Psychological Bulletin
2011, Vol. 137, No. 5, 856–879

© 2011 American Psychological Association
0033-2909/11/$12.00 DOI: 10.1037/a0023678

The Reliability and Validity of Discrete and Continuous Measures of Psychopathology: A Quantitative Review

Kristian E. Markon and Michael Chmielewski
University of Iowa

Christopher J. Miller
University of Minnesota

In 2 meta-analyses involving 58 studies and 59,575 participants, we quantitatively summarized the relative reliability and validity of continuous (i.e., dimensional) and discrete (i.e., categorical) measures of psychopathology. Overall, results suggest an expected 15% increase in reliability and 37% increase in validity through adoption of a continuous over discrete measure of psychopathology alone. This increase occurs across all types of samples and forms of psychopathology, with little evidence for exceptions. For typical observed effect sizes, the increase in validity is sufficient to almost halve sample sizes necessary to achieve standard power levels. With important caveats, the current results, considered with previous research, provide sufficient empirical and theoretical basis to assume a priori that continuous measurement of psychopathology is more reliable and valid. Use of continuous measures in psychopathology assessment has widespread theoretical and practical benefits in research and clinical settings.

Keywords: reliability, validity, meta-analysis, psychiatric classification, diagnosis

Whether or not psychological constructs should be treated as discrete (e.g., in terms of discrete disease states or diagnoses) or continuous (e.g., in terms of spectra or trait levels) is an enduring question in psychology, especially in the area of psychopathology (Eysenck, 1970; Flett, Vredenburg, & Krames, 1997; Gangestad & Snyder, 1985; Lewis, 1938; Meehl, 1992; Pickles & Angold, 2003). The issue has arguably attracted increased attention in recent years, possibly due to methodological advances that have facilitated empirical inquiry into the question (e.g., Markon & Krueger, 2006; Meehl & Yonce, 1994, 1996; Ruscio, Ruscio, & Meron, 2007), as well as anticipated changes to official psychiatric nomenclature (e.g., Helzer et al., 2008; Krueger, Markon, Patrick, & Iacono, 2005; Widiger & Samuel, 2005). Addressing these questions has important implications for a number of areas of psychology, including psychopathology theory, general assessment, and clinical practice.

One specific issue in this area is whether observed measures of psychopathology per se are best treated as discrete or continuous.[1] That is, independent of the nature of the underlying constructs, how do observed measures of psychopathology compare in their reliability and validity? Our purpose in this paper is to inform this discussion by quantifying the relative reliabilities and validities achieved by discrete and continuous measures of psychopathology and to examine conditions under which the relative performance of the two forms of assessment might differ. As has been noted by various authors (Watson, 2003), the answer to whether psychopathology is best assessed with discrete or continuous measures ultimately rests on the relative empirical performance of the two paradigms. Quantifying the relative reliabilities and validities of the two forms of measurement across different conditions has important applied implications for researchers and clinicians across a number of areas.

Comparing Discrete and Continuous Measures of Psychopathology

Comparisons of discrete and continuous measures of psychopathology have received increased attention in recent years, in part because of the development of methods for empirically doing so. Application of these methods has raised awareness of the importance of distributional considerations in conceptualizing and assessing psychopathology and of the possibility of appealing to empirical evidence in adjudicating between discrete and continuous paradigms. Although these methods were developed to address questions about latent distributions of constructs, interest in these issues has subsequently extended to questions about observed distributions of measures as well.

This article was published Online First May 16, 2011.

Kristian E. Markon and Michael Chmielewski, Department of Psychology, University of Iowa; Christopher J. Miller, Division of Biostatistics, University of Minnesota.

Correspondence concerning this article should be addressed to Kristian E. Markon, Department of Psychology, University of Iowa, Iowa City, IA 52242. E-mail: kristian-markon@uiowa.edu

[1] It is important to note that the terms *discrete* and *continuous* can be vague in meaning. Throughout the paper, it is assumed that a discrete measure is a dichotomous, binary, or diagnostic measure taking two values; a continuous measure is a count or interval-scaled measure or an ordinal measure having several values. We acknowledge that use of these terms in this way is imperfect, but it is nevertheless appropriate for at least two reasons. First, most discussions of discreteness and continuousness in psychopathology research and theory focus on these cases, especially given diagnostic conventions in official nosologies. Second, given the statistical literature on this topic, we believe that major conclusions that are made on the basis of these definitions will generalize to more nuanced cases, with minor caveats.

Statistical Tests of Latent Discreteness Versus Continuousness

Much of the recent empirical work on discrete versus continuous assessment of psychopathology can be traced to the development of taxometric methods by Meehl and others (e.g., Meehl & Golden, 1982; Ruscio et al., 2007; Waller & Meehl, 1998). These methods are based on the premise that discrete underlying groups will induce discontinuities in the moments (e.g., means and covariances) of observed measures, which can be used to identify discreteness (for overviews, see Beauchaine, 2007; Waller & Meehl, 1998). A large literature on taxometrics has developed, with results expectably varying across constructs and studies. Quantitative, meta-analytic summaries of the literature are rare, but qualitative reviews have been conducted, with conclusions that have varied by domain. The most comprehensive qualitative reviews (Haslam, 2003a, 2003b, 2007), for example, have concluded that the strongest evidence for continuity has been found for certain forms of internalizing psychopathology (e.g., social anxiety and posttraumatic stress disorder) and the strongest evidence for discreteness has been found for forms of dissociation, antisociality, and schizotypy. Beauchaine (2007) highlighted the importance of methodological considerations in reviewing the taxometrics literature, noting, for example, that although eating disorders have appeared taxonic in nature in some studies, they behave dimensionally when appropriate sampling is used.

Another approach to comparing discrete and continuous assessments of psychopathology is statistical comparison of models in which the latent variable is discrete to models in which the latent variable is continuous (Lubke & Neale, 2006; Markon & Krueger, 2005; Schmitt, Mehta, Aggen, Kubarych, & Neale, 2006). In contrast to taxometric approaches, these latent variable models require the specification of explicit probabilistic models relating the observed variables to latent variables. Research has demonstrated that these models can be directly compared and successfully distinguished using likelihood-based methods (Lubke & Neale, 2006; Markon & Krueger, 2005; Schmitt et al., 2006). Because these methods are newer, fewer studies have adopted this approach in comparisons between discrete and continuous measures of psychopathology. Comparisons of discrete and continuous latent variable models have indicated, however, that externalizing psychopathology (e.g., substance use, antisociality) is best modeled in terms of an underling continuous normal distribution (Markon & Krueger, 2005). Latent variable models freely estimating the distribution of depression have suggested, similarly, that it is continuous and roughly normal in shape (Schmitt et al. 2006).

Psychometric Properties of Discrete Versus Continuous Measures

As interest in explicit comparisons between discrete and continuous models of psychopathology has increased, so has interest in the relative psychometric properties of discrete and continuous measures. Although questions about the properties of latent variables are ultimately distinct from questions about observed measures of those variables (e.g., it is possible to construct discrete observed measures of continuous latent variables and vice versa), they are related in recognizing that the optimal representation of a construct, either at the latent or the observed level, might be discrete or continuous. The use of methods such as those just described, although focused on latent properties of constructs, has arguably increased awareness of the fact that empirical considerations can also be used to evaluate assumptions about the optimal scale for observed measures of psychopathology.

Reliability. Reliability has been and continues to be an integral criterion for evaluating the appropriateness of psychopathology measures and assessment devices (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; American Psychiatric Association, 2006; Hunsley & Mash, 2007). Historically, reliability has been a major driver of changes in official psychiatric nosology, with the emphasis on reliability—especially interrater reliability—in the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.; *DSM–III*; American Psychiatric Association, 1980) and its descendants representing a major shift in research and practice in the area of psychopathology assessment (Compton & Guze, 1995). With upcoming revisions to official nomenclature in the form of the *DSM–5* and *ICD–11,* issues surrounding the role of reliability in assessment of psychopathology continue to be a major focus of discussion (Hyman, 2010).

Many researchers have commented on discrete versus continuous scales of measurement in discussing the reliability of psychopathology assessments. Baca-Garcia et al. (2007), for example, observing widespread low levels of diagnostic stability in a large study of clinical practice settings, suggested that low reliability of diagnoses might in part result from the use of discrete diagnostic criteria that fail to recognize continuous variation in patients' presentations. Challenges surrounding how to represent continuous variation in psychopathology over time using discrete diagnoses have been central in diagnostic theory and practice in various areas of psychopathology. The *DSM–5* Neurocognitive Disorders Workgroup, for example, has proposed replacing the diagnosis of dementia with two diagnoses, major neurocognitive disorder and minor neurocognitive disorder, in part to better reflect longitudinal trajectories in cognitive functioning (*DSM–5* Neurocognitive Disorders Work Group, 2010; see also Mitchell & Shiri-Feshki, 2009; Ritchie & Touchon, 2000).

More broadly, issues surrounding the reliability of continuous and discrete measures are critical to the validity of assessments, at the individual or group level. Messick (1995), for example, suggested that construct validity be considered broadly, noting that "communality among . . . indicators [is] taken to imply the operation of the construct to the degree that discriminant evidence discounts the intrusion of alternative constructs as plausible rival hypotheses" (p. 742). According to this view, invalidity derives from construct-irrelevant variance in measures, where "the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance" (p. 742). Low reliability, in this paradigm, can be seen as a direct source of invalidity, in that unreliable measures comprise variance not directly related to the construct of interest. For example, low interrater reliability represents a direct threat to the validity of an assessment, in that it implies a substantial portion of the variance in someone's observed standing reflects the particular assessor or clinician, rather than the individual's actual level of psychopathology (for a similar perspective, see Smith, 2005). To the extent that discrete or continuous measures of psychopathology systematically differ in their reliability, they will differ in their validity as well.

Validity. The use of discrete versus continuous indicators of psychopathology may have broader implications for construct validity, beyond the effects of reliability per se. Issues surrounding optimal scales of measurement and of the role of discrete versus continuous indicators in particular are often intimately tied to considerations of construct validity and related concepts in the psychopathology literature (e.g., Kendell & Jablensky, 2003; Robins & Guze, 1970). At a very abstract level, the scale of measurement used to assess a psychopathology construct can be seen as part of the theoretical framework (Landy, 1986; Smith, 2005) or nomological network (Cronbach & Meehl, 1955) surrounding that construct. Whether a construct should be assessed with discrete or continuous measures comprises another aspect of the construct's theory and can be evaluated in terms of basic standards of construct validation (Waldman & Lilienfeld, 2001).

