounding the initiation of therapy, and so on. Nonetheless, administering an intake assessment battery seems reasonable given high rates of psychiatric comorbidity (Regier et al., 1990; Sartorius, Ustun, Lecruiber, & Wittchen, 1996) and evidence that half of clients keep secrets from their therapists

(Farber, Berano, & Capobianco. 2004). The costs of such assessment would presumably be outweighed by the benefits of detecting salient issues and concerns (e.g., dangerousness to self or others, possible substance abuse, sexual matters, etc.) that might complicate treatment of the more obvious presenting problems. In effect, an intake assessment can help make a determination that unidentified psychological issues warrant treatment in their own right. The relative benefits of psychiatric screenings would ostensibly be even greater with clients who have vague and ambiguous presenting problems.

Despite the reasonableness of the above arguments, we are unaware of any treatment utility research that has examined the cost-benefit ratio of using screening batteries. In order to do so, a manipulated assessment experiment could be conducted in which half of the clients at an outpatient mental health clinic are randomly assigned to complete a comprehensive assessment battery. The treatment utility of the assessment could be determined by comparing client responsiveness to therapy. If those clients who were administered the screening battery improve more than their control group counterparts, its treatment utility is empirically supported.

It is important to note that screening batteries are just one example of what might be evaluated by a manipulated assessment design. Recall that this design asks whether the inclusion of some form of assessment for identifying focal problems and issues in therapy results in improved treatment outcomes. The independent variable in such experiments could be as broad as an extensive psychiatric screening or as narrow as the administration of a brief questionnaire (e.g., the Hopelessness Scale of Beck and colleagues, 1974).

Treatment guidance. Treatment determination is seldom an issue in providing psychological services to those with pervasive developmental disorders.

This is particularly the case, for example, in targeting self-stimulatory and/or self-injurious behaviors displayed by autistic children. While the need for intervention and its goals are thus readily apparent (e.g., to significantly reduce, if not eliminate head banging), how to most efficaciously guide treatment is not. For this reason, we will use clinical work with children displaying developmental disorders to illustrate the utility of assessment in guiding individualized treatment programs.

Functional analysis represents the oldest and most quintessential contextualistic approach to assessing behavior (Ferster, 1965). Treatment determination is met by identifying a target behavior in need of modification, followed by the systematic manipulation of antecedent and consequential events that are suspected of functioning as its controlling variables (Iwata, Dorsey, Silfer, Bauman, & Richman, 1982). For example, hand biting occasioned by the presentation of a nonpreferred task might serve an avoidant function. To evaluate this possibility, changes in the frequency of the target behavior could be documented in response to the systematic introduction and withdrawal of the task. The findings of such mini-experiments are then used to guide the development of treatment plans, such as teaching the child a more appropriate way of requesting a break. A complete functional analysis would also “test” for other controlling variables, such as receipt of attention, sensory reinforcement (i.e., hand biting is maintained by its sensory consequences), and tangible reinforcement (e.g., the child stops biting his hand when offered a favorite toy). Insofar as the same topographical behavior (e.g., hand biting) can serve different functions for different children, and different topographical behaviors (e.g., hand banging and hand biting) can serve the same function (e.g., attention-getting) both within and across children, functional analyses must be conducted on a case-by-case basis. The data from the functional analysis thereby

guide a “customized” treatment program for each child, even among those displaying the same target behavior.

Direct observations of target behaviors in which correlated antecedent and consequential events are systematically tracked (but not experimentally manipulated as during a functional analysis), and detailed interviewing of caretakers and teachers of the autistic child about their observations, provide two less costly and more efficient strategies for identifying controlling variables (O’Neill, Horner, Albin, Story, & Sprague, 1990). Unlike the use of screening batteries with psychiatric outpatients, there have been some efforts to evaluate the utility of functional analyses in designing and guiding treatment programs for autistic children and others with pervasive development disorders. The treatment utility of functional analyses appears to be moderated by the severity of behavioral problems being assessed. While programs to modify severe target behaviors based on functional analyses have been shown to be more efficacious than those guided by other means (Nelson-Gray, 2003), such as behavioral observations and assessment interviews, the same pattern does not seem to hold in providing services with children displaying milder developmental disabilities and behavioral problems. For example, English and Anderson (2006) reported that interventions based on descriptive observations of target behaviors as detailed by O’Neill et al. (1990) for three children with developmental delays were more effective than those derived from functional analyses. The children displayed mild intellectual disabilities and appeared to be less severely impaired than participants in other projects investigating the treatment utility of functional analyses (e.g., Repp, Felce, & Barton, 1988).

The evidence base for the treatment utility of functional analyses is sobering. After all, if the one assessment strategy that is widely recognized as the “poster child”

for contextualized approaches to psychological assessment has limited treatment utility, how likely is it that the use of other forms of assessment enhance treatment outcome? Our response is that this is an empirical question that can only be addressed by more manipulated assessment experiments and not by generalizing what is known to date about the treatment utility of functional analyses to other measures. To underscore our point, consider the possible treatment utility of projective techniques.

The reliability and validity of projective techniques, when used in unstructured ways, has long been recognized as being weak (Eysenck, 1959; Jensen, 1965). This has led some to argue that projective techniques should not be regarded and evaluated as psychological tests, but as clinical tools (Anastasi & Urbina, 1997). This point of view is similar to, but less expansive than our own that all forms of psychological assessment, including standardized tests, are usefully seen as tools; and also echoes that articulated by proponents of projective techniques. For example, defenders of the Rorschach typically have claimed that its clinical utility outweighs its poor psychometric properties (Meyer, 1999). While some have argued that there is sufficient empirical support for the clinical utility of the Rorschach (Viglione, 1999), others have pointed out that there is an absence of research that has addressed this matter, particularly when applying a strong definition of “clinical utility” that is indistinguishable from treatment utility (Hunsley & Bailey, 1999).

Our “take home message” at this point is that the treatment utility of the Rorschach, despite claims to the contrary, remains unknown. But with relatively few exceptions (e.g., functional analyses) and in the absence of systematic treatment utility research, the same must be regrettably said both now and in the future about clinical assessment more broadly. With regard to the Rorschach, a manipulated

assessment experiment could be conducted where practitioners who routinely use the Rorschach only administer it to a randomly selected subset of their clients. Do those clients who are administered the test show better therapeutic outcomes than the control group clients? If they do, preliminary support would be provided for the treatment utility of the Rorschach, although follow-up experimental research would be necessary to fully understand the effect. Several explanations are plausible, including a placebo, self-fulfilling prophecy type effect related to clinicians' belief that having a Rorschach protocol enables them to be more effective therapists. Would they be just as effective if given bogus protocols? Another possibility is that having access to a completed protocol is not the critical variable; rather the critical variable is the testing itself, during which therapists pick up on more subtle client behavioral cues. Before moving on, we again would remind our readers that the same questions and issues raised about projective techniques also apply to other therapeutic approaches and related forms of assessment. For example, does the use of the Valued Life Questionnaire (VLQ; Wilson, Sandoz, Kitchens, & Roberts, 2010) in guiding committed action homework assignments within acceptance and commitment therapy (ACT; Hayes, Stroschal, & Wilson, 2012) improve treatment outcomes?

Treatment Selection

The term "treatment selection" refers to the process of choosing a preferred therapeutic option from an array of two or more alternatives. It is preceded by treatment determination (e.g., deciding to treat depression) and can provide a therapeutic context within which relevant issues of treatment guidance can also emerge (e.g., use of the VLQ to guide behavioral activation as the treatment selected to treat depression). Treatment selection can be made on the basis of conceptual reasons (e.g., a new approach is consistent with an agency or therapist's theoretical

orientation), practical considerations (e.g., the therapist is more familiar with one approach than other options that enjoy the same level of empirical support), or relevant research.

