MEASUREMENT, *4*(3), 188–197

REJOINDER

New Coke, Rosetta Stones, and Functional Data Analysis: Recommendations for Developing and Validating New Measures of Depression

Darcy A. Santor
School of Psychology
University of Ottawa, Ontario, Canada
Provincial Centre of Excellence in Child and Youth Medical Health
Ottawa, Ontario, Canada

In this rejoinder, I outline 6 recommendations that may guide the continued development and validation of measures of depression. These are (a) articulate a formal theory of signs and symptoms; (b) differentiate complex theoretical goals from pragmatic evaluation needs; (c) invest heavily in new methods and analytic models; (d) calibrate all measures against a common set of items or indicators; (e) seek period expert consensus on theory, measures, and analytic models; and (f) establish a registry of scales and measures to facilitate knowledge exchange. With hundreds of measures now in existence, few of which are used routinely, and many of which have not been revised in decades, continued progress in the assessment of depression will depend on a clear differentiation of both scientific and evaluation goals, on our ability to utilize new methods and technologies systematically, and on our willingness to establish a common reference point for comparison.

The invited commentaries are a welcome supplement to our efforts (Santor, Gregus, & Welch, this issue) to examine the characteristics of measures of depres-

Correspondence should be addressed to Darcy A. Santor, School of Psychology, University of Ottawa, Lamoureaux Hall, 145 Jean-Jacques-Lussier St., P. O. Box 450, Station A, Ottawa, Ontario, K1N 6N5, Canada. E-mail: dsantor@uottawa.ca

sion developed in the past 80 years and the frequency with which those measures have been used. Collectively, these commentaries highlight a broad range of conceptual and methodological challenges to determine how existing measures of depression are selected and evaluated, as well as the manner in which new measures of depression will be developed. The call for prescriptive recommendations from Tanner and Haaga (this issue) is timely and deserves response from experts in the fields. In this rejoinder, I highlight some of the many important issues raised by the commentators and outline six recommendations concerning how measures might be developed and validated in the coming years.

RECOMMENDATION 1: ARTICULATE AND REVISE A FORMAL THEORY OF SIGNS AND SYMPTOMS

Defining and operationalizing depression is an ongoing endeavor that should reflect the result of careful investigation at various levels of analysis, ranging from neurocellular mechanisms to broad cultural differences (Sternberg, 2004; Sternberg & Grigorenko, 2001). The debate on (a) what depression is, (b) what symptoms characterize the disorder, (c) which signs may best serve as markers of the underlying construct, (d) how these symptoms are best operationalized, (e) the manner in which indicators or items perform across a continuum of severity from mood states to illness, and (f) which items or indicators will maximize the manner in which depression is identified (inclusionary signs and symptoms) and distinguished (exclusionary signs and symptoms) from other constructs and syndromes will be resolved on the basis of a number of theoretical and pragmatic considerations.

Historical accounts and empirical reviews suggest that the symptoms that characterize depression are likely to change. Early accounts of depression tended not to assess symptoms of guilt (Jackson, 1986), and our own research suggests that the composition of scales has changed over time (Santor et al., this issue). Advances in both ethological (Gilbert, 1992; Price, 1967) and neurobiological models (Nestler et al., 2002) of depression have yet to influence the construction of measures of depression, but it is reasonable to assume that attempts to adopt constructs such as social withdrawal and feeling entrapped (Gilbert, 1992) or treatment response (Alda, 1999) may be considered in future depression scales or proposed as specific subtypes of depression, just as *hopelessness* depression has evolved out of learned helplessness and attributional theories of depression (Alloy, Abramson, Metalsky, & Hartlage, 1988).

Recently, I have proposed that in designing or evaluating any psychometric measure, scale developers should articulate a formal theory of signs and symptoms designed to answer a number of questions (cf. Santor, 2005), including (a) what the ontological status of the construct being assessed is or is assumed to be (i.e., is de-

pression a mood state or a symptom of an illness, which is either continuous or categorical in nature); (b) what symptoms, signs, and associated conditions should and should not be included in our definitions and measures; (c) how those symptoms are related to the underlying construct and should be combined to derive an index of severity; (d) how many and what type of questions should be used to operationalize each symptom domain being assessed; and (e) how those symptoms are expected to be expressed differently in different groups, whether the groups are differentiated with respect to age, gender, culture, comorbid condition, or any other characteristic.