Framing the process of construct validation in terms of hypothesis testing (Landy, 1986), one can frame discussions about the use of discrete versus continuous measures in this way: Which form of measurement provides the greatest explanatory and statistical power? To the extent that the most meaningful variability in an observed measure is represented by discrete classes of psychopathology, class assignment should be reliable and should demonstrate theoretically meaningful criterion-related validity. Moreover, to the extent that discrete and continuous measurement paradigms have comparable construct validity, they should demonstrate comparable reliability and validity. Watson (2003), for example, provided evidence that observed discrete measures may have relatively low retest reliability relative to continuous measures, even when formal tests of latent discreteness suggest the presence of underlying groups.

It is important to note in this regard that the measurement properties of an observed measure are only one part of the theory concerning the construct. Discovering that a continuous measurement scale is more optimal than a discrete measurement scale does not necessarily invalidate the notion that the underlying construct is discrete in some way. In fact, such findings might be used to refine and expand the theory regarding the nature of the latent discreteness. Lenzeweger, McLachlan, and Rubin (2007), for example, have proposed the use of mixture models to represent classes of schizotypic risk. In this type of model, observed measures of schizotypy reflect latent mixtures of continuous subpopulations, with valid continuous variability existing within classes. Under such a model, finding that continuous measures are more reliable and valid would not invalidate the notion that there are discrete liability classes but would instead reinforce the notion that there is valid variation within classes that cannot be ignored at the observed level. This can be compared to general cognitive ability, which is generally accepted to be continuously distributed, even though discrete liability classes can be identified, each with observable etiologies (e.g., environmental or genetic agents of large effect). Such a framework would help explain observations that schizotypy and related traits share features of continuous as well as discrete distributions (Linscott & van Os, 2010).

The relative explanatory and statistical power of discrete and continuous measures as revealed by reliability and validity studies is also important from a practical perspective. As Kendell and Jablensky (2003) noted, independent of any theoretical considerations, measures of psychopathology have important practical, clinical utility in their ability to provide information about status, prognosis, and associations with other variables. Even if one ignores issues surrounding the validity of the constructs as instantiated in a measurement theory, the measures themselves have quantifiable utility in a purely statistical sense. Much of official psychiatric nosology comprises discrete indicators in the form of diagnoses, a measurement assumption that affects the provision of clinical care as well as research on psychopathology assessment, description, etiology, and outcome. Understanding how discrete and continuous measures differ in their reliability and validity, in terms of their statistical power, helps frame an understanding of how their use affects clinical and research practice. This issue has particular import currently, given that revisions to official nomenclature, under development, have emphasized greater inclusion of continuous measurement elements (e.g., Regier, 2007).

Evidence Regarding the Reliability and Validity of Discrete and Continuous Psychopathology Measures

Given the importance of how discrete and continuous measures compare in terms of their reliability and validity, it is not surprising that substantial literatures about this issue have developed, within the methodological as well as applied communities. A brief overview of these two literatures is informative in understanding why uncertainty surrounding the reliability of discrete and continuous measures of psychopathology persists.

Methodological Considerations

Much of the relevant methodological literature on the issue derives from work on the practice of discretization, or creation of discrete indicators from continuous measures. In general, work on this issue has suggested that discretizing a continuous variable reduces its expected correlations with other variables, whether those are measures of different constructs or measures of the same construct at different times. The reduction in the size of the correlation, moreover, can be predicted from the proportion of cases being assigned to each group (e.g., Kraemer, 1979; MacCallum, Zhang, Preacher, & Rucker, 2002).

In light of these conclusions, DeCoster, Iselin, and Gallucci (2009) recently reviewed putative rationales for using discrete versus continuous measures and evaluated them in a series of simulation studies. Overall, DeCoster et al. concluded that continuous measures generally produced greater correlations and are to be preferred a priori, except in three cases, where a discrete measure might be equally acceptable: first, when the intent is to evaluate the performance of a discrete measure; second, when using extreme groups analysis; and third, when the discrete measure reflects true underlying discreteness, is reliable, and mirrors the actual distribution of the latent variable.

This last scenario arguably is the primary reason why confusion about discrete versus continuous assessment of psychopathology persists. The first scenario is somewhat self-evident, and the second arguably represents a specialized application requiring continuous assessment at some point. As many psychopathology constructs have traditionally been thought of in terms of disease or illness states, however, they raise the third possibility: that a discrete measure might demonstrate greatest reliability and validity when it is assessing discrete constructs in a way that appropriately reflects the distribution of the underlying groups. The converse is also sometimes assumed: that discrete constructs might demonstrate greatest validity when assessed with discrete indicators (assuming that all within-class variance is error; cf. Ruscio & Ruscio, 2002).

Although common, this last set of assumptions—that discrete psychopathology constructs are best assessed through discrete indicators or vice versa—may not always be well justified. That is, although it might initially seem somewhat paradoxical, even when latent constructs are discrete, assessing them using continuous indicators is often better from a psychometric perspective. This conclusion is arguably supported by the results of DeCoster et al. (2009), as they found that even when underlying variables were truly discrete, there were many conditions in which discrete manifest indicators performed worse than continuous indicators. Similarly, it has been shown that the power to directly detect latent mixtures and classes generally increases when observed indicators are continuous rather than discrete (Lubke & Neale, 2008), or, similarly, when indicators increase in the precision of their scale of measurement (Markon & Krueger, 2006). Continuous measures generally reflect more information about the nature or level of the latent variable (e.g., distribution in a population, trait level in an individual) than discrete measures, in that they offer more observed values over which information about the latent variable is sampled.

Empirical Evidence

The practical importance of knowing how discrete and continuous versions of psychopathology measures perform, together with the theoretical implications of their relative performance, has led to a substantial body of empirical literature on the relative reliability and validity of psychopathology measures. Many of these studies have been more pragmatic in motivation, seeking to document characteristics of discrete and continuous versions of a measure (e.g., Grant et al., 2003) or to compare alternative measures available (e.g., Skodol, Oldham, Rosnick, Kellman, & Hyler, 1991). Other studies have been motivated more by an interest in what is implied about the underlying constructs (e.g., Prisciandaro & Roberts, 2009).

There have been a small number of quantitative reviews of this literature, which have been mixed in their conclusions. Clark (1999), for example, reviewing measures of personality pathology, found results consistent with most methodological work, in that median correlations involving discrete personality disorder measures were lower than correlations involving their continuous counterparts. These results were consistent with an earlier, more qualitative review by Widiger (1992) finding that all studies considering discrete and continuous measures of personality disorder produced results favoring continuous measures. Moreland and Dumas (2008), in contrast, found that discrete and continuous measures of disruptive preschool behavior provided comparable reliability and validity, with those of discrete assessments actually being somewhat larger in some cases.

Rationale and Goals of the Current Work

Given the practical and theoretical importance of the relative reliabilities and validities of discrete and continuous psychopathology measures, we sought to quantify typical values encountered empirically and to examine possible moderators of these values. Doing so would help provide a benchmark of what to expect in clinical and research settings and might help inform discussion about the role of the two types of measures in psychopathology. We conducted two meta-analyses, one of reliability and the other of validity, and focused in both meta-analyses on studies that have compared discrete and continuous measures of the same construct in the same samples. By focusing on studies comparing discrete and continuous measures of the same construct in the same samples, we were able to examine reliability and validity estimates in two ways: first by considering the estimates individually and second by examining differences within pairs of discrete and continuous estimates. This latter approach allows for a relatively direct estimate of the effect of using a discrete versus continuous measure of psychopathology.

Meta-Analysis 1: Reliability

Method

Samples. Numerous studies have reported reliabilities of psychopathology measures, either continuously or categorically assessed. We therefore limited the meta-analysis to include only those studies reporting reliabilities of continuous and discrete measures of the same constructs in the same samples. We also included only studies reporting reliabilities of measures as used to assess actual empirical samples of individuals (i.e., studies reporting reliabilities using vignettes or case prototypes were excluded).

Studies were located by searching for relevant search terms (e.g., *continuous and categorical, reliability, test–retest*) in PsycINFO, Google, Google Scholar, and PubMed. Works citing potential studies were also searched, as were the references of potential studies. Finally, the references of chapters and review papers on categorical and continuous assessment of psychopathology were searched for empirical papers.

Ultimately, 488 effect size estimates from 31 studies, representing a total of 4,200 participants, were included in the meta-analysis. Of these studies, 14 included a test–retest design component, nine included the review of interview protocols (e.g., audio or video recordings, or left unspecified), eight included live observation of interview, four included joint interviews, and one involved independent raters with knowledge of the target. Four studies included clinical samples, and eight studies included nonclinical samples. Of the studies, 27 were conducted in English, two in Dutch, one in German, and one in Italian. References for the final list of studies are given in Appendix A.

Effect size estimates and moderators.

Effect size estimates. Nearly all studies reported kappas as reliability statistics for categorical measures, with the exception of one study that reported a Pearson correlation (Watson, 2003). Although most studies reported intraclass correlations (ICCs) as reliability statistics for continuous measures, a large number of studies reported Pearsons for this purpose. Although kappas and ICCs are directly comparable (Fleiss & Cohen, 1973), this is not the case with Pearson correlations. Therefore, we empirically estimated what the ICCs would have been for studies reporting Pearson correlations.

This was done by fitting a local likelihood regression model (Tibshirani & Hastie, 1987) to Monte Carlo simulation data designed to reflect actual observed study characteristics. Ten thousand simulated ICC–Pearson pairs were randomly generated under bivariate normal population models. The characteristics of the populations (e.g., variance components due to rater and target) and characteristics of the samples (e.g., sample sizes) were randomly selected from parameters from five studies where these values were available or could be calculated (Blanchard, Horan, & Collins, 2005; Ferro, Klein, Schwartz, Kasch, & Leader, 1998; Nazikian, Rudd, Edwards, & Jackson, 1990; Watson, 2003; Zimmerman, Pfohl, Coryell, Stangl, & Corenthal, 1988). The local likelihood regression model fits to these 10,000 ICC–Pearson pairs were used to predict ICC values in subsequent analyses for those studies originally reporting Pearsons. The predicted ICCs were generally very similar to the Pearsons, especially for smaller values of the Pearsons, albeit almost uniformly smaller.[2]

Moderators. Two possible moderators of reliability were examined: whether the sample was clinical or nonclinical and what type of construct was being measured. Samples comprising patients or individuals selected directly on the basis of elevated psychopathology were classified as clinical; other samples were classified as nonclinical. Constructs were classified according to their relationship with four higher order psychopathology factors as reported in previous literature (Krueger & Markon, 2006; Markon, 2010b): internalizing (e.g., depression), externalizing (e.g., antisocial personality disorder, substance use disorders), thought disorder (e.g., schizophrenia), and pathological introversion (e.g., avoidant personality disorder). A category of other was created for constructs not clearly falling into those categories (e.g., personality disorder not otherwise specified, inappropriate sexual behavior, dementia). In addition, in models of individual effect sizes used to estimate average reliabilities, two additional moderators were examined: the continuous versus discrete nature of the measure and the time between assessments. The time between assessments was coded in days; assessments based on recorded material (e.g., video or audio recordings) were treated as occurring at the same time.