The type of research that we believe is most likely to increase the utility of assessment used for the purpose of treatment selection is that investigating moderating variables, or characteristics that influence the direction or magnitude of the relationships between the independent variable of psychological interventions and the dependent variable of therapeutic outcomes (Kazdin, 2007). Sometimes moderating variables are demographic in nature (e.g. gender, age, racial/ethnic status), such as suggested earlier in our discussion of genomic testing of cancerous tumors, and are of limited psychological relevance except to those who study gender and age-related differences. Other times, as indicated by our overview of the treatment utility of functional analyses, moderating variables (e.g., severity of behavioral problems) are of both psychological and pragmatic importance. Unfortunately, the assessment of moderating psychological variables can be considerably more challenging than evaluating those that are demographic in nature. Not only must the relevant variables themselves first be identified, but ways of assessing them that maximize treatment utility must also either be chosen or developed.

A correlational treatment utility design known as “post hoc identification of dimensions” (Hayes et al., 1986) is one method that can be used to identify psychologically-relevant moderating variables associated with therapeutic responsiveness. For example, Masuda et al. (2007) and Zettle (2003) reported that those who initially scored high on psychological inflexibility benefitted most from ACT, in terms of reduced stigmatizing attitudes and mathematics anxiety, respectively. By contrast, *low* inflexibility was associated with improvement in the mental health of

call center employees participating in a work reorganization intervention designed to enhance job control (Bond, Flaxman, & Bunce, 2008). More recently, Forman and colleagues (2013) found that susceptibility to eating cues, depression, and emotional eating positively moderated responsiveness to an acceptance-based behavioral treatment for obesity.

Unfortunately, the above findings may be of limited practical utility for at least two reasons. First, while the findings may point to a psychological profile of those who might be preferred candidates for a specific treatment, such as ACT, they don't necessarily directly point to comparably efficacious options for those who are not. To approach this objective, post hoc analyses are preferable within studies in which the therapeutic benefits of two or more interventions have been shown to be essentially equivalent. For example, within a large research program for treatment of depression (Sotsky et al., 1991), low social dysfunction predicted superior response to interpersonal therapy (Klerman, Weissman, Rounsaville, & Chevron, 1984), while low cognitive dysfunction was associated with greater responsiveness to cognitive therapy (Beck, Rush, Shaw, Emery, 1979). Second, while such findings suggest that treatment of depression might be optimized by assessing both social and cognitive functioning of clients and matching them to therapeutic options on that basis, a manipulated match design would be necessary to empirically evaluate this possibility.

Previous research by one of us (RDZ) provides an illustration of how correlational and experimental treatment utility designs can be combined in the manner described to address the question of whether outcomes can be enhanced by matching client moderating variables to therapeutic options. An initial study found that group and individual cognitive therapy of depression were equally efficacious (Zettle, Hafflich, & Reynolds, 1992). A post hoc analysis, however, indicated that

pretreatment scores on the Sociotropy-Autonomy Scale (SAS; Bieling, Beck, & Brown, 2000) were differentially associated with treatment outcome. Specifically, depressed participants high in sociotropy, reflective of relying on social relationships for gratification and support, responded better to group than individual therapy, while their counterparts high in autonomy, or a tendency to derive gratification from personal accomplishments, showed an opposite pattern. Based on these findings, a subsequent manipulated match experiment was conducted in which depressed participants were either matched or mismatched to group versus individual cognitive therapy based on their pretreatment SAS scores (Zettle & Herring, 1995). As expected, a higher proportion of matched participants displayed marked improvement at follow-up, thereby providing empirical support for the treatment utility of the SAS in matching depressed clients to the format of cognitive therapy.

Similar to the earlier-cited example of genomic testing of cancer cells, knowledge of psychologically-relevant variables that moderate psychotherapeutic outcomes also may enable mental health professionals to offer a “personalized approach” with their clients. By assessing such moderating variables, treatment options may be selected that are better matched to individual clients.

Section 3: Treatment Outcome and Process

Psychometrics is one of those areas that appears irrelevant to a practitioner. And it often is. Yet if you are a practitioner, you are stuck with psychometrics, because if you want to get better at improving people’s lives, you need to be able to answer two measurement questions, one concerning outcome and the other involving process.

First, how would we know if we were creating positive outcomes and improving someone’s life? What does that even mean? Is “improvement” measured in

reductions in anxiety, depression, and self-harm? Does it mean increases in positive indices, such as vitality and resilience? Does it mean activation of value-congruent behavior, regardless of the emotional consequences? Second, how do we know what process, in particular, leads to improvement? That is, what are we doing with our clients that actually produces change? If we cannot measure outcomes and intervention processes hypothesized to lead to those outcomes, it becomes difficult if not impossible for practitioners to become better at their craft.

Outcome Measure Selection

As suggested earlier, psychometric considerations are important in selecting an appropriate measure for assessing treatment outcome.

Reliability. Psychological measures are typically evaluated for their internal as well as temporal consistency. There are several ways of evaluating whether therapeutic change can be considered clinically significant. Test-retest reliability is of particular importance if the reliable change index (RCI) of Jacobson and Truax (1991) is used to do so. Any pre to posttreatment improvement in an outcome measure should exceed what would be expected by chance given its test-retest reliability. Measures that display higher levels of test-retest reliability accordingly require less of an absolute change for such improvement to be regarded as clinically significant. In selecting an outcome measure, it is important that test-retest reliability is assessed over a span of time comparable to the length of many treatment protocols (e.g., 12 weekly sessions) in order for that reliability estimate to be of most use to the practitioner.

Validity. A significant RCI suggests that the degree of therapeutic change is meaningful and unlikely to be an artifact of the temporal instability of the particular measure used to assess it. However, such a finding does not address the question of

what exactly has changed as a result of the intervention. Discriminant validity (Campbell & Fiske, 1957) may be of particular importance in helping the practitioner identify specific ingredients of change.

Recall our earlier discussion of common versus unique variance in an assertiveness inventory containing items that assess both assertiveness and rumination. The inventory might fail to demonstrate sufficiently discriminant validity by being correlated too highly with a questionnaire such as the Ruminative Response Scale (Nolen-Hoeksema, Morrow, & Fredrickson, 1993). Such a finding should be of concern to both the basic researcher and the clinician. For the basic researcher, the construct validity of the assertiveness inventory comes into question. For the practitioner, the concern is more practical. If the goal of therapy was to increase assertiveness, does reliable change in the assertiveness inventory mean that this goal was attained or does this apparent improvement instead reflect reduced rumination? One strategy to minimize this issue is to develop measures that are unidimensional and are sensitive to change in the therapeutic target. Another strategy is to administer multiple outcome measures, with some being specific to the goal of therapy and others being of less relevance to that objective. For example, an ACT intervention for depression might assess for reductions in both depression and anxiety.