The benefits of articulating each of these components offers, at a minimum, clarity in comparing the myriad of measures that currently exist and, ideally, guide the systematic development of new measures. A number of the commentaries to our article raise issues concerning various features of the construct, namely whether depression is primarily a category or continuum (Pepper & Nieuwsma, this issue), how depression is related to other affective states (Joseph, this issue), whether depression is primarily multidimensional or unidimensional (Simms, this issue), to what extent the construct can be sufficiently operationalized, as intended (Kane, this issue), and whether the construct should vary according to the respondent demographics (Tanner & Haaga, this issue).

As noted by Kane (this issue), operationalizing the construct as intended is often difficult, given that the theory is frequently insufficiently explicit or elaborated (cf. Cronbach, 1989). The purpose of articulating a formal theory of signs and symptoms for each construct and measure is to make explicit the manner in which the construct should be operationalized (cf. Santor, 2005). As implied by Joseph (this issue), the context in which depression is assessed should also consider the relation between health and illness.

Assessing both illness and health as part of a single continuum makes a clear theoretical statement about their interrelation—one tends to exclude the other. There are likely to be both advantages and disadvantages in assessing both illness- and wellness-related constructs within a single scale such as the Center for Epidemiologic Depression Scale (CES-D; Radloff, 1977). Indeed, assessing a broader construct that includes both illness and wellness will likely produce a broader range of scores, which may account for the greater discriminability of the CES-D relative to the Beck Depression Inventory (BDI) in nonclinical populations (Santor, Zuroff, Ramsay, Cervantes & Palacios, 1995), but may also account for the great variability in cut scores (see Santor, 2000, for a brief review).

Definitions of health and illness are not, however, necessarily exclusive. As defined in the U.S. Surgeon General Report on Mental Health (U.S. Surgeon General Office, 1999), *mental illness* refers to "all diagnosable mental disorders, which are health conditions that are characterized by alterations in thinking, mood, or behavior (or some combination thereof) associated with distress and/or impaired functioning," whereas *mental health* is viewed as "a state of successful performance of

mental function, resulting in productive activities, fulfilling relationships with other people, and the ability to adapt to change and to cope with adversity." Although health and illness are certainly interrelated, they may be best assessed as distinct components rather than measured as bipolar opposites.

Prescribing which symptoms should be included, the number of items allocated to assess each symptom, and which of those symptoms should be given status as core symptoms will be a fine debate, requiring expert counsel and a clear statement regarding the purpose of assessment, whether to diagnose, screen, assess severity, or evaluate change. Core symptoms should be evaluated with an equal number of indicators or items, and what constitutes a "central" or core feature should be determined empirically. Results of our empirical review (Santor et al., this issue) showed that more items were allocated, on average, to assessing worthlessness in measures of depression created over the past 80 years than other symptoms, which we interpreted as evidence that scale developers have generally acted as if worthlessness were a core symptom of depression. Given that worthlessness is a core mechanism in many cognitive models of depression, worthlessness could easily be considered a core feature of depression alongside sad mood and anhedonia. However, this should be substantiated empirically, which we are currently pursuing.

RECOMMENDATION 2: DIFFERENTIATE COMPLEX SCIENTIFIC NEEDS FROM PRAGMATIC EVALUATION GOALS

Given the heterogeneity of symptoms and the array of distinct biological processes involved in regulating the various drives, capacities and biorhythms that are frequently disrupted in individuals with depression, it should not be surprising to find a degree of multidimensionality in responses to scale items. Several measures created over the past several decades have conceptualized depression as a multidimensional construct, and psychometric analyses have identified distinct factors in many of the most popular measures of depression, including the CES-D and the BDI.

Fulfilling the theoretical or substantive goal to understand the nature and degree to which specific mood, cognitive, and biological symptoms of depression are interrelated should be distinguished from and balanced against the pragmatic goal of identifying a smaller set of largely unidimensional items that are ideal for evaluating the effectiveness of treatments or for screening individuals for the presence or absence of illness or difficulties. As we have argued elsewhere, these are fundamentally different purposes that require scale items or indicators to function in very different ways, which may be best achieved only through different measures (cf. Santor & Coyne, 1997).