Finally, to examine how reliabilities differed between continuous and discrete measures across different types of reliability, we conducted paired effect size analyses separately for two types of reliability. The first form of reliability, which we refer to as single-occasion reliability, results from studies in which two different measurements derive from a single assessment occasion, with the different measurements being based on the same behavioral information. Examples of this include interrater reliability studies in which two raters use the same audio or video recordings, studies involving interview-observer sessions, or other scenarios in which the two measurements are of the same behavior. The second form of reliability, which we refer to as dual-occasion reliability, results from studies in which the two different measurements derive from two assessment occasions, with the different measurements being based on different behavior. Examples of this include studies incorporating a test–retest component and studies of different informants using the same measure to make ratings on a single target. Among the studies, 16 provided single-occasion reliabilities, 11 provided dual-occasion reliabilities, and four provided both.

Analyses. Two approaches were adopted in analysis: analyses of the effect sizes per se and analyses of the differences between the paired effect sizes obtained using discrete and continuous measures. Modeling of individual effect sizes allowed for estimation of typical reliabilities observed across studies. Moreover, as each study included two estimates of a given effect size—one using discrete measurement and another using continuous measurement—it was possible to obtain relatively direct estimates of the effect of type of assessment and possible moderators by analyzing differences within pairs.

Given the variety of measures, samples, and constructs examined, significant heterogeneity in effect size estimates was assumed a priori. As such, effect size estimates and moderators were modeled with mixed-effect regression, which allows for estimation of overall fixed population effects as well as random effects associated with each study. In all analyses, effect size estimates were modeled as nested within study. Mixed-effect models were estimated with the lme4 package (Bates & Sarkar, 2006) for R (R Development Core Team, 2010), with hybrid Monte Carlo (MC) bootstrap and permutation methods used to construct confidence intervals and perform hypothesis tests on possible moderators.

Analysis of individual effect sizes. In analyses of individual effect sizes, each effect size estimate was weighted by the inverse of its estimated variance, with variances for effect size estimates used as reported in Bloch and Kraemer (1989) and Bonett (2002). As variances for kappa estimates depend on the base rates of the variables, study-specific base rates were used when possible; when these were not available, mean base rates in other studies of the same construct were used. In cases where no other studies of the construct were available, the mean base rate over all constructs was used.

Confidence intervals for the coefficients of fixed effect terms were computed with a hybrid MC–bootstrap procedure. In this approach, 10,000 bootstrap replications were used to calculate confidence intervals. In each bootstrap replication, however, predicted ICCs were randomly imputed from Pearsons using the prediction model described above, to simulate the effects of estimating ICCs from Pearson correlations in the data.

Analysis of paired effect sizes. In addition to analyzing the effect sizes individually, we conducted analyses on the within-study differences between continuous and discrete effect size estimates. Differences between continuous and discrete reliability estimates were calculated for each pair, and these differences were modeled with mixed-effect models as just described. Two mixed-effect models were examined: a model of the binomial probabilities of the continuous estimate being greater than the discrete estimate and a model of the magnitude of the differences between

[2] We explored the use of other effect size metrics in analyses and found that estimates were similar regardless of the metric used. For example, in contrast to the continuousness effect of .171 reported in Table 1, a value of .131 was produced by transforming the kappas and ICCs to Pearsons by adjusting for base rate or mean-level differences in assessments. We reported analyses on a kappa–ICC metric because most studies reported results in this format. Although other ICC metrics might be used (e.g., nonparametric ICCs; Rothery, 1979), we believe this metric is standard in the psychopathology literature, familiar to most in the field, and relatively well understood statistically.

the two estimates. In each model, the study sample sizes were used as weights.

Tests of moderators were performed in two ways: through MC–bootstrap evaluations of the log-likelihood values under different models and through MC–permutation tests of likelihood ratio statistics. The former provided an evaluation of whether each term added significantly to the modeling of differences in effect sizes; the latter provided an evaluation of whether removing each term significantly decreased the performance of the models.

MC–bootstrap *p* values of log-likelihoods were calculated for each moderator by first obtaining the MC–bootstrap distribution of the log-likelihood under a model including only an intercept, using the hybrid MC–bootstrap procedure described above. The log-likelihoods of models with each moderator added were then compared to this MC–bootstrap distribution to obtain *p* values for the moderator terms, to determine whether the log-likelihood significantly increased relative to a model with only an intercept. In addition, the significance of the intercept was evaluated by randomly changing the direction of the difference between continuous and discrete assessments in each bootstrap replication, to simulate a scenario in which there is no difference between continuous and discrete assessments.

Hybrid MC–permutation likelihood ratio tests (LRTs) were conducted by evaluating the distribution of the LRT under a null model. In each of 10,000 permutation replications, each effect size difference was randomly reassigned to a different study, which eliminated any association between moderators and the effect sizes (Higgins & Thompson, 2004). In addition, as in the MC–bootstrap simulations, in each permutation replication, predicted ICCs were randomly imputed from Pearsons with the prediction model described above. LRT statistics were calculated for a full model including intercept and both moderator terms across permutation replications, simulating the distribution of the LRT statistics under the null hypothesis.

Results

Analyses of individual effect sizes. Parameter estimates from the mixed-effect model of individual effect sizes are presented in Table 1, and corresponding meta-analytic reliability estimates are shown in Table 2. Values in Table 1 represent fixed effect parameters coded relative to a discrete measure of externalizing in a nonclinical sample. It is important to note that the parameter estimates and confidence intervals presented in Table 1 should not be used to make inferences about the significance of any particular effect: They reflect inferences about a specific coefficient, not the term as a whole (e.g., the apparent significances of coefficients will change with recoding of the predictors). Moreover, they reflect a model where all terms in the regression are assumed to be valid predictors and therefore likely to overstate the statistical significance of some of the effects. These parameter estimates and their bootstrap confidence intervals are nevertheless presented to provide completeness and to give a general sense of the direction of the effects and the degree of uncertainty associated with each estimate. For example, consistent with expectations, reliabilities generally decreased with longer test–retest intervals, and reliabilities were greater overall in clinical samples, where pathological behaviors targeted by the measures were more likely to be observed.

Table 1
Reliability Meta-Analysis: Mixed-Effect Modeling of Individual Effect Sizes

		95% CI	
Effect	β	Lower	Upper
Intercept	.6897	.6066	.7466
Type of measure			
Discrete	.0000		
Continuous	.1708	.1076	.2539
Type of construct			
Externalizing	.0000		
Internalizing	−.0126	−.0638	.0443
Pathological introversion	−.0091	−.0577	.0375
Thought disorder	−.0486	−.1273	.0061
Other	−.0107	−.2188	.1381
Type of sample			
Nonclinical	.0000		
Clinical	.1273	.0443	.1942
Retest interval (days)	−.0005	−.0006	−.0004
Interaction: Measure × Construct			
Continuous externalizing	.0000		
Continuous internalizing	.0073	−.0503	.0620
Continuous pathological introversion	.0230	−.0246	.0768
Continuous thought disorder	.0224	−.0377	.1007
Continuous other	.0279	−.1490	.2461
Interaction: Measure × Sample			
Continuous nonclinical	.0000		
Continuous clinical	−.0966	−.1817	−.0234

Note. Parameter estimates are fixed-effect terms from the mixed-effect models of individual effect sizes, as described in the text. Confidence intervals are based on Monte Carlo–bootstrap procedures described in the text. Effects are generally coded relative to a baseline condition, represented by the intercept, of a discrete measure of externalizing used in a nonclinical sample.

Estimated reliabilities in Table 2 are based on the fixed effect terms of the mixed-effect model and reflect overall effects controlling for random effects due to study. As is evident, the reliabilities are significantly greater for the continuous measures than for discrete measures across all observed study characteristics. Overall, the estimated reliabilities under the model suggest that compared to use of discrete measures, use of continuous measures increases reliability by approximately 15%.

Analyses of paired effect sizes. The observed pairs of continuous and discrete reliability estimates are plotted in Figure 1. In the figure, the size of the circles is proportional to the sample size used in the calculation of the estimates. The greater size of the continuous estimates is evident in the rightward shift of the points relative to the diagonal. The continuous and discrete estimates were also significantly correlated (Spearman's $\rho = .664$, $p < .0001$), although there are some relatively large continuous reliability estimates that have discrete estimates near zero.

In general, results of the mixed-effect modeling of differences between paired effect sizes indicated that continuous measures were more reliable than discrete measures and that this phenomenon was not significantly moderated by type of phenotype or sample (see Table 3). Continuous measures were significantly more reliable than their discrete counterparts, both in terms of the probability of a given continuous measure being more reliable than its discrete counterpart ($p = .000$) and in terms of magnitude of the

Table 2
Reliability Estimates Based on Mixed-Effect Modeling of Individual Effect Sizes

Parameter	Overall	Discrete	Continuous
Overall	.766	.711	.820
Type of construct			
Externalizing	.768	.719	.817
Internalizing	.774	.727	.821
Pathological introversion	.795	.736	.853
Thought disorder	.749	.686	.812
Other	.691	.640	.742
Type of sample			
Nonclinical	.738	.646	.830
Clinical	.774	.731	.818

Note. Reliability estimates are predicted values of the model based on fixed-effect parameters from Table 1.

difference (p = .000). The type of psychopathology construct did not significantly affect the probability of a continuous measure being more reliable than a discrete measure (MC–bootstrap p = .430; MC–permutation p = .372), nor the magnitude of the difference (MC–bootstrap p = .410; MC–permutation p = .348). Similarly, type of sample did not significantly affect the probability of a continuous measure being more reliable than a discrete measure (MC–bootstrap p = .498; MC–permutation p = .519), nor the magnitude of the difference, although this latter effect was nearly significant (MC–bootstrap p = .440; MC–permutation p = .058).