Process Measure Selection

Let's now take a more detailed look at how traditional psychometrics can aid the practical goal of understanding what process or mechanism might account for therapeutic change. If changes in a process measure, such as a defusion inventory, can be shown to mediate therapeutic improvement (e.g., lessened depression) by reliably preceding it, repeated administration of it provides useful feedback to practitioners. Reductions in defusion suggest that the therapy is having its intended

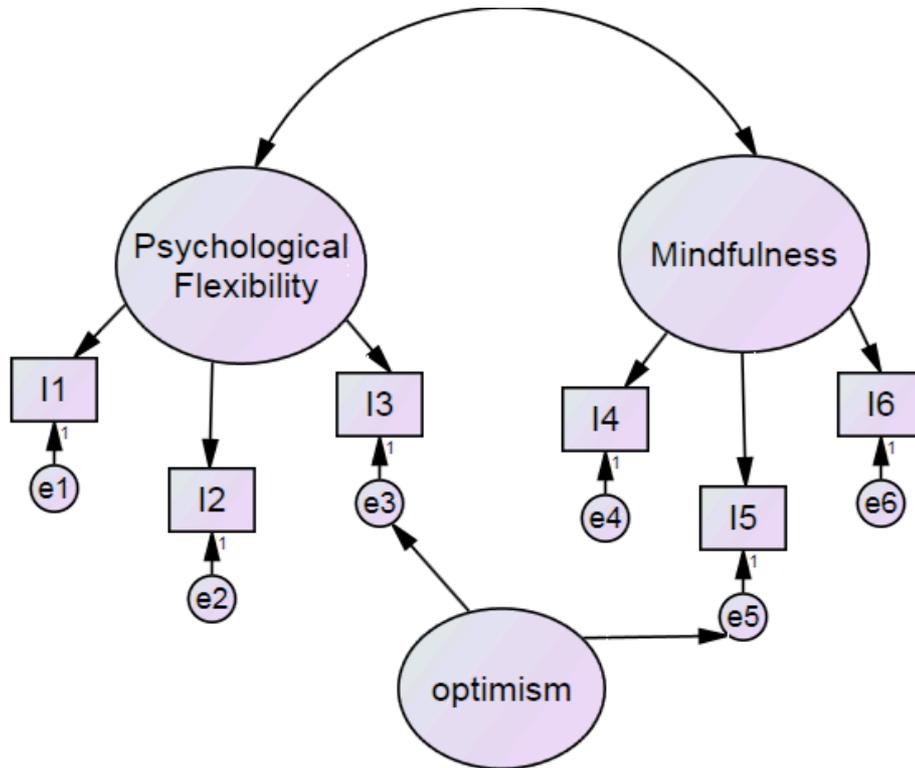
effect, even if similar reductions in depression are not yet apparent, and that it is advisable to “stay the course.”

Intervention packages such as ACT and CBT include many possible “ingredients.” Process research can help the practitioner hone in on which of these ingredients is most essential to positive outcomes. For example, ACT has shown that positive change occurs by promoting psychological flexibility and not experiential control (Dalrymple & Herbert, 2007). ACT also has been shown to target different processes than CBT (see Ciarrochi, Billich, & Godsell, 2010, for a review). CBT appears to promote self-confidence and not psychological flexibility, whereas the reverse is true for ACT (Lappalainen et al., 2007). Research on mindfulness-based cognitive therapy suggests that promoting self-compassion may be the key to therapeutic change (Kuyken et al., 2010).

Given the potential utility of process measures, it is worth taking a closer look at the psychometrics used to develop them. Figure 1 illustrates a traditional psychometric measurement model involving three possible process variables: (a) psychological flexibility, (b) mindfulness, and (c) optimism. Items 1 through 6 are observed indicators, or more simply items from the flexibility and mindfulness scales. According to psychometric theory, these indicators are assumed to be caused by underlying, correlated latent constructs, indicated by circles. Thus, psychological flexibility influences how people respond to items 1 through 3 and mindfulness influences items 4 through 6. In the below figure, items 3 and 5 are unintentionally linked to a third latent variable, optimism. This kind of crossloading is common (Marsh et al., 2009). Given that there are thousands of constructs, it is unlikely any test item a human can create will be an indicator of one and only one process variable.

For example, the first item from the measure of psychological flexibility ---“I am in control of my life”---is likely to reflect both flexibility and self-efficacy.

Figure 1. Hypothetical psychometric measurement model.



The quality of a measurement model from a psychometric perspective can be assessed via fit indices (Hooper, Coughlan, & Mullen, 2008). For example, imagine that optimism was not represented in the Figure 1 model. Then, the constraint would be that items 1 through 3 are caused only by psychological flexibility, and items 4 through 6 are caused only by mindfulness. This model is likely to fit poorly, because the correlation between the unique variance in item 3 (E3) and item 5 (E5) is constrained to be 0. There are many other constraints in Figure 1 that might be incorrect, such as the constraint that E1 does not correlate with E4. In general, the more parameters that are incorrectly constrained to be 0, the worse the model fit.

Measurement model evaluation allows us to assess the extent that items and factors (psychological flexibility and mindfulness) overlap. It allows the practitioner to ask questions like, “If the therapy I provide targets psychological flexibility, am I also likely to increase mindfulness?” If someone claims that an intervention such as ACT works by increasing psychological flexibility, then the measurement model in Figure 1 would suggest that an alternative interpretation needs to be explored. Perhaps the intervention works by increasing optimism. Thus, a clear measurement model, when combined with intervention outcomes, can guide the practitioner towards what is and is not being changed in a client.

When we look at assessment through the lens of functional contextualism we often respond negatively to the typical measurement model in Figure 1, because it seems to imply that there are latent, unchanging things that underpin responses. We assume that the responses to items on an inventory are just behavior under the influence of context, and therefore often can be altered, sometimes even rapidly, by changing controlling variables. For example, it may be possible for an intervention like ACT to alter the very structure of the measure. Imagine an inventory that has two items: (a) “I am able to persist at my goals, even when I feel distressed” and (b) “I am able to control my feelings.” Prior to therapy, these items are likely to be positively correlated and psychometrically might be said to reflect an underlying construct called “emotion regulation.” However, after the intervention, when people have been taught that emotional control is a problem, the two items may no longer correlate. Thus, the two-item emotion regulation scale would cease to be a reliable estimate of emotion regulation.

One basic tension between elemental realism and functional contextualism often comes down to this: the latter emphasizes that change is possible, while the

former often speaks in terms of stable constructs, e.g., the notion of “true scores” discussed in the first section this chapter. However, while many may argue for the stability of traits, there is nothing inherent in the elemental realism position that implies that change is impossible. For example, the common notion of testing for “measurement invariance” in a longitudinal design implies the possibility that measures may change their structure from year to year. The elemental realist recognizes that a particular measurement model may hold only in a particular context (time and place).

In summary, classic psychometric research does not automatically lead to treatment utility, but there is no reason why it can’t support treatment utility. Practitioners need to be able to assess outcomes and processes of change in order to be able to identify what is and is not working. If a measure does not meet traditional psychometric standards, or the measurement model is mis-specified, then the practitioner is unlikely to get precise, useful information from the measure. For example, if a measure of mindfulness lacks predictive validity, it is unlikely that changing responses to that measure will cause positive outcomes. If two therapies claim to be distinct, but the measures of their core process variables and purported mechanisms of change lack discriminant validity, then researchers within each therapy type will have a great deal of trouble showing distinctiveness. Finally, if a measure is heterogeneous and reflects many constructs, then it will be hard to interpret the meaning of scores, and if an intervention changes those scores, it will be hard to know exactly what changed. There is nothing so practical as reliable and valid measures in determining how and if a desired therapeutic goal has been realized.

Identifying processes of change in longitudinal research. For a process measure to mediate treatment outcome, three conditions need to be met. First, the

measure must be correlated with the outcome. Second, measures of suspected processes must be shown to temporally proceed, or be antecedent to the outcome. Third, one must rule out the possibility that a third variable accounts for the relationship between the process measure and the outcome.