Scales and measures represent much of the technology used to achieve both of these goals. They streamline how information is managed and simplify how decisions are made (cf. Santor & Ramsay, 1998). However, scales that show substantial multidimensionality will complicate these decisions. Multidimensionality is, however, an issue of degree, and the degree of multidimensionality observed in scales such as the CES-D should be estimated routinely, though the use of statistics, such as a target coefficient (Marsh & Hocevar, 1985), which quantifies the extent to which first-order factors, of which there may be many, can be accounted for by a higher order, second factor. Given that items for most scales are selected to be internally consistent, it should not be surprising that the bulk of variance in item responses will be attributed to the second, single-order factor rather than the multitude of first-order factors.

RECOMMENDATION 3: INVEST HEAVILY IN NEW METHODS AND ANALYTIC MODELS

Over the past 8 decades, the predominant measure of scale performance has been, without exception, some form of internal reliability, consistency, or index of model fit or sufficiency. High reliabilities were achieved by Woodworth in 1918 (cf. Garret & Schenk, 1928), with what is arguably the first published measure of depression, the Psychoneurotic Inventory, and the vast majority of measures published to date have had adequate to good levels of reliability. However, given that most published measures will have high internal consistencies or show adequate fit or sufficiency in confirmatory models, choosing which measure of depression performs best on the basis of traditional performance indexes will be difficult.

Psychometric methods traditionally used to initially select, shorten, and revise item pools have generally emphasized the relation of items to underlying components of variation rather than other equally important indexes of performance, such as (a) the ability of short forms to reproduce the characteristics of the original scale (cf. Santor, Zuroff, & Fielding, 1997), or (b) the ability of items to discriminate among individual differences in severity throughout the entire continuum of severity (Santor, De Brota, Gelwicks, & Englehardt, 2006; Santor & Ramsay, 1998).

There is good evidence to support new methodologies and analytic models that may offer better ways to develop and select candidate items for measures of depression, better ways to assess change in symptoms over time, and better ways to model the changes that do occur. As noted by Tennen (this issue), there is now clear benefit in assessing clinical outcomes daily and modeling change over a large number of time points, rather than from just baseline to follow-up (Gunthert, Cohen, Butler, & Beck, 2005; Kranzler, Armeli, Feinn, & Tennen,

2004; Tennen, Affleck, Coyne, Larsen, & DeLongis, in press). In our own work, we have used growth curves to model change in symptoms (Santor & Segal, 2001) showing that rates of improvement during the first 10 weeks of treatment were superior to either follow-up scores or absolute improvement in predicting symptom return after 3 and 6 months. My colleagues at Eli Lilly and I have recently used item response models to rationally select a subset of items from the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960) that optimally differentiate individual differences in depressive severity that were subsequently shown to be more sensitive to treatment changes over time (Santor et al., 2006). Finally, by modeling scale performance as a function of scale length, we have been able to develop a short form of the Depressive Experiences Questionnaire (Blatt, D'Afflitti, & Quinlan, 1976; Santor et al., 1997) that preserves the psychometric characteristics of the original scale where other shortened versions of the scale have failed (Bagby, Parker, Joffe, & Buis, 1994; Viglione, Lovette, & Gottlieb, 1995; Welkowitz, Lish, & Bond, 1985).

All techniques designed to more closely model the performance of test scores as a function of some other variable, whether that be severity, time, or scale length, belong to a broader branch of statistics and mathematics, called *functional data analysis* (Ramsay & Dalzell, 1991), where the focus of interest is on modeling the entire observed function rather than a string of data points or observations, possibly as few as one or two. Several different methods and analytic techniques, including *item response models* (Lord, 1980), *growth curve modeling* (Singer & Willett, 2003), and *daily process methods* (Tennen, Affleck, Armeli, & Carney, 2000), share this common goal of modeling the dynamic features of an underlying process (cf. Neufeld, 1999) and offer a wealth of new analytic models available to assess item and scale performance.

Unfortunately, the amount of investment in methods and technologies to identify the treatment effect that interventions are believed to produce is extremely small relative to the amount invested in developing new pharmacological agents, even though the identification of those effects depends on the quality of the methods and technologies used, which tend to be adopted slowly and in general have been vastly underutilized.