Paired effect size analyses conducted separately for single- and dual-occasion reliabilities are presented in Tables 4 and 5, respectively. As can be seen, the pattern of results for each form of reliability was similar to the pattern of findings for the reliabilities considered together in that the type of construct did not appear to moderate the probability of a continuous measure being more reliable than a discrete measure or the magnitude of the difference. Sample type, in contrast, did significantly moderate the magnitude of the difference but only for single-occasion reliabilities (for single-occasion reliabilities, MC–bootstrap p = .364 and MC–permutation p = .007; for dual-occasion reliabilities, MC–bootstrap p = .483 and MC–permutation p = .328). The magnitude of the difference between continuous and discrete reliabilities was significantly smaller in clinical samples (with estimated discrete and continuous reliabilities of .687 and .895, respectively) than in nonclinical samples (with estimated reliabilities of .826 and .909) but only among single-occasion studies. Sample type did not moderate the probability of observing a continuous reliability greater than a discrete reliability for either type of reliability.

In general, the pattern of results considered separately by type of reliability suggests that the nearly significant effect of sample type among all studies considered together was largely due to studies reporting single-occasion reliabilities (e.g., interrater reliability studies). It is important to note, however, that the continuous–discrete difference was much smaller in dual-occasion studies than in single-occasion studies (with intercepts of .070 and .364, respectively) and that dual-occasion reliabilities were smaller than single-occasion reliabilities (with overall estimated reliabilities of .686 and .847, respectively). The lack of effect of sample type among dual-occasion reliabilities may represent a sort of floor effect in this regard; conversely, the greater ceiling among single-occasion reliabilities may provide more opportunity for the effect to manifest.

With the effect estimates reported in Table 3, it was possible to calculate overall model-implied differences in continuous and discrete reliabilities. For example, the overall model-implied probability of observing a continuous measure with greater reliability than a discrete measure was .679. Similarly, the model-implied overall difference in reliabilities between continuous and discrete measures was approximately .141, slightly larger than what is implied by the estimated reliabilities in Table 2.

Trends by construct. To explore the possibility that specific constructs (e.g., alcohol use problems) might demonstrate evidence of equal reliability for continuous and discrete measures even though superordinate groupings of constructs (e.g., externalizing) do not, we examined the probability of observing a continuous measure reliability greater than a discrete measure reliability by specific construct. We found that the probability of observing a greater reliability with one form of measure or the other strongly depended on the number of replications of findings involving the construct. This trend is illustrated in Figure 2; as can be seen, as the number of effect sizes reported on a construct increased, the probability of observing greater reliabilities with discrete measures decreased.

Discretization models. As noted by multiple authors (Kraemer, 1979; MacCallum et al., 2002), one can estimate the expected decrease in magnitude of a correlation as a result of discretizing an underlying continuous variable. Assuming an underlying bivariate normal distribution, using the original continuous correlation and

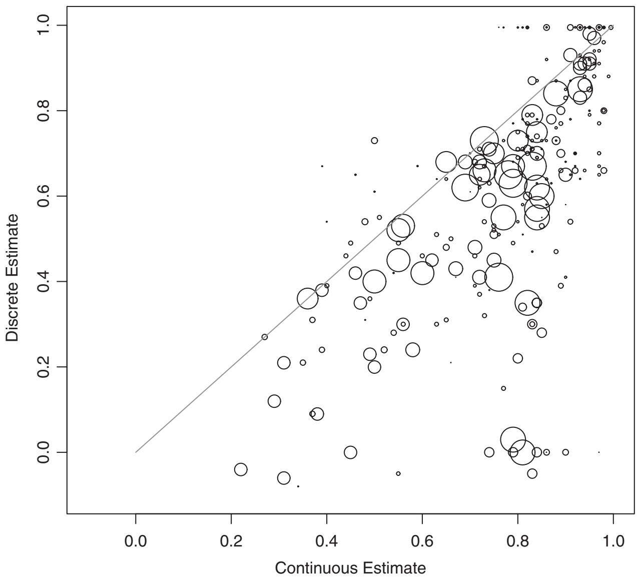

Figure 1. Scatter plot of reliability estimates; each point represents a pair of estimates, one using a continuous measure and the other using a discrete measure. The size of circles is proportional to the sample size used to calculate the estimate. The light solid line is drawn for a reference point and represents what would be expected if the continuous and discrete estimates were equal. To the extent points are lower than the line, continuous estimates are greater; to the extent points are above the line, discrete estimates are greater.

Table 3
Reliability Meta-Analysis: Mixed-Effect Modeling of Differences in Effect Sizes, All Reliabilities

						95% CI	
Effect	*lnL*	*p*	χ^2	*p*	β	Lower	Upper
Binomial model							
Intercept	−1,777.31	.000			.944	0.779	1.154
Type of construct	−1,733.23	.430	88.16	.372			
Externalizing					.000		
Internalizing					.011	−0.140	0.161
Pathological introversion					−.001	−0.133	0.119
Thought disorder					−.076	−0.204	0.039
Other					−.273	−0.717	0.124
Type of sample	−1,774.37	.498	5.88	.519			
Nonclinical					.000		
Clinical					−.215	−0.506	−0.017
Difference model							
Intercept	−521.42	.000			.227	0.171	0.296
Type of construct	−517.37	.410	4.264	.348			
Externalizing					.000		
Internalizing					−.019	−0.112	0.065
Pathological introversion					−.024	−0.106	0.064
Thought disorder					.046	−0.040	0.128
Other					−.082	−0.193	0.136
Type of sample	−519.29	.440	8.11	.058			
Nonclinical					.000		
Clinical					−.114	−0.185	−0.058

Note. Parameter estimates are fixed-effect terms from the mixed-effect models of differences within pairs of effect sizes, as described in the text. The binomial model modeled the probability of observing a continuous effect size greater than its discrete counterpart; the difference model modeled the magnitude of difference between continuous effect sizes and their discrete counterparts. *P* values for the log-likelihood and likelihood ratio χ^2 and confidence intervals for the effect parameters β are based on MC–permutation and MC–bootstrap procedures described in the text. Effects are coded relative to a baseline condition, represented by the intercept, of a measure of externalizing used in a nonclinical sample. MC = Monte Carlo.

the proportion of individuals in the dichotomized groups (or, equivalently, the points of discretization on the normal distribution), one can estimate the correlation involving the discretized variable. Kraemer provided these formulas for kappa in continuous distributions that have been discretized; MacCallum et al. provided these formulas for Pearson-family correlations.

We applied the formulas of Kraemer (1979) to the kappa–Pearson pairs, in order to determine how the observed discrete reliabilities compare to what would be expected under a simple discretization model. Sample-specific base rates, mean base rates for the constructs involved, or mean base rates observed in the meta-analysis were used as available; in cases where base rates were available separately for the two assessments, the first base rate was used. The results of these analyses are illustrated in Figure 3. As is evident in the figure, there was significant variability around the values predicted by a simple discretization model (variance of the difference = .034), possibly due to errors in the assumption of normality, assumed base rates, and the assumed underlying continuous reliability, as well as to errors in the discretization model itself. The mean observed–expected difference in discrete reliabilities was .099; the median difference was .147; the observed and expected values were significantly different according a Wilcoxon signed-rank test ($W = 4{,}323$; $p < .001$). Therefore, under the assumptions of the discretization model, the observed discrete reliabilities were actually typically somewhat larger than would be expected under discretization of an underlying normal distribution.

Discussion

Overall, the results of this meta-analysis underscore previous empirical and methodological findings that continuous measures are more reliable than discrete measures across a wide range of settings. Our meta-analysis extends these findings by demonstrating that this difference in reliability is invariant across different forms of psychopathology and is generally robust to changes in sample characteristics. In general, researchers have a 68% chance of observing greater test–retest reliabilities with continuous measures and can expect a 15% increase in the size of the reliabilities with use of continuous measures. Analyses of specific constructs suggest, moreover, that deviations from this trend for any given construct are likely to be due to sampling variation.

Analyses of differences between continuous and discrete measures by different types of reliability indicated that the advantage of continuous measures of psychopathology may be attenuated in clinical populations. However, this effect seemed limited to reliabilities in which the two assessments were based on the same behavioral information (e.g., interrater reliability) and did not generalize to cases where assessments were based on different behavioral information (e.g., test–retest reliability). These results suggest that discrete assessments of psychopathology are more consistent in cases where base rates of the target phenomena are greater and where the behavior being assessed is held constant across assessors. When the behavior is allowed to vary and when the phenomena of interest are less prevalent, the advantage of continuous

Table 4
Reliability Meta-Analysis: Mixed-Effect Modeling of Differences in Effect Sizes, Single-Occasion Reliabilities

Effect	*lnL*	*p*	χ^2	*p*	β	95% CI Lower	95% CI Upper
Binomial model							
Intercept	−703.82	.000			.978	0.698	1.257
Type of construct	−695.21	.466	17.22	.877			
Externalizing					.000		
Internalizing					−.034	−0.256	0.171
Pathological introversion					−.076	−0.283	0.107
Thought disorder					−.014	−0.170	0.150
Other					.361	0.112	0.628
Type of sample	−700.74	.483	6.15	.328			
Nonclinical					.000		
Clinical					−.313	−0.683	0.038
Difference model							
Intercept	−312.49	.003			.364	0.256	0.470
Type of construct	−306.41	.336	12.17	.185			
Externalizing					.000		
Internalizing					−.075	−0.210	0.058
Pathological introversion					−.085	−0.204	0.070
Thought disorder					.057	−0.073	0.179
Other					.151	0.067	0.241
Type of sample	−307.82	.364	9.33	.007			
Nonclinical					.000		
Clinical					−.243	−0.351	−0.134

Note. Parameter estimates are fixed-effect terms from the mixed-effect models of differences within pairs of effect sizes, as described in the text. The binomial model modeled the probability of observing a continuous effect size greater than its discrete counterpart; the difference model modeled the magnitude of difference between continuous effect sizes and their discrete counterparts. *P* values for the log-likelihood and likelihood ratio χ^2 and confidence intervals for the effect parameters β are based on MC–permutation and MC–bootstrap procedures described in the text. Effects are coded relative to a baseline condition, represented by the intercept, of a measure of externalizing used in a nonclinical sample. MC = Monte Carlo.

measures is greater. It should be noted, moreover, that even in clinical samples where behavioral information is held fixed across raters, continuous measures were more reliable, albeit less so.