It is commonly thought that experiments are always the best method for satisfying all three of these conditions. Random assignment supposedly solves the third variable problem, by distributing the unmeasured third variable equally between conditions. In section 1 of this chapter, we provided examples of strong experimental designs for evaluating treatment utility. However, experimental designs get somewhat more difficult when it comes to analyzing complex questions, and experiments don't always have the perfect answer (Bullock, Green, & Ha, 2010). Let's say that you show that an ACT intervention improves psychological flexibility, and psychological flexibility mediates the relationship between intervention and some positive outcomes such as emotional well-being. If we assume randomization worked perfectly, then it is reasonable to conclude that the intervention caused changes in the mediator and outcome. However, to have strong support for mediation, we must also show that the intervention affected only psychological flexibility and no other mediator (Bullock et al., 2010). In the above example, there could be an unmeasured mediator that explains the link between psychological flexibility and emotional well-being. For example, perhaps the intervention increased clients' sense of hope. If changes in hope correlate with changes in flexibility, then it may be hope is promoting outcome, rather than psychological flexibility.

We, therefore, do not see experiments as the only tool for addressing causal questions. Rather, we propose that longitudinal designs can complement the experimental designs. Because longitudinal research does not manipulate variables, it

is always subject to the criticism that a third variable could explain the link between the hypothesized antecedent and consequent variable. Controlling for confounds can reduce, but never completely eliminate, the third variable problem.

However, longitudinal designs have at least two strengths. First, one can examine a wide range of therapeutic process variables in a single study. For example, one could look at the effects of self-compassion, self-esteem, mindfulness, hope, and psychological flexibility on the positive transition from high school to adult life. It would be difficult for an experimental paradigm to simultaneously manipulate each of these five explanatory variables. The longitudinal design is excellent at narrowing down the possible list of process variables that are relevant to any outcome. Those variables that uniquely predict positive change are identified as candidates for further intervention research.

A second strength of a longitudinal design is that it can be used to detect heterogeneity in effects between participants. One difficulty with much experimental mediational research is that it assumes that links between intervention, mediator, and outcome are the same for all participants. If there is heterogeneity in links, then conclusions based on the average subject response can be misleading (McNulty & Fincham, 2011). Longitudinal designs with many repeated measures can be used to identify heterogeneity by estimating the positive effects of the mediator (say increased psychological flexibility) on the outcome within subject. Then we can set about identifying the contextual factors that predict benefit or lack of benefit for different types of clients, and we can potentially use this information to guide experimental, treatment utility research.

It is also possible to combine experimental and longitudinal designs by utilizing interventions that repeatedly manipulate the mediator and the outcome

(Bullock et al., 2010). For example, one could repeatedly manipulate daily mindfulness and examine daily positive affect. This approach would allow one to assess the effect of mindfulness interventions within subject, and thus identify those who fail to respond to the intervention and identify the contextual factors that predict such failure. Like all designs, this design has its assumptions, including assumptions that within-person effects don't change and that the effect of the mediator on the outcome has worn off before the next manipulation is administered (Bullock et al., 2010).

Ecological momentary assessment. Ecological momentary assessment (EMA) refers to a cluster of longitudinal data collection approaches that include experience sampling methods (ESM), ambulatory assessment, and diary methods. In an EMA study, data are collected from individuals in real world environments as they go about their daily life, resulting in greater *ecological* validity than other methodological approaches (Shiffman, Stone, & Hufford, 2008). The assessments are *momentary*, focusing either on the current state of an individual's experience, or very recent states, thus minimizing the impact of retrospective biases. There are several strengths of EMA for the contextual behavioral scientist.

Assessing variability of behavior. One index of flexible responding is variability. Does the client use the same response in every situation (inflexible) or do responses vary from situation to situation? EMA allows one to assess not just the mean of responses, but also the variability of response, a statistic that is impossible to calculate in a one-off trait measure (Kashdan, Ferrisizidis, Collins, & Muraven, 2010).

Linking changes in context to changes in outcomes. EMA can be used to measure antecedent events in people's lives, their behavioral reactions to those events (e.g., coping strategies), and the consequences of those behavioral reactions (e.g.,

levels of stress, mental health, affect, value-consistent behavior, etc.). See Delespaul and van Os (2002) for an illustration. . For example, EMA can be used to record the extent that people experience daily positive and negative social interactions, their coping response (e.g., reappraisal, mindfulness, suppression), and the consequence of those responses on their mental health. In many ways, this data collection mimics the data collected on the traditional thought record that is used by many therapeutic traditions as a means of assessing and building awareness of *in the moment* chains of experience.

EMA as a clinical intervention. A small but growing literature focuses on extending EMA methods to clinical intervention research. This literature can be broken down into examining the use of EMA for the purposes of *assessment* versus its use as an *intervention*. Using EMA as an assessment tool with clients engaged in psychotherapy can be useful for many of the same reasons that single-case methodologies are popular among clinicians. Such assessment methods can have the impact of increasing client motivation through consistent monitoring, and can be based on behaviors and experiences (processes) that are derived from a client's clinical formulation (Shiffman et al., 2008). EMA data allow for the analyses of the therapy process as it occurs, revealing possible opportunities for intervention , or for “taking stock” at given points so as to aid clinical decision making regarding the progress of therapy (Cohen et al., 2008).

The use of the EMA methodology as an *intervention* strategy has the potential to bring the therapy room more closely into the client's life. Recently, Heron and Smyth (2010) systematically reviewed 27 EMA intervention studies (EMIs) that had provided treatment for such problems as smoking cessation, weight loss, anxiety disorders, diabetes self-management, eating disorder symptoms, and general health

behaviors (e.g., exercise). The review found evidence that EMI can be effectively implemented across a range of clinical problems and health conditions, and that the method has high acceptability among clients. Of note, trials of cognitive behavioral therapy (CBT) with and without supplementary EMI for anxiety disorders and weight loss have found that in general, clients in CBT+EMI groups are able to achieve the same levels of treatment efficacy as standalone CBT with approximately half of the face-to-face sessions, and that EMI augmentation improved the overall efficacy of the face-to-face interventions.

Section 4: Mapping Interventions to Measures: A Field Guide

Self-report measures are the most widely used means of evaluating interventions. They are also the most regularly criticized (Haefel & Howard, 2010). Many view them as inherently biased and less valid than behavioral or biological measures. We shall argue here, though, that the case is not so simple. Self-reports can be considered behavioral samples, just like those obtained from observing behavior or measuring blood pressure. As such, they are neither inherently superior nor inferior to other behavioral methods. It is clear that self-reports can be valid measures of moods, attributions, plans, attitudes, and beliefs, and can be as good or even better predictors of behavior than behavioral measures (Haefel & Howard, 2010). We are not saying that all self-report measures are created equally, only that they cannot be treated as inherently worse than other types of measures. Each method of measurement is likely to have its strengths and weaknesses.

The biggest challenge in discussing measures is that there are so many of them, and we don't always know how they relate to each other or how they relate to particular interventions. Gross's (1998, 2002) process model of emotional regulation provides one organizing framework. According to this model, emotion can be

regulated at several stages. Antecedent-focused regulation involves strategies that occur before emotional responses become fully activated, and include altering the situation (situation selection or modification) and altering responses to the situation (attentional deployment and cognitive change, or reappraisal). Response-focused strategies refer to what the person does once the emotion is underway, such as hiding expressive behavior, suppressing emotions, and using drugs or exercise to alter the physiological components of emotion.

Gross's model has one major limitation from a contextual behavioral science (CBS) perspective. It puts emotion regulation at the center of the measurement universe, whereas a CBS proponent would put behavior at the center. The Gross perspective emphasizes the idea that emotions are the key therapeutic challenge and the central question is how are emotions moderated, downregulated, hidden, and managed? In contrast, a contextualist approach views behavioral activation as the key therapeutic challenge, and the central question is, how do we help people to behave flexibly in a way that increases their sense of meaning, purpose, and vitality? The regulation of emotion, as defined by Gross, may be a subset of factors that influence value-consistent action, but would not include everything of relevance.