RECOMMENDATION 4: CALIBRATE ALL MEASURES AGAINST A COMMON ROSETTA STONE

Given the large number of measures of depression currently used with any frequency, any ongoing program of research that directly compares the performance of more than one or two measures at a time will be costly and time consuming. Accordingly, there is a need to develop a method by which all measures

can be calibrated and compared efficiently and effectively. In earlier work, my colleagues and I illustrated this methodology showing how scores on the BDI and CES-D could be equated using a common metric (Santor et al., 1995), which not only facilitates a comparison of how scales perform at different levels of severity but also allows researchers and clinicians to translate scores from one measure to another.

Were every scale developed and validated with respect to a common set of indicators, the need for large validation studies, where several measures are compared simultaneously, would be diminished. Although reaching a consensus on what items should constitute this "anchor" measure or "Rosetta stone" will be difficult, this should not eclipse the need for a common set of anchors, which itself may even evolve over time. So long as there remains a common core set of indicators, scores on various measures can be equated, which is essential to evaluating the relative performance of different measures and to preserving the continuity of knowledge as measures develop and are revised over time.

The reluctance to move from well understood, albeit imperfect, measures of depression such as the HRSD (Hamilton, 1960) to potentially better but untested measures is understandable. Effect sizes in clinical trials can be small, and the risk of effect sizes diminishing even further is concerning. The absence of such an anchor measure may be one of the main impediments to adopting new measures of depression within treatment trials but perhaps one of the main vehicles for facilitating this change.

RECOMMENDATION 5: ESTABLISH A CONSENSUS CONFERENCE

Although progress in the assessment of depression has been considerable, there are a number of enduring questions concerning most, if not all, measures of depression. The need for periodic consensus—from researchers, clinicians, industry, and regulators—is essential to ensure the field's continued progress. As many will be acutely aware, identifying a newer, better product and marketing it successfully are two very different endeavors. Adopting the "New Coke" as the gold standard over traditionally used measures will require (a) clear evidence that the newer measure is superior to existing measures (for which the necessary analytic methods now exist), as well as (b) consensus from a broad range of stakeholders.

Establishing a consensus conference is therefore essential to determine, in a proactive manner, and revise, when appropriate, which indicators, items, or symptoms constitute our best understanding of what depression is, what the best measures are to achieve both our substantive and evaluation goals, what items and features of depression should compose a routinely administered anchor measure, and

what the best methods are to assess both differences in severity and symptom changes over time.

RECOMMENDATION 6: ESTABLISH A REGISTRY OF SCALES AND MEASURES TO FACILITATE KNOWLEDGE EXCHANGE

Scales, inventories, and diagnostic interviews constitute one of the most pervasive and influential technologies used in psychology (cf. Santor & Ramsay, 1998). The number of studies appearing annually that report on the development, validation, or revision of some scale or measure is large, and there are few areas in psychology that have not used such measures to examine the importance of individual differences to the phenomenon being investigated. The American Psychological Association (APA) has estimated recently that some 20,000 new measures appear every year, many of which may only have been used once (APA, 1993).

Maximizing the benefit obtained from those measures currently in existence and minimizing the possibility of creating additional measures unnecessarily will depend on the efficient exchange of knowledge among all stakeholders. To this end, the existing database of depression measures will be maintained and updated online at www.scalesandmeasures.net.

CONCLUSION

There are good reasons for constructing different measures of depression. Depression has been conceptualized in different ways and may be expressed differently in various populations and contexts. However, the abundance of measures has increased both the complexity and quantity of information we possess. With hundreds of measures now in existence, few of which are used routinely, and many of which have not been revised in decades, continued progress in the assessment of depression will depend on a clear differentiation of both scientific and evaluation goals, on our ability to utilize new methods and technologies systematically, and on our willingness to establish a common reference point of comparison. Failure to equate measures with regard to differences in scale discriminability or differential item functioning means that the results gathered from these measures may be misleading and the conclusions drawn possibly unwarranted.

Many of these recommendations are being implemented by the Depression Inventory Development Project, which is a multistakeholder panel of experts from university, industry, and clinical practice that is developing a new clini-

cian-rated measure of depressive severity being evaluated with techniques based on item response theory.

REFERENCES

Alda, M. (1999). Pharmacogenetics of lithium response in bipolar disorder. *Journal of Psychiatry & Neuroscience, 24,* 154–158.

Alloy, L. B., Abramson, L. Y., Metalsky, G. I., & Hartlage, S. (1988). The hopelessness theory of depression: Attributional aspects. *The British Journal of Clinical Psychiatry, 27,* 5–21.