Meta-Analysis 2: Validity

Method

Samples. As in the reliability meta-analysis, studies were located by searching for relevant search terms (e.g., *continuous and categorical, validity*) in PsycINFO, Google, Google Scholar, and PubMed. Works citing potential studies were also searched, as were the references of potential studies. Finally, the references of chapters and review papers on categorical and continuous assessment of psychopathology were searched for empirical papers. In keeping with the search methodology of the reliability meta-analysis, only those studies reporting continuous and categorical validity measures of the same constructs in the same samples were included.

In total, 652 effect size estimates from 27 studies, representing a total of 55,375 participants, were included in the meta-analysis. Twenty studies included clinical samples, and seven studies included nonclinical samples. Twenty-three studies were conducted in English, and one study each was conducted in Spanish, Dutch, Chinese, and Italian. References for the final list of studies are given in Appendix B.

Effect size estimates and moderators.

Effect size estimates. Most studies reported discrete estimates using Pearsons, multiple correlations, kappas, or R^2, although odds ratios, log odds, phi correlations, and beta coefficients were also reported. Most studies reported continuous estimates using Pearsons, R^2 statistics, and multiple correlations, although odds ratios, phi correlations, log odds, and ICCs were reported as well. To put the kappas, ICCs, and Pearsons on a comparable scale, we transformed the kappas to estimated Pearsons using methods described in Bloch and Kraemer (1989, p. 279; cf. Cohen, 1960, and Maxwell, 1977, Equations 1a and 3). Kappas were adjusted for base rate or mean-level differences in assessments using study-specific base rates corresponding to the estimate; when these were not available, we used the mean base rate of the phenotype in other studies or, when that was not available, the mean base rate across studies.[3]

Moderators. As in the reliability meta-analysis, two moderators of validity were tested: which type of construct was measured and whether the sample was clinical or nonclinical in composition. Constructs were classified as in the reliability meta-analysis, according to their relationship with four higher order psychopathology factors—internalizing, externalizing, thought disorder, and pathological introversion—as reported in previous literature

[3] As in the reliability meta-analysis, we explored the use of other effect size metrics in analyses and found again that estimates were similar regardless of the metric used. For example, in contrast to the continuousness effect of .088 reported in Table 6, a value of .096 was produced by using ICCs predicted from Pearsons. We reported analyses on a Pearson metric because most studies reported results in this format.

Table 5
Reliability Meta-Analysis: Mixed-Effect Modeling of Differences in Effect Sizes, Dual-Occasion Reliabilities

						95% CI	
Effect	*lnL*	*p*	χ^2	*p*	β	Lower	Upper
Binomial model							
Intercept	−1,066.70	.000			.828	0.564	1.043
Type of construct	−979.60	.345	174.21	.313			
Externalizing					.000		
Internalizing					.057	−0.170	0.291
Pathological introversion					.080	−0.077	0.270
Thought disorder					−.174	−0.400	0.002
Other					−.292	−0.631	0.105
Type of sample	−1,066.65	.501	0.105	.807			
Nonclinical					.000		
Clinical					−.034	−0.243	0.208
Difference model							
Intercept	−199.105	.000			.070	0.005	0.123
Type of construct	−196.584	.366	5.04	.445			
Externalizing					.000		
Internalizing					.036	−0.080	0.130
Pathological introversion					.042	−0.022	0.112
Thought disorder					.010	−0.071	0.092
Other					−.080	−0.237	0.147
Type of sample	−198.966	.474	0.276	.620			
Nonclinical					.000		
Clinical					.035	−0.024	0.098

Note. Parameter estimates are fixed-effect terms from the mixed-effect models of differences within pairs of effect sizes, as described in the text. The binomial model modeled the probability of observing a continuous effect size greater than its discrete counterpart; the difference model modeled the magnitude of difference between continuous effect sizes and their discrete counterparts. *P* values for the log-likelihood and likelihood ratio χ^2 and confidence intervals for the effect parameters β are based on MC–permutation and MC–bootstrap procedures described in the text. Effects are coded relative to a baseline condition, represented by the intercept, of a measure of externalizing used in a nonclinical sample. MC = Monte Carlo.

(Krueger & Markon, 2006; Markon, 2010b). There was an additional category for studies in which the validity of multiple constructs was examined simultaneously.

Analyses. As in the reliability meta-analysis, two approaches were adopted in analysis of validities: analyses of the effect sizes per se and analyses of the differences between the paired effect sizes obtained using discrete and continuous measures. Mixed-effect modeling with MC–bootstrap and MC–permutation inference was again used in each approach, with effect sizes modeled as nested within study.

Analysis of individual effect sizes. Mixed-effect models were again estimated with the lme4 package (Bates & Sarkar, 2006) for R (R Development Core Team, 2010). Each effect size estimate was weighted by the inverse of its estimated variance, with variances for effect size estimates used as reported in Gurland (1968) and Gurland and Milton (1970). Confidence intervals were computed for coefficients of the fixed effect terms with the MC–bootstrap procedure utilized in the reliability meta-analysis. Due to heterogeneity in effect sizes reported across studies, analyses of individual effect sizes were restricted to those in a correlation metric (i.e., beta weights and odds ratios were excluded). The results of these analyses were used to meta-analytically estimate average validities encountered in the literature.

Analysis of paired effect sizes. To evaluate the effect of moderators, we conducted analyses on the within-study differences between continuous and discrete effect size estimates. Two mixed-effect models were examined: a model of the binomial probabilities of the continuous estimate being greater than the discrete estimate and a model of the magnitude of the differences between the two estimates. Study sample sizes were again used as weights. As in the analyses of individual effect sizes, models of the magnitudes of differences were restricted to those pairs of effect sizes in a correlation metric. However, models of the binomial probabilities included all effect sizes. As in the reliability analysis, the significance of moderator terms was evaluated with MC–bootstrap *p* values for the log-likelihoods and MC–permutation *p* values for the LRT statistics.

Results

Analyses of individual effect sizes. Parameter estimates from the mixed-effect model of individual effect sizes are presented in Table 6, and corresponding meta-analytic validity estimates are shown in Table 7. Values in Table 6 represent fixed effect parameters coded relative to a discrete measure of externalizing in a nonclinical sample. (As in the reliability meta-analysis, caution is warranted in interpreting the coefficients in Table 6, as they should not be used to test the significance of terms; the values in the table are presented for the sake of completeness.) Again, the validities are greater for the continuous measures than for discrete measures across all observed study characteristics. The estimated validities under the model suggest that compared to use of discrete measures, use of continuous measures increases validity by approximately 37%. An increase is apparent for every type of psychopathology and for both types of sample.

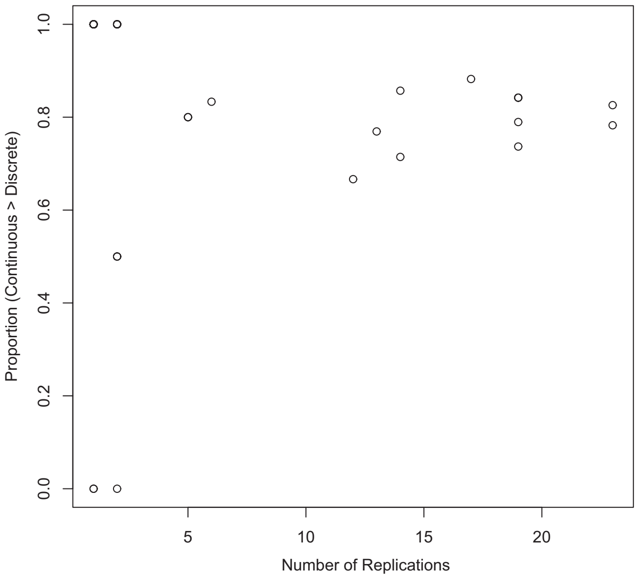

Figure 2. Plot of the proportion of reported reliabilities for which the continuous measure estimate is greater than the discrete measure estimate, as a function of the number of replications. Each point represents a different construct.

Analyses of paired effect sizes. The observed pairs of continuous and discrete validity estimates are plotted in Figure 4. In the figure, the size of the circles is proportional to the sample size used in the calculation of the estimates. The greater size of the continuous estimates is evident in the rightward shift of the points relative to the diagonal. The continuous and discrete estimates were also significantly correlated (Spearman's $\rho = .697$, $p < .0001$).

In general, results of the mixed-effect modeling of differences between paired effect sizes indicated that continuous measures had greater validities than discrete measures and that this phenomenon was not significantly moderated by type of phenotype or sample. (As in the individual effect size analyses, it is important to note that the confidence intervals for the individual coefficients are presented in Table 8 for the sake of completeness and should not be used to interpret significance.) Continuous measures had significantly greater validities than did their discrete counterparts, both in terms of the probability of a given continuous measure having greater validity than its discrete counterpart ($p = .000$) and in terms of magnitude of the difference ($p = .000$). The type of psychopathology construct did not significantly affect the probability of a continuous measure having greater validity than a discrete measure (MC–bootstrap $p = .475$; MC–permutation $p = .815$) or the magnitude of the difference (MC–bootstrap $p = .423$; MC–permutation $p = .297$). Similarly, type of sample did not significantly affect the probability of a continuous measure having greater validity than a discrete measure (MC–bootstrap $p = .482$; MC–permutation $p = .523$) or the magnitude of the difference (MC–bootstrap $p = .487$; MC–permutation $p = .687$).

Using the effect estimates in Table 8, the overall model-implied probability of observing a continuous measure with greater validity than a discrete measure was .699. Similarly, the model-implied overall difference in validities between continuous and discrete measures was .107, similar to the difference implied by the estimated validities in Table 7.

Trends by construct. To explore the possibility that specific constructs (e.g., alcohol use problems) might demonstrate evidence of equal validity for continuous and discrete measures even though superordinate groupings of constructs (e.g., externalizing) do not, we examined the probability of observing a continuous measure with a validity greater than that of a discrete measure for the constructs examined across studies. As in the analysis of reliabilities, the probability of observing a greater validity with a continuous measure was found to depend on the number of replications of findings involving the construct. This trend is illustrated in Figure 5. As can be seen, as the number of effect sizes reported on a construct increased, the probability of observing greater validities with continuous measures increased and, conversely, the probability of observing greater validities with discrete measures decreased.

This trend was not as pronounced as it was in the reliability meta-analysis, however. The point in the lowermost right portion of the plot, for example, represents obsessive-compulsive personality pathology. Continuous and discrete measures of obsessive-compulsive personality pathology were approximately equally likely to demonstrate larger validities, despite the fact that effect sizes were reported on the construct over twenty times. Whether this pattern would remain with further replications is unclear.