As an alternative to the Gross emotion regulation model, we present the choice point model of behavioral regulation (Ciarrochi, Bailey, & Harris, 2014). This model makes use of some of the Gross distinctions, but puts valued behavior as the primary outcome. The model is illustrated in Figure 2, along with hypothesized measures that map to each process type in the model (see Appendix A for measure references).

A choice point is a moment in time when it is possible to choose between values-consistent and values-inconsistent behavior. This model begins with an understanding of the situation (bottom of model), which describes when, where, and

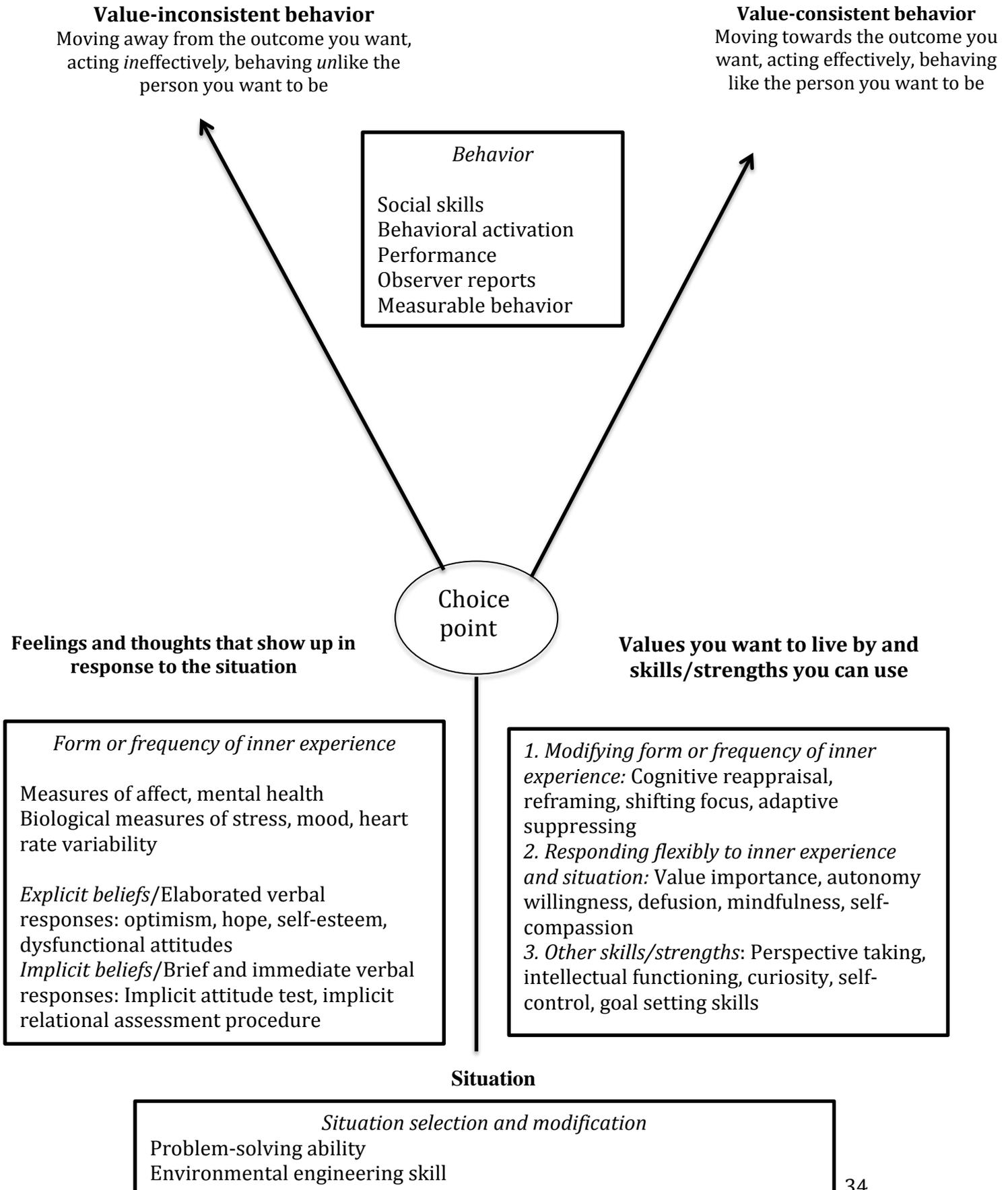
in what circumstances the client finds it challenging to engage in values-based behavior. Some examples might be a divorce, a problem at work, a medical diagnosis, death of a loved one, an upcoming exam or speech, a past regrettable action, a “temptation”, and so on. Problem solving therapy and stimulus control (e.g., removing the temptation) are ways of modifying or improving the situation (Nezu & Nezu, 2001).

Not every situation can be modified, and so when faced with a challenging situation, we often experience difficult thoughts and feelings, such as sadness, anxiety, hopelessness, and self-criticism, as illustrated in the lower left corner of Figure 2. These are the typical clinical outcomes assessed in randomized control trials. We make a distinction between verbal behavior that is relatively “explicit,” slow and elaborated, and is generally measured with self-report measures, and verbal behavior that is relatively “implicit,” quick and immediate, and measured using reaction time tasks (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). The elaborated verbal responses are presumably more susceptible to social desirable responding than the brief responses.

The lower right corner panel of Figure 2 illustrates our responses to the situation and to the inner experience, and these responses can be roughly mapped to different therapy types. Traditional cognitive behavior therapy emphasizes the direct modification of the form or frequency of inner experiences, through interventions such as cognitive reappraisal, reframing, and practicing positive coping statements (Beck, 1995; Ciarrochi & Bailey, 2008). Mindfulness-based therapies such as ACT de-emphasize the direct modifying of inner experience, and instead focus on helping clients respond flexibly to inner experience and the situation (Hayes, Strosahl, & Wilson, 2012). Thus, people learn to have difficult thoughts and feelings, as they are,

without modification, and to choose value-consistent action. Finally, a host of interventions, such as those found in positive psychology, seek to promote a wide range of skills and strengths, such as self-control, perspective taking, and goal setting (Kashdan & Ciarrochi, 2013). It is important to note that most therapies probably target all three of the value/skill/strength components in the bottom right panel of Figure 2. They differ perhaps only in emphasis.

Figure 2. Choice point model and points of measurement and intervention.



Note that the choice point model makes no a-priori assumption as to whether a particular strategy or skill is inherently good or bad. Thus, suppression, in some contexts may support value-consistent behavior, whereas other times it may promote value-inconsistent behavior. Each skill and strength on the right side of the model is tied to the situation in which it is used and the consequences of using it (value-consistent versus value-inconsistent). For example, if suppressing was a toxic strategy for a particular client in a particular context, it would not be listed under the skills/strengths side of the choice point. For more examples on using the choice point model in interventions, please see Ciarrochi et al. (2014).

This choice point model of behavioral regulation goes beyond emotions to capturing any factor that might increase value-consistent behaviors and therefore might be measured. For example, the bottom left panel could include unhelpful beliefs and rules (“I must never eat carbohydrates”), and the bottom right could encompass a skill to undermine the power of that rule (e.g., mindfully stepping back from the rule and noticing that it doesn’t force you to act in a certain way.) There are many skills that could appear within the bottom right panel that do not have to do with regulating feelings, such as empathy, character strengths, transcendent sense of self, intellectual skills, etc. The model is not intended to be a map of physical structures in the brain, but rather a practical map of different types of intervention and intervention measurement. Please see the Appendix for measures that appear to map to each part of the choice point model.