American Psychological Association. (1993). Internal report on measurement. *American Psychologist, 48,* 182.

Bagby, R. M., Parker, J. D. A., Joffe, R. T., & Buis, T. (1994). Reconstruction and validation of the Depressive Experiences Questionnaire. *Assessment, 1,* 59–68.

Blatt, S. J., D'Afflitti, J. P., & Quinlan, D. M. (1976). Experiences of depression in normal young adults. *Journal of Abnormal Psychology, 85,* 383–389.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Garret, H. E., & Schenk, M. R. (1928). A study of the discriminative value of the Woodworth Personal Data Sheet. *Journal of General Psychology, 1,* 459–471.

Gilbert, P. (1992). *Depression: The evolution of powerlessness*. New York: Guilford.

Gunthert, K., Cohen, L., Butler, A., & Beck, J. (2005). Predictive role of daily coping and affective reactivity in cognitive therapy outcome: Application of a daily process design to psychotherapy research. *Behavior Therapy, 36,* 77–88.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Jackson, S. (1986). Melancholia and depression: From Hippocratic times to modern times. New Haven, CT: Yale University Press.

Kranzler, H., Armeli, S., Feinn, R., & Tennen, H. (2004). Targeted naltrexone treatment moderates the relations between mood and drinking behavior among early problem drinkers. *Journal of Consulting and Clinical Psychology, 72,* 317–327.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept first- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

Nestler, E. J., Barrot, M., DiLeone, R. J., Eisch, A. J., Gold, S. J., & Monteggia, L. M. (2002). Neurobiology of depression. *Neuron, 34,* 13–25.

Neufeld, R. W. J. (1999). Dynamic differentials of stress and coping. *Psychological Review, 106,* 385–397.

Price, J. C. (1967). Hypothesis: The dominance hierarchy and the evolution of mental illness. *Lancet, 2,* 243–246.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B, 53,* 539–572.

Santor, D. A. (2000). The Center for Epidemiologic Studies Depression Scale. In A. E. Kazdin (Chief Ed.), *Encyclopedia of Psychology.* Washington, DC: American Psychological Association.

Santor, D. A. (2005). Using and evaluating psychometric measures. In J. Miles & P. Gilbert (Eds.), *Handbook of research methods for clinical and health psychology* (pp. 95–101). Oxford, England: Oxford University Press.

Santor, D. A., & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9,* 233–243.

Santor, D. A., De Brota, D., Gelwicks, S., & Englehardt, N. (2006). Optimizing the ability of the Hamilton Depression Rating Scale to discriminate across levels of severity and between antidepressants and placebos. Manuscript submitted for publication.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359.

Santor, D. A., & Segal, Z. V. (2001). Rates of change in cognitive behavior therapy for depression: Predicting symptom return over a 12-month follow-up. *Cognitive Therapy and Research, 25,* 117–135.

Santor, D. A., Zuroff, D. C., & Fielding, A. (1997). Analysis and revision of the DEQ: Examining scale performance as a function of scale length. *Journal of Personality Assessment, 69,* 145–163.

Santor, D. A., Zuroff, D. C., Ramsay, J. O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment, 7,* 131–139.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Sternberg, R. J. (2004). The role of biological and environmental contexts in the integration of psychology : A reply to Posner and Rothbart. *Canadian Psychology, 45,* 279–283.

Sternberg, R. J., & Grigorenko, E. L. (2001). Unified psychology. *American Psychologist, 56,* 1069–1079.

Tennen, H., Affleck, G., Armeli, S., & Carney, M. A. (2000). A daily process approach to coping: Linking theory, research and practice. *American Psychologist, 55,* 626–636.

Tennen, H., Affleck, G., Coyne, J. C., Larsen, R. J., & DeLongis, A. (in press). Paper and plastic in daily diary research. *Psychological Methods*.

U.S. Surgeon General's Office. (1999). *Mental Health: A Report of the Surgeon General*. Washington, DC: Department of Health and Human Services, U.S. Public Health Service.

Viglione, D. J., Lovette, G. J., & Gottlieb, R. (1995). Depressive Experiences Questionnaire: Exploration of the underlying theory. *Journal of Personality Assessment, 65,* 91–99.

Welkowitz, J., Lish, J. D., & Bond, R. N. (1985). The depressive experiences questionnaire: Revision and validation. *Journal of Personality Assessment, 49,* 89–94.