Discretization models. As in the reliability meta-analysis, the formulas of MacCallum et al. (2002) were applied to the discrete–continuous pairs, in order to determine how the observed validities for discrete variables would compare to what would be expected

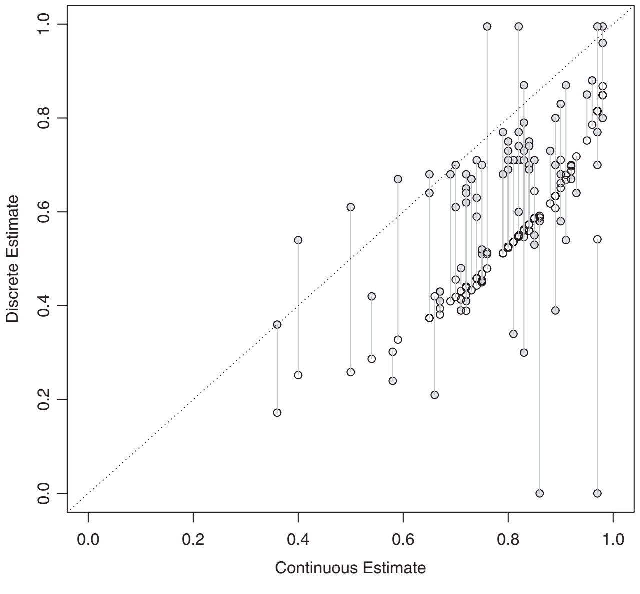

Figure 3. Plot of discrete reliabilities as a function of the continuous reliabilities, for observed values and predicted values based on simple dichotomization of a bivariate normal distribution (Kraemer, 1979). Only kappa–Pearson pairs are shown. Filled circles correspond to observed values; open circles are predicted values based on a simple dichotomization model. Light solid lines indicate observed–predicted pairs; the dark dotted line illustrates what would be expected if the discrete and continuous estimates were identical.

Table 6
Validity Meta-Analysis: Mixed-Effect Modeling of Individual Effect Sizes

Effect	β	95% CI	
		Lower	Upper
Intercept	.2002	.1316	.2500
Type of measure			
Discrete	.0000		
Continuous	.0883	.0249	.1933
Type of construct			
Externalizing	.0000		
Internalizing	.0250	−.0509	.0988
Pathological introversion	.0250	−.0477	.0911
Thought disorder	.0128	−.0608	.0765
Multiple	.0212	−.0453	.0801
Type of sample			
Nonclinical	.0000		
Clinical	.1040	.0396	.1729
Interaction: Measure × Construct			
Continuous externalizing	.0000		
Continuous internalizing	.0304	−.0849	.0948
Continuous pathological introversion	−.0242	−.1355	.0565
Continuous thought disorder	−.0776	−.1756	.0142
Continuous multiple	−.1586	−.2649	−.0564
Interaction: Measure × Sample			
Continuous nonclinical	.0000		
Continuous clinical	.1114	−.0115	.1599

Note. Parameter estimates are fixed-effect terms from the mixed-effect models described in the text. *P* values and confidence intervals are based on Monte Carlo–bootstrap procedures described in the text. Effects are generally coded relative to a baseline condition, represented by the intercept, of a discrete measure of externalizing used in a nonclinical sample.

under a simple discretization model. Sample-specific base rates, mean base rates for the constructs involved, or mean base rates observed in the meta-analysis were used as available; in cases where base rates were available separately for the two assessments, the first base rate was used. In order to make use of MacCallum et al.'s (2002) results, we restricted analyses to effect sizes involving continuous outcomes.

The results of these analyses are illustrated in Figure 6. As is evident in the figure, there was substantial variability around the values predicted by a simple discretization model, although less than was observed for the reliabilities (variance of the difference = .009). The mean observed–expected difference in discrete validities was .045; the median difference was .055; these differences were not significant according to a Wilcoxon rank sum test (W = 3699; p = .089). As was the case for the reliability meta-analysis, the observed discrete validities were typically somewhat larger than would be expected under discretization of an underlying normal distribution. In general, however, validities were more consistent with a discretization model than was the case for the reliabilities. In fact, in contrast to the reliabilities, they did not differ significantly from what would be expected under such a model.

Table 7
Validity Estimates Based on Mixed-Effect Modeling of Individual Effect Sizes

Parameter	Overall	Discrete	Continuous
Overall	.364	.305	.419
Type of construct			
Externalizing	.326	.254	.395
Internalizing	.420	.317	.514
Pathological introversion	.402	.321	.477
Thought disorder	.369	.311	.424
Multiple	.348	.324	.372
Type of sample			
Nonclinical	.247	.205	.288
Clinical	.381	.320	.438

Note. Validity estimates are predicted values of the model based on fixed-effect parameters from Table 6.

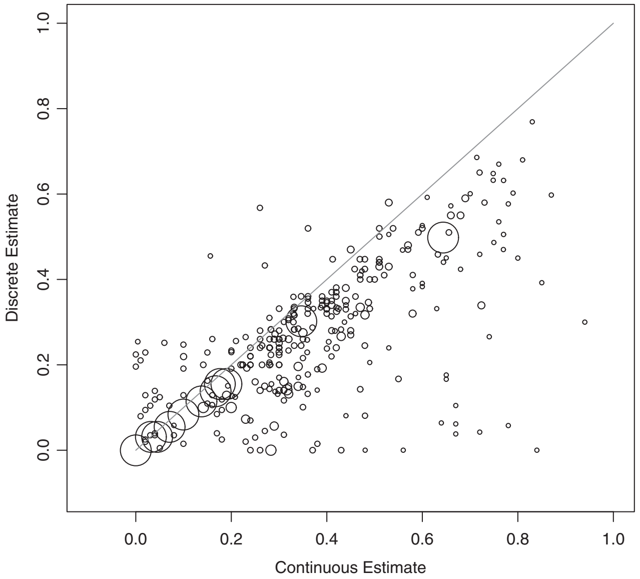

Figure 4. Scatter plot of validity estimates; each point represents a pair of estimates, one using a continuous measure and the other using a discrete measure. The size of circles is proportional to the sample size used to calculate the estimate. The light solid line is drawn for a reference point and represents what would be expected if the continuous and discrete estimates were equal. To the extent points are lower than the line, continuous estimates are greater; to the extent points are above the line, discrete estimates are greater.

Discussion

As with the reliability meta-analysis, this meta-analysis of validities supports previous findings that continuous measures are more valid than discrete measures across a wide range of settings. Our meta-analysis extends these findings by demonstrating that this difference in validity is largely invariant across different spectra of psychopathology and is robust to the clinical status of the sample. The pattern of results found here for validities parallels that of the reliabilities, suggesting that researchers have an approximately 70% chance of observing greater validities with continuous measures and on average could expect a 37% increase in validity through adoption of a continuous measure alone.

Viewed from the perspective of statistical power, the current results suggest that use of continuous measures reduces by half the

Table 8
Validity Meta-Analysis: Mixed-Effect Modeling of Differences in Effect Sizes

						95% CI	
Effect	*lnL*	*p*	χ^2	*p*	β	Lower	Upper
Binomial model							
Intercept	−20,166.19	.000			.911	0.808	1.008
Type of construct	−20,066.34	.475	199.69	.815			
Externalizing					.000		
Internalizing					.074	−0.046	0.225
Pathological introversion					−.030	−0.171	0.097
Thought disorder					−.065	−0.168	0.018
Multiple					.025	−0.110	0.126
Type of sample	−20,165.92	.482	0.530	.523			
Nonclinical					.000		
Clinical					−.080	−0.192	0.057
Difference model							
Intercept	−708.29	.000			.111	0.066	0.151
Type of construct	−704.28	.423	8.03	.297			
Externalizing					.000		
Internalizing					−.018	−0.075	0.035
Pathological introversion					−.043	−0.102	0.021
Thought disorder					−.056	−0.113	−0.004
Multiple					−.121	−0.166	−0.074
Type of sample	−708.19	.487	0.202	.687			
Nonclinical					.000		
Clinical					.064	0.014	0.119

Note. Parameter estimates are fixed-effect terms from the mixed-effect models of differences within pairs of effect sizes, as described in the text. The binomial model modeled the probability of observing a continuous effect size greater than its discrete counterpart; the difference model modeled the magnitude of difference between continuous effect sizes and their discrete counterparts. *P* values for the log-likelihood and likelihood ratio χ^2 and confidence intervals for the effect parameters β are based on MC–permutation and MC–bootstrap procedures described in the text. Effects are coded relative to a baseline condition, represented by the intercept, of a measure of externalizing used in a nonclinical sample. MC = Monte Carlo.

sample sizes necessary to achieve traditional significance levels for the typical effect sizes observed (e.g., achieving 80% power to detect the validities in Table 5 at a significance level of .05 would require approximately 43 individuals for a continuous measure vs. 82 individuals for a discrete measure). Although the studies examined here typically had more than enough power to detect these effect sizes, there are many situations in which issues of power and sample size are critical. The current results suggest that adoption of continuous measures could greatly improve power in such situations.

General Discussion

Overall, the results of these two meta-analyses—involving a total of 58 studies and 59,575 participants—suggest that continuous measures of psychopathology are more reliable and valid across a wide range of settings, with little evidence of any significant exceptions. Individuals have roughly a 70% chance of observing greater reliabilities and validities with continuous measures and can expect an increase of 15% in reliability and 37% increase in validity simply by adopting a continuous over a discrete measure of psychopathology.

These results support a large body of empirical and theoretical literature suggesting that continuous measures generally outperform discrete measures, and echo the many calls for more widespread adoption of continuous measures in clinical and research settings (e.g., Clark, 1999; Krueger et al., 2005; Trull & Durrett, 2005; Widiger & Samuel, 2005). These results extend prior research by demonstrating the empirical effect of using continuous measures across different settings, precisely quantifying the magnitude of this effect, and demonstrating that it is robust to differences in construct and sample.

The weight of evidence summarized here suggests that, in the absence of a specific rationale for the contrary, continuous measures of psychopathology should be preferred over discrete measures a priori, insofar as they increase reliability and validity. Our results are broadly consistent with the recommendations of DeCoster et al. (2009), in this regard, in that in the absence of any evidence to the contrary, use of continuous measures is likely to maximize or nearly maximize the strength of observed relationships with other measures.