Section 5: Promising New Directions in Contextual Behavioral Measurement

This final section focuses on two promising measurement approaches for contextual behavioral science. First, measures of heart rate variability (HRV) may

provide an objective, biological index of people's ability to flexibly respond to positive and negative situations, and may therefore be an important tool for assessing the biological consequences of interventions. Second, the implicit relational assessment procedure (IRAP) provides an observation measure of verbal behavior, a construct that is central to all cognitive behavior therapies, but that has, up to this point, been largely assessed via self-report measures. The IRAP may be a useful, nonreactive tool for assessing the effect of interventions on client attitudes, beliefs, and other verbal behavior. Because other chapters within this handbook (see . . .) discuss the IRAP in greater detail, our coverage of it here will primarily focus on its possible use within clinical assessment.

Heart Rate Variability (HRV)

HRV is a measure of beat-to-beat temporal changes in heart rate. HRV reflects the output of the central autonomic network that comprises the prefrontal cortex, cingulate cortex, insula, amygdala, and brainstem regions (Kemp & Quintana, 2013). Heart rate variability can be low because of excessive activity of the sympathetic system, underactivity of the parasympathetic system, or both. Low HRV, then, can be understood as reduced differentiation between contexts that elicit parasympathetic activity (i.e., positive) and those that elicit sympathetic activity (negative). For example, imagine a workaholic on a beach, stressing about a deadline, insensitive to all the peace, calm, and beauty around him. Consistent with this interpretation of HRV, Ruiz-Padial, Sollers, Vila, and Thayer (2003) found that those with highest HRV showed the most differentiated emotion-modulated start effects, with little startle response to pleasant pictures, more response for neutral pictures, and the greatest response to negative pictures. In general, HRV has been linked to low self-regulatory strength, reduced motivation to engage in social situations, and lower

psychological flexibility in the face of stressors (Appelhans & Luecken, 2006; Kemp & Quintana, 2013; Segerstrom & Nes, 2007). In the long run, lower HRV leads to reduced immune function, inflammation, cardiovascular disease, and mortality (Kemp & Quintana, 2013).

We believe measuring HRV has several strengths within contextual behavioral intervention approaches. First, it appears to be a biological measure of sensitivity to positive, neutral, and negative contexts. CBS interventions such as ACT are intended to increase responsiveness to environmental feedback, and so HRV would seem to be an excellent way of assessing this outcome. Second, HRV does not rely on self-report and is therefore relatively objective and difficult to fake. Third, HRV links broadly to both physical and psychological health. Thus, if a practitioner seeks to improve both physical and psychological functioning, HRV may offer a simple index of improvement.

Implicit Relational Assessment Procedure

Relational frame theory (RFT) emphasizes the importance of learning and applying relational frames in human language and cognition (Hayes, Barnes-Holmes, & Roche, 2001). The Implicit Relational Assessment Procedure (IRAP) provides a method for directly observing and recording verbal behavior, and variables that influence that behavior (Barnes-Holmes, Barnes-Holmes, et al., 2010; Hayes, Barnes-Holmes, & Wilson, 2012). A number of studies have demonstrated the utility of the IRAP in domains such as food preferences (Barnes-Holmes, Murtagh, Barnes Holmes, & Stewart, 2010), depression (Hussey & Barnes-Holmes, 2012), self-esteem (Timko, England, Herbert, & Forman, 2010), eating disorders (Parling, Cernvall, Stewart, Barnes-Holmes, & Ghaderi, 2012), and formation of new attitudes (Hughes & Barnes-Holmes, 2011). The IRAP also appears to be very difficult to fake which

offers a potential advantage of this method over questionnaires, interviews, and implicit cognition measures (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007).

The Implicit Association Test (IAT) offers an alternative to the IRAP (Greenwald, McGhee, & Schwartz, 1998), but it does have a major limitation from a behavioral perspective. It infers attitudes by taking the difference between negative responses (e.g., “Thin people are bad”) and positive response (“Thin people are good”). From a CBS perspective, these are two different behaviors, not two items reflecting a single hidden construct called “attitude.” If one uses the IAT and finds no evidence for a negative attitude, this does not indicate the absence of positive and negative responses. A null result in the IAT could mean that people either have no verbal reactions to thin people or they could have anti-thin and pro-thin reactions. In the latter case, the two responses would counteract each other and amount to no bias in the IAT. In contrast, the IRAP would be able to detect both responses. For example, Roddy, Stewart, and Barnes-Holmes (2010) found that the overall IRAP score correlated significantly with the IAT, suggesting pro-slim/anti-fat attitudes. However, the IRAP included four trial types that showed that the participants demonstrated a so-called “pro-slim” response, but there was no “anti-fat” response.

Concerning the potential link between the IRAP and therapy, the IRAP has been used to predict treatment outcome for cocaine dependence (Carpenter, Martinez, Vadhan, Barnes-Holmes, & Nunes, 2012) and increases in weight, body dissatisfaction, and disordered eating over a span of 7.8 months (Juarascio et al., 2011). Cullen, Barnes-Holmes, and Barnes-Holmes (2009) found that presenting pictures of admired old people influenced how participants responded on the IRAP. The images reduced the pro-young IRAP effect, and reversed the anti-old IRAP

effect, that was maintained when retested 24 hours later. Also, Hooper, Villatte, Neofotistou, and McHugh (2010) found that a mindfulness induction influenced performance on an experiential avoidance IRAP; the mindfulness group experienced lower levels of experiential avoidance after the induction when compared to a thought suppression group. Hence, the IRAP offers CBS researchers a way to objectively measure verbal behavior and examine the influence of context on that behavior.

Summary and Conclusions

With the recent passing of Gordon Paul, we are reminded of the challenging question he posed to clinical researchers and practitioners now nearly a half century ago: “What treatment, by whom, is most effective for this individual, with that specific problem, under which set of circumstances, and how does it come about?” (Paul, 1969, p. 44). This chapter sought to provide answers to most parts of this question (except that dealing with “by whom” or ideal therapist characteristics).

Our discussion of treatment determination and guidance addressed the question, “*What treatment with what specific problem is likely to be the most effective?*”. Our discussion of moderating variables addressed the question of “*What is likely to be most effective for a particular individual.*” None of these questions, of course, can be adequately answered without the administration of reliable and valid outcome measures. In addition, if we do have sound outcome measures and demonstrate positive outcomes, we then also must have psychometrically sound process measures to address the question, “*How does it come about?*”.

Considerable progress has been made over the past five decades in providing empirically-based answers to Paul’s question. This is the good news. The bad news is that addressing the question is an ongoing process and there is still much more critical work to be done by both clinical researchers and practitioners. There are, for

example, still remarkably few treatment utility studies, and we don't know the extent that measures improve treatment outcomes. Our belief is that contextual behavioral science can make a significant contribution in improving the utility of clinical assessment. Our hope is that this chapter may make at least a small contribution to this larger process.

Appendix

Process and Example Measures

Situation Selection and Modification

D'Zurilla, T., Nezu, A., & Maydeu-Olivares, A. (2004). Social problem solving: Theory and assessment. In E. Change, T. D'Zurilla, & L. Sanna (Eds.), *Social problem solving: Theory, research, and training* (pp. 11-27). Washington, DC: American Psychological Association.

Form and Frequency of Inner Experience

Mental Health and Well-Being

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71-75.

Goldberg, D. (1978). *Manual of the General Health Questionnaire*. Windsor, United Kingdom: National Foundation for Education Research.