Explaining the Relative Performance of Discrete and Continuous Measures

Given that continuous measures of psychopathology are generally more reliable and valid, it is reasonable to ask why this is the case. A related and perhaps more important question is why the magnitude of the difference in performance is as large—or small—as it is and what causes variation in this difference. For example, if the analyses suggest that there is roughly a 70% chance of observing a greater reliability or validity when using continuous measures, this implies that there is a 30% chance of observing a greater value using discrete measures—which is somewhat larger than might be expected given some simulations reported in the methodological literature (e.g., MacCallum et al., 2002). Under-

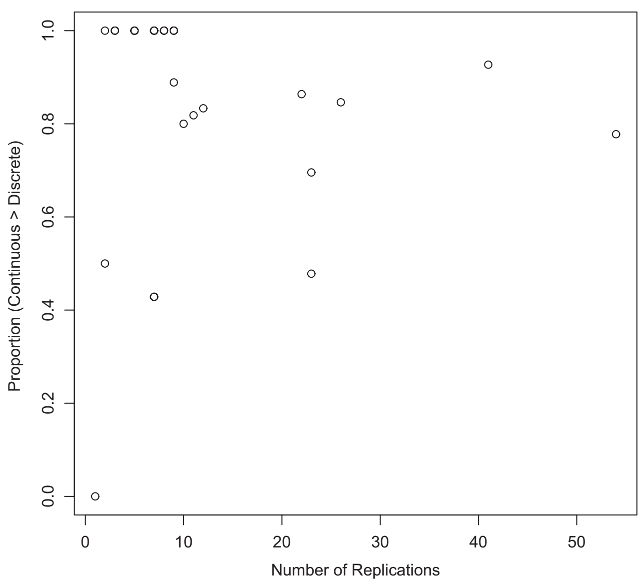

Figure 5. Plot of the proportion of reported validities for which the continuous measure estimate is greater than the discrete measure estimate, as a function of the number of replications. Each point represents a different construct.

standing why this is the case might help inform psychopathology theory and research.

Loss of information. At some level, the most likely explanation for the widespread psychometric superiority of continuous measures of psychopathology is the same that has been proposed for other measures: loss of information. When a construct lies on a continuum and is artificially discretized, valid variation between individuals is often eliminated. Individuals who are relatively different might be reclassified as having the same diagnostic status, and individuals who are relatively similar might be reclassified as having different diagnostic status. This discretization process obscures individuals' standing on the latent variable, lowering reliabilities and validities (excellent discussions of this are provided by DeCoster et al., 2009, and MacCallum et al., 2002).

Although discussions in the literature have often suggested that psychopathology might represent an important exception to this phenomenon (e.g., DeCoster et al., 2009; MacCallum et al., 2002; Ruscio & Ruscio, 2002), the current meta-analyses suggest this is not the case. Across a wide variety of measures and constructs, continuous measures are generally more reliable and valid, suggesting that relative to use of continuous measures, use of discrete measures incur a loss of information. The extra precision in the scale of measurement of continuous measures provides additional, valuable information that improves statistical power.

At the same time, it is important to note that the magnitude of the loss of information with discretization of psychopathology measures might be less than expected under simple discretization models (i.e., those in which the discrete measures represent dichotomized continuous normal distributions; e.g., Kraemer, 1979; MacCallum et al., 2002). In both the reliability and the validity analyses, observed values on discrete measures were somewhat larger than expected under a simple discretization process. These results should be interpreted with a great deal of caution, as this difference was not significant in the validity analysis, and it is unclear how well the assumptions of the models (e.g., normality of the continuous variables) are met. This difference does however raise the possibility that discrete measures of psychopathology lose somewhat less information than might be expected under a simple bivariate normal threshold model.

One interpretation of this, discussed in the following section, is that psychopathology constructs sometimes represent mixtures of continuous subpopulations, comprising valid continuous variance as well as discrete features. Under this scenario, continuous measures provide valid information that is not entirely lost with discretization. Another possibility is a psychometric variant of the file drawer problem: that reports of statistics are selective when discrete measures are being used (e.g., not reporting reliabilities for measures that fail to attain conventional standards). A related possibility is that authors might sometimes choose thresholds of discrete measures in such a way as to maximize validities in any particular sample, which would lead to upwardly biased estimates of effects (cf. Marshall et al., 2000). This might be especially relevant to psychopathology measures, as a single instrument might be scored in multiple ways to provide discrete as well continuous measurements and thereby be subject to multiple standards or afford flexibilities in scoring.

The importance of appropriate thresholds. Thresholds (e.g., diagnostic thresholds, cutpoints) are critical to understanding how well a discrete measure will perform relative to a continuous measure. The loss of information associated with discrete measures when a continuous variable is discretized is strongly related to how the thresholds are established (e.g., Felsenstein & Pötzel-

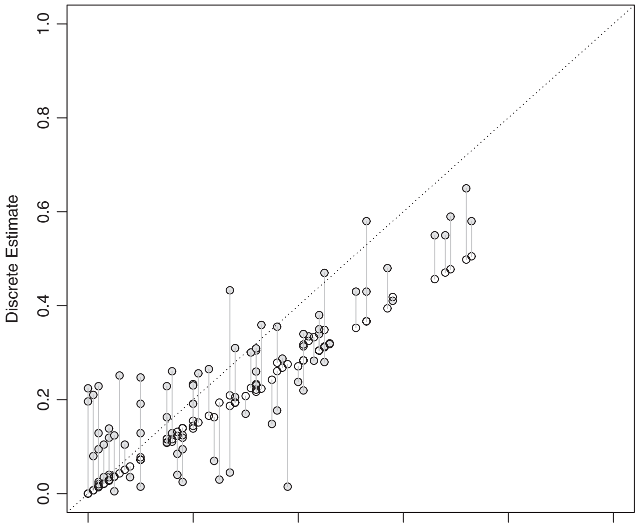

Figure 6. Plot of discrete validities as a function of the continuous validities, for observed values and predicted values based on simple dichotomization of a bivariate normal distribution (MacCallum, Zhang, Preacher, & Rucker, 2002). Filled circles correspond to observed values; open circles are predicted values based on a simple dichotomization model. Light solid lines indicate observed–predicted pairs; the dark dotted line illustrates what would be expected if the discrete and continuous estimates were identical.

berger, 1998; MacCallum et al., 2002). More generally, the appropriate use of thresholds is critical to maximizing the reliability and validity obtained with discrete measures (DeCoster et al., 2009). The studies examined here, likely to be representative of actual practice, utilized diverse approaches to establishing thresholds: Many targeted conventional criteria established by official nosologies such as the *DSM* or *ICD* (e.g., Chanen et al., 2004; Zimmerman & Coryell, 1989), some used signal detection theory methods (e.g., Ross, Gavin, & Skinner, 2003), and some used thresholds established by other statistical methodologies (e.g., Goedeker & Tiffany, 2008; Watson, 2003). As not all of the studies included in the meta-analyses used optimal thresholds, the current results might underestimate what might be achievable by discrete measures relative to continuous measures in terms of their psychometric characteristics.

In this regard, our results highlight the need for more rigorous consideration of how to best make discrete assessment decisions when they are necessary or desired for whatever reason. Given that relative to continuous measures, discrete indicators are likely to incur some loss in statistical power, greater attention should be placed on how to optimally develop them to minimize this loss. Signal detection theory methods (McFall & Treat, 1999; Swets, 1988), which have been used to establish cutoffs in psychopathology research for some time (e.g., Murphy et al., 1987), represent one underutilized approach to establishing such thresholds. However, signal detection theory approaches to thresholds rely on comparing a continuous measure to a diagnostic standard. This creates a certain circularity—how does one establish the threshold for the standard?—suggesting other methods are also needed (Faraone & Tsuang, 1994). Methods based on latent variables, such as taxometric approaches to establishing thresholds (e.g., Ruscio & Ruscio, 2002; Waller & Meehl, 1998) and mixture models (McLachlan & Peel, 2000), represent promising possibilities in this regard. In general, more theoretical and empirical attention should be focused on how to best make discrete assessments of observed psychopathology and, more broadly, how this should be instantiated in nosologies that can be used in clinical and research settings (for a detailed discussion of this issue, see Kamphuis & Noordhof, 2009).

The role of sample characteristics. Certain features of the current results help clarify why continuous indicators of psychopathology might be relatively more reliable and valid in some settings than others. For instance, these results suggest that the reliabilities of discrete and continuous indicators differ more in nonclinical than in clinical settings, at least when behaviors are held constant across assessment occasions. One interpretation of such results is that relatively less severe, subclinical behaviors are evidenced in nonclinical settings in a way that can be captured by continuous but not discrete measures of psychopathology. Consistent with this interpretation, in single-occasion assessment scenarios, discrete reliabilities decreased much more in nonclinical settings than did continuous reliabilities, the latter changing relatively little across sample type. Continuous measures of psychopathology generally appear more reliable—there were no conditions when the opposite appeared true—but this may be especially true in settings where psychopathology is less severe or prevalent or where continuous measures can otherwise better accommodate informative behaviors that are evidenced, across a greater range.

Differences in reliability versus validity. One interesting set of questions relates to whether and why the advantage of continuous measures might be different for reliability versus validity statistics. It is important to note in this regard that even though continuous measures produced greater reliabilities and validities in terms of relative magnitude (15% and 37% increases, respectively), the absolute magnitude of the increases was extremely similar for reliabilities and validities (.109 and .114, respectively). In some sense, then, the apparently greater advantage for validities is due to the fact that the typical validities (.364) were smaller than the reliabilities (.766), so the same absolute increase had a larger relative effect.

Nevertheless, various factors might be differentially relevant to the reliability versus validity advantages of continuous versus discrete measures of psychopathology. Similar to what was already noted, for example, reports of reliabilities often serve a different function than those of validities and are subject to different standards in that regard. Reporting on and using poor measures usually has little utility, whereas reporting on validities has often has scientific utility regardless of their magnitude. Truncation in what is reported about reliabilities might affect what is concluded about discrete versus continuous measures in this regard, at least relative to what is concluded about validities. Another possibility is alluded to in the observation that the advantage of continuous measures is attenuated in clinical samples only for single-occasion reliability designs: that many reliability designs decrease the number of sources of variance affecting a measure, which reduces the advantage of continuous measures relative to what might be observed for validities.

It is likely that the decreased reliability of discrete measures is contributing in part to their decreased validity. However, quantifying the extent to which this is the case and the extent to which other factors are contributing to the decreased validity of discrete measures is challenging. As Carey and Gottesman (1978) noted, the relationship between reliability and validity is complex for dichotomous measures and depends on specific parameters of the situations encountered. In fact, they showed that increases in reliability can sometimes be associated with decreases in validity under certain conditions, an observation they cited as reason for caution in overemphasizing reliability at the expense of validity (cf. Hyman, 2010). Determining how and why relative decrements in the reliability of discrete measures relate to decrements in validity will require additional research, but such studies will probably reveal a great deal about how to optimize assessments of psychopathology and the nature of the underlying constructs.

Random sampling variation. Finally, it is important to draw attention to the role of random variation (e.g., sampling error) in observing larger or smaller reliabilities or validities in discrete versus continuous measures. MacCallum et al. (2002) illustrated this phenomenon extremely well: Even when dichotomization decreases associations in discrete measures at the population level, it is possible to observe larger values with discrete versus continuous measures simply by chance. The trends illustrated in Figures 3 and 5 reinforce this observation by showing that, as the number of reports on a construct increase, reports of greater reliabilities and validities using continuous measures tend to outnumber the reverse at greater rates. Psychiatric measurement may be similar to other areas of science in this regard, in that many initial reports of a phenomenon can be attributable to chance, and subsequent replications are critically important (Ioannidis, 2005).

Varieties of Discreteness and Continuousness in Psychopathology Theory

The current results indicate that it is generally reasonable to assume a priori that continuous measures of psychopathology are more reliable and valid. It is important to emphasize, however, that characteristics of observed measures are only one perspective on how psychopathology can be understood to be discrete or continuous. For instance, finding that continuous observed measures are more reliable or valid does not preclude discreteness in the distribution of the latent construct itself. Also, it is important to emphasize that even if a latent construct or its manifestation in a set of observed measures is continuous, it may nevertheless demonstrate discreteness in the form of the construct's relationships with other variables.

Latent versus manifest discreteness and continuousness. It is important to distinguish between continuousness and discreteness in the latent constructs themselves versus in the observed, manifest indicators. One can have continuous indicators of discrete constructs or vice versa. It is important to note that even though our results suggest that continuous indicators are generally more reliable and valid, this does not preclude discreteness of some sort in the latent variables themselves. As noted in the introduction, it is often better, seemingly paradoxically, from a psychometric perspective to assess discrete latent variables with continuous indicators (Lubke & Neale, 2008; Markon & Krueger, 2006).

The widespread superiority of continuous measures in our analyses in terms of reliability and validity does suggest that continuity in the observed measures is capturing some valid continuity at the latent level. However, it is possible to explain this phenomenon in terms of a mixture of latent continuous subpopulations. Under this scenario, discrete groups do exist, but there is valid continuous variation within each group. In the taxometrics literature, this within-group variance has sometimes been referred to as "nuisance covariance" (Waller & Meehl, 1998). Considered in the context of the current results, this label is somewhat misleading in that continuous measures are likely able to capitalize on this latent within-group variance to increase statistical power at the observed level. Although the within-group variance does create challenges in distinguishing between the latent groups (relative to the situation where there is no within-group variance) and is nuisance variance in this sense, it nevertheless provides important, valid information from an assessment perspective. Continuing with the example of general cognitive ability, imagine an individual carrying a catastrophic mutation (e.g., a repeat expansion mutation; La Spada & Taylor, 2010): although the individual is a member of a discrete etiological class, his or her actual measured general cognitive ability, which is measured continuously, has important predictive status.

These mixture models hold great promise as a way to explain latent discreteness in psychopathology and are generally consistent with our results. The fact that reliabilities and validities were actually slightly larger in our analyses than would be predicted on the basis of dichotomization alone, for example, is consistent with such a scenario. Also consistent with this scenario was the tendency for discrete and continuous reliabilities to be more similar in clinical samples, under certain circumstances—the attenuated difference might reflect discrete subpopulations observable in clinical samples but only rarely if at all in nonclinical samples. Mixture models—which arguably form the basis for taxometric approaches but are not limited to those methods—have been explicitly proposed for various forms of psychopathology, such as psychosis proneness (Lenzenweger, McLachlan, & Rubin, 2007). Mixture models would help explain observations that the available evidence supports discrete as well as continuous features of psychosis (Linscott & van Os, 2010).

It is possible in this sense for a construct to possess both discrete and continuous features and for constructs to occupy a sort of middle position in a continuum from perfectly discrete to perfectly continuous. De Boeck, Wilson, and Acton (2005), for example, have proposed that the degree of discreteness or continuousness of a variable can be formulated in terms of the joint degree of within- and between-group variation of a construct. Relatively discrete constructs, in this framework, have large between-group variation and relatively little within-group variation; relatively continuous constructs, in contrast, have small between-group variation and relatively large within-group variation. It is conceivable for a psychopathology construct to occupy some middle position in this framework, possessing valid and reliable continuous features within groups that nevertheless differ discretely.

Psychometric versus phenomenological discreteness and continuousness. Another useful distinction can be made between what Flett et al. (1997) referred to as psychometric and phenomenological continuity. The former refers to the continuity of the construct itself, in terms of its distribution and the manifestation of that distribution in observed variables; the latter refers to the continuity of the construct in terms of its relationships with other variables. Even if a construct is most usefully conceived of as continuous in a psychometric sense, it may exhibit nonlinearities in its relationships with other variables that make it discontinuous in a phenomenological sense. A form of psychopathology may be psychometrically continuous, for example, but nevertheless accelerate impairment at some level of severity, or it may itself have resulted from some etiology that accelerates in effects beyond some threshold. These nonlinearities in either effects or causes of the constructs could define a discontinuity in the construct, not in terms of the distribution of the construct itself but in how it relates to other variables within its nomological network.

One of the current authors, for example, recently explored phenomenological continuity and discontinuity in the relationship between internalizing psychopathology and impairment (Markon, 2010a). Under one of the evaluated hypotheses, as internalizing psychopathology increases, impairment increases at some constant rate until "things fall apart" and difficulty in adaptive functioning accelerates. Under the other hypothesis, the relationship between internalizing and impairment increases uniformly and linearly, such that there is no point at which impairment discernibly accelerates. A comparison of the two models suggested this later model—reflecting phenomenological continuity in the relationship between internalizing and impairment—fit better according to a variety of criteria and across gender. Whether other forms of phenomenological discontinuities can be identified for internalizing psychopathology, either in etiologies or outcome, is an important area for future research.

Delineating nonlinearities in relationships with antecedents and consequences of constructs rather than identifying discontinuities in the constructs themselves is an important direction for future inquiry. The discreteness or continuousness of any given construct reflects an element of the theory of that construct and can be explicated in terms of its causes and effects in a nomological network. Questions regarding psychometric continuity are equally important but represent only one form of continuity that can be examined.

Implications for Classification and Nosology

These results have important implications for classification of psychopathology, whether with regard to official nosologies, such as the *DSM* or *ICD,* or with regard to classification more broadly. Current official nosologies adopt a discrete approach to conceptualizing and assessing psychopathology, both reflecting and influencing the development of theory and research on mental illness. The results of these two meta-analyses suggest that, to the extent that the studies surveyed are representative of existing research, adoption of discrete measurements has probably resulted in a decrement in explanatory power relative to what would have been achieved with use of continuous measures. Adoption of a continuous approach to psychopathology assessment is likely to increase observed reliabilities and validities, enhancing the development of theory and research in these areas.

It is worthwhile in this regard to consider the role that similar considerations of reliability and validity played in the development of current nosology. By increasing focus on explicit, quantitative criteria for assessment, *DSM–III* instituted a major change in approach to psychiatric nosology, forming the foundation for current practice (Blashfield, 1982; Compton & Guze, 1995). Although many factors contributed to the development of *DSM–III* (American Psychiatric Association, 1980), one major consideration was a desire to increase the reliability and validity of psychiatric diagnosis. The results of these meta-analyses suggest that reliability and validity of psychiatric assessments might similarly be improved in future nosologies by adopting a continuous assessment paradigm.

Our results are particularly relevant to what Kendell and Jablensky (2003) referred to as the utility of classification systems. That is, independent of whether the indicators provided by a psychiatric system reflect some underlying true state of nature, they may nevertheless provide important, clinically useful predictive power in a purely statistical sense. To the extent that increasing utility is a goal of nosologies such as the *DSM* or *ICD,* the current results suggest that adopting continuous measures will generally do so more efficiently. Writing about the importance of utility in the *DSM–5,* Mullins-Sweatt and Widiger (2009) emphasized a similar point: that nosological features increasing validity from a theory development standpoint are likely to increase utility as well. As Kendell and Jablensky suggested, utility is as important in clinical settings as research settings—if not more important, given its practical salience. For example, in some sense, in a clinical setting, understanding the true etiologies underlying an individual's problems might matter less than predicting and improving outcome in that individual per se. Such considerations highlight the utility of continuous measures for such purposes.

Many researchers have written about the costs associated with psychopathology not well represented by traditional discrete, binary indicators. For example, studies of depression—one of the most common forms of psychopathology—have shown that depression below diagnostic thresholds is associated with significant impairment at levels comparable to those for depression above diagnostic thresholds (e.g., Broadhead, Blazer, George, & Tse, 1990; Gotlib, Lewinsohn, & Seeley, 1995; Johnson, Weissman, & Klerman, 1992; Judd, Paulus, Wells, & Rapaport, 1996; Sherbourne et al., 1994). Moreover, this depression–impairment relationship is gradual, with incremental increases in depression being associated with incremental increases in impairment. As a consequence, some researchers (e.g., Fergusson, Horwood, Ridder, & Beautrais, 2005; Pickles et al. 2001) have argued that depressive severity, rather than depressive diagnosis per se, is critical to predicting levels of impairment and outcome. Our results reinforce these observations by showing that this graduated, incremental approach to assessment of psychopathology increases statistical power for a wide variety of psychopathologies and correlates. Adopting a more continuous measurement paradigm is arguably the simplest way to successfully represent within classification systems the costs associated with these intermediate states of psychopathology.

Limitations and Caveats

Power. Although these meta-analyses help quantify the relative reliability and validity of continuous and discrete measures of psychopathology, a number of limitations of the current research must be kept in mind. One important consideration is whether the current analyses were adequately powered to identify moderation of the difference between continuous and discrete measures by the variables examined (i.e., type of psychopathology and sample). Although this might seem counterintuitive in the context of a meta-analysis, it is an important issue to consider.

In order to determine whether the meta-analyses were sufficiently powered to detect moderating effects, we estimated power for the paired effect size analyses, assuming population effect sizes that were equal in magnitude but opposite in direction to the main effects. Using these assumed values allowed