Keyes, C. (2006). Mental health in adolescence: Is American's youth flourishing. *The American Journal of Orthopsychiatry*, *76*, 395-402.

Lovibond, S., & Lovibond, P. (1995). *Manual for the Depression Anxiety Stress Scales*. Sydney, Australia: The Psychology Foundation of Australia, Inc.

Implicit Beliefs/Brief and Immediate Verbal Responses

Barnes-Holmes, D., Murtagh, L., Barnes Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes towards meat and vegetables in vegetarians and meat-eaters. *The Psychological Record*, *60*, 287-306.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social psychology*, *79*, 631-43.

Explicit Beliefs/Elaborated Verbal Responses

Brown, G. P., Hammen, C. L., Craske, M. G., & Wickens, T. D. (1995). Dimensions of dysfunctional attitudes as vulnerabilities to depressive symptoms. *Journal of Abnormal Psychology, 104*, 431-435.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Snyder, C. R., Harris, C., Anderson, J. R., Holleran, S. A., Irving, L. M., Sigmon, S. T., . . . Harney, P. (1991). The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology, 60*, 570-585.

Efforts to Modify Form or Frequency of Inner Experience

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56*, 267-283

Gratz, K. L., & Roemer, L. (2008). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment, 26*, 41-54.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*, 348-362.

Responding Flexibly to Inner Experience

Baer, R. A., Smith, G. S., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment, 13*, 27-45.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., . . . Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire - II: A revised measure of psychological flexibility and experiential avoidance. *Behavior Therapy*, *42*, 676-688.

Forman, E., Herbert, J., Juarascio, A., Yeomans, P., Zebell, J., Goetter, E., & Moitra, E. (2012). The Drexel Defusion Scale: A new measure of experiential distancing. *Journal of Contextual Behavioral Science*, *1*, 55-65.

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity*, *2*, 223-250.

Values and Personal Strivings

Sheldon, K. M., & Kasser, T.. (2001). Goals, congruence, and positive well-being: New empirical support for humanistic theories. *Journal of Humanistic Psychology*, *41*, 30-50.

Veage, S., Ciarrochi, J., & Heaven, P. C. L. (2011). Importance, pressure, and success: Dimensions of values and their links to personality. *Personality and Individual Differences*, *50*, 1180-1185.

Wilson, K. G., Sandoz, E. K., Kitchens, J., & Roberts, M. E. (2010). The Valued Living Questionnaire: Defining and measuring valued action within a behavioral framework. *The Psychological Record*, *60*, 249-272.

Other Skills/Strengths

Davis, M. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113-126.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. New York, NY and Washington, DC: Oxford University Press and American Psychological Association.

Observable Behavior

Manos, R. C., Kanter, J. W., & Busch, A. M. (2010). A critical review of assessment strategies to measure the behavioral activation model of depression. *Clinical Psychology Review, 30*, 547–61.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*, 33–66

Note: Many of the measures mentioned in the values section assess value-consistent activity.

References

- American Education Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle, NJ: Pearson.
- Andre, F., Bachelot, T., Commo, F., Campone, M., Arnedos, M., Dieras, V., . . . Bonnefoi, H. (2014). Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: A multicentre, prospective trial (SAFIRO1/UNICANCER). *Lancet Oncology*, *15*, 267-274.
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the implicit relational assessment procedure (IRAP) and the relational elaboration and coherence model (REC). *The Psychological Record*, *60*, 527-542.
- Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes towards meat and vegetables in vegetarians and meat-eaters. *The Psychological Record*, *60*, 287-306.
- Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. New York, NY: Guilford.
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York, NY: Guilford.
- Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). Measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, *42*, 861-865.
- Bieling, P. J., Beck, A. T., & Brown, G. K. (2000). The Sociotropy-Autonomy Scale: Structure and implications. *Cognitive Therapy and Research*, *24*, 763-780.

- Bond, F. W., Flaxman, P. E., & Bunce, D. (2008). The influence of psychological flexibility on work redesign: Mediated moderation of a work reorganization intervention. *Journal of Applied Psychology, 93*, 645-634.
- Bullock, J., Green, D., & Ha, S. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550-558.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Carpenter, K. M., Martinez, D., Vadhan, N., Barnes-Holmes, D., & Nunes, E. V. (2012). Measures of attentional bias and relational responding are associated with behavioral treatment outcome for cocaine dependence. *The American Journal of Drug and Alcohol Abuse, 38*, 146-154.
- Ciarrochi, J., & Bailey, A. (2008). *A CBT-practitioner's guide to ACT: How to bridge the gap between cognitive behavioral therapy and acceptance and commitment therapy*. Oakland, CA: New Harbinger.
- Ciarrochi, J., Bailey, A., & Harris, R. (2014). *The weight escape*. Sydney, Australia: Penguin.
- Ciarrochi, J., Bilich, L., & Godsel, C. (2010). Psychological flexibility as a mechanism of change in acceptance and commitment therapy. In R. Baer (Ed.), *Assessing mindfulness and acceptance: Illuminating the processes of change* (pp. 51-76). Oakland, CA: New Harbinger.

- Cohen, L. H., Gunthert, K., Butler, A., Parrish, B., Wenze, S., & Beck, J. (2008). Negative affective spillover from daily events predicts early response to cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 76*, 955.
- Cullen, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2009). The implicit relational assessment procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record, 59*, 591-620.
- Dalrymple, K. L., & Herbert, J. D. (2007). Acceptance and commitment therapy for generalized social anxiety disorder: A pilot study. *Behavior Modification, 31*, 543-568.
- Delespaul, P., & van Os, J. (2002). Determinants of occurrence and recovery from hallucinations in daily life. *Social Psychiatry and Psychiatric Epidemiology, 37*, 97-104.
- English, C. L., & Anderson, C. M. (2006). Evaluation of the treatment utility of the analog functional analysis and the structured descriptive assessment. *Journal of Positive Behavior Interventions, 8*, 212-229.
- Eysenck, H. J. (1959). Review of the Rorschach Inkblot Test. In O. K. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 276-278). Highland Park, NJ: Gryphon Press.
- Farber, B. A., Berano, K. C., & Capobianco, J. A. (2004). Clients' perceptions of the process and consequences of self-disclosure in psychotherapy. *Journal of Counseling Psychology, 51*, 340-346.
- Ferster, C. (1965). Classification of behavior pathology. In L. Krasner & L. P. Ullman (Eds.), *Research in behavior modification* (pp. 6-26). New York, NY: Holt, Rinehart, & Winston.

- Forman, E. M., Butryn, M. L., Juarascio, A. S., Bradley, L. E., Lowe, M. R., Herbert, J. D., & Shaw, J. A. (2013). The Mind Your Health Project: A randomized controlled trial of an innovative behavioral treatment for obesity. *Obesity, 21*, 1119-1126.
- Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480.
- Gross, J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*, 271-299.
- Gross, J. (2002). Emotion regulation: Affective, cognitive and social consequences. *Psychophysiology, 39*, 281-291.
- Haefffel, G., & Howard, G. (2010). Self-report: Psychology's four-letter word. *American Journal of Psychology, 123*, 181-188.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Kluwer Academic/ Plenum.
- Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science 1*, 1-16.
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1986). Evaluating the quality of behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 461-503). New York, NY; Guilford.

- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42*, 963-974.
- Hayes, S. C., Strosahl, K., & Wilson, K. G. (2012). *Acceptance and commitment therapy: The process and practice of mindful change* (2nd ed.). New York, NY: Guilford.
- Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology, 15*, 1-39.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 53-60.
- Hooper, N., Villatte, M., Neofotistou, E., & McHugh, L. (2010). The effects of mindfulness versus thought suppression on implicit and explicit measures of experiential avoidance. *The International Journal of Behavioral Consultation and Therapy, 6*, 233-243.
- Hughes, S., & Barnes-Holmes, D. (2011). On the formation and persistence of implicit attitudes: New evidence from the Implicit Relational Assessment Procedure. *The Psychological Record, 61*, 391-410.
- Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11*, 266-277.
- Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice, 19*, 573-582.

- Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities, 2*, 3-20.
- 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.
- Jensen, A. R. (1965). Review of the Rorschach Inkblot Test. In O. K. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 501-509). Highland Park, NJ: Gryphon Press.
- Juarascio, A., Forman, E., Timko, C. A., Herbert, J., Butryn, M., & Lowe, M. (2011). Implicit internalization of the thin ideal as a predictor of increases in weight, body dissatisfaction and disordered eating. *Eating Behaviors 12*, 207-213.
- Kashdan, T. B., & Ciarrochi, J. (Eds.). (2013). *Mindfulness, acceptance, and positive psychology: The seven foundations of well-being*. Oakland, CA: New Harbinger.
- Kashdan, T. B., Ferrisizidis, P., Collins, R. L., & Muraven, M. (2010). Emotion differentiation as resilience against excessive alcohol use: An ecological momentary assessment in underage social drinkers. *Psychological Science, 21*, 1341-1347.
- Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1-27.
- Kemp, A., & Quintana, D. (2013). The relationship between mental and physical health: Insights from the study of heart rate variability. *International Journal of Psychophysiology, 89*, 288-296.

- Klerman, G. L., Weissman, M. M., Rounsaville, B. J., & Chevron, E. S. (1984). *Interpersonal therapy of depression*. New York, NY: Basic Books.
- Kline, R. (2010). *Principles and practice of structural equation modelling* (3rd ed.). New York, NY: Guilford.
- Kuyken, W., Watkins, E., Holden, E., White, K., Taylor, R., Byford, S., . . . Dalglish, T. How does mindfulness-based cognitive therapy work? *Behavior Research and Therapy*, *48*, 1105-1112.
- Lappalainen, R., Lehtonen, T., Skarp, E., Taubert, E., Ojanen, M., & Hayes, S. C. (2007). The impact of CBT and ACT models using psychology trainee therapists. *Behavior Modification*, *31*, 488-511.
- Marsh, H., Muthen, B., Asparouhove, T., Ludtke, O., Robitzsch, A., Morin, A., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Applications to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439-476.
- Marshal, S., Parker, P., Ciarrochi, J., & Heaven, P. (2013). Is self-esteem a cause or consequence of social support: A five year, five wave longitudinal study amongst adolescence. [Submitted]. *Child Development*.
- Masuda, A., Hayes, S. C., Fletcher, L. B., Seignourel, P. J., Bunting, K., Herbst, S. A., . . . Lillis, J. (2007). Impact of acceptance and commitment therapy versus education on stigma toward people with psychological disorders. *Behaviour Research and Therapy*, *45*, 2764-2722.
- McKenna, I., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the fake-ability of the Implicit Relational Assessment Procedure (IRAP): The first study. *International Journal of Psychology and Psychological Testing*, *7*, 253-268.

- McNulty, J., & Fincham, F. (2011). Beyond positive psychology? Toward a contextual view of psychological processes and well-being. *American Psychologist, 67*, 101-110.
- Meyer, G. J. (1999). Introduction to the special series on the utility of the Rorschach for clinical assessment. *Psychological Assessment, 11*, 235-239.
- Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., & Narod, S. A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: Randomised screening trial. *BMJ, 348*, g366.
- Millon, T. (2006). *Millon Clinical Multiaxial Inventory-III manual*. Minneapolis, MN: NCS Pearson.
- Nelson, R. O., & Hayes, S. C. (1986). The nature of behavioral assessment. In R.O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 3-41). New York, NY: Guilford.
- Nelson-Gray, R. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15*, 521-531.
- Nezu, A. M., & Nezu, C. M. (2001). Problem solving therapy. *Journal of Psychotherapy Integration, 11*, 187-205.
- Nolen-Hoeksema, S., Morrow, J., & Fredrickson, B. L. (1993). Response styles and the duration of depressed mood. *Journal of Abnormal Psychology, 102*, 20-28.
- Nunnally, J. (1967). *Psychometric theory*. New York, NY; McGraw-Hill.
- O'Neill, R. E., Horner, R. H., Albin, R. W., Story, K., Sprague, J. R. (1990). *Functional analysis of problem behavior: A practical assessment guide*. Sycamore, IL: Sycamore Publishing.

- Parling, T., Cernvall, M., Stewart, I., Barnes-Holmes, D., & Ghaderi, A. (2012). Using the Implicit Relational Assessment Procedure to compare implicit pro-thin/anti-fat attitudes of patients with anorexia nervosa and nonclinical controls. *Eating Disorders* 20, 127-143.
- Paul, G. L. (1969). Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status* (pp. 29-62). New York, NY: McGraw-Hill.
- Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse: Results from the Epidemiologic Catchment Area (ECA) Study. *Journal of the American Medical Association*, 264(19), 2511-2518.
- Repp, A. C., Felce, D., & Barton, L. E. (1988). Basing the treatment of stereotypic and self-injurious behaviors on hypotheses of their causes. *Journal of Applied Behavior Analysis*, 21, 281-289.
- Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both: Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology*, 15, 416-425.
- Ruiz-Padial, E., Sollers, J., Vila, J., & Thayer, J. (2003). The rhythm of the heart in the blink of an eye: Emotion-modulated startle magnitude covaries with heart rate variability. *Psychophysiology*, 40, 306-313.
- Sartorius, N., Ustun, T. B., Lecruiber, Y., & Wittchen, H-U. (1996). Depression comorbid with anxiety: Results from the WHO study on "Psychological disorders in primary health care." *The British Journal of Psychiatry*, 168, 38-43.

- Segerstrom, S. C., & Nes, L. S. (2007). Heart rate variability reflects self-regulatory strength, effort, and fatigue. *Psychological Science, 18*, 275-281.
- Shiffman, S., Stone, A., & Hufford, M. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1-32.
- Sotsky, S. S., Glass, D. R., Shea, T., Pilknois, P. A., Collins, J. F., Elkin, I., . . . Oliveri, M, E. (1991). Patient predictors of response to psychotherapy and pharmacotherapy: Findings in the NIMH Treatment of Depression Collaborative Research Program. *The American Journal of Psychiatry, 148*, 997-1008.
- Timko, C. A., England, E. L., Herbert, J. D., & Forman, E. (2010). The Implicit Relational Assessment Procedure as a measure of self-esteem. *The Psychological Record, 60*, 679-698.
- Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11*, 251-265.
- Wilson, K. G., Sandoz, E. K., Kitchens, J. , & Roberts, M. (2010). The Valued Living Questionnaire: Defining and measuring valued action within a behavioral framework. *The Psychological Record, 60*, 249-272.
- Zettle, R. D. (2003). Acceptance and commitment therapy (ACT) vs. systematic desensitization in treatment of mathematics anxiety. *The Psychological Record, 53*, 197-215.
- Zettle, R. D., Haflich, J. L., & Reynolds, R. A. (1992). Responsivity of cognitive therapy as a function of treatment format and client personality dimensions. *Journal of Clinical Psychology, 48*, 787-797.

Zettle, R. D., & Herring, E. L. (1995). Treatment utility of the sociotropy/autonomy distinction: Implications for cognitive therapy. *Journal of Clinical Psychology, 51*, 280-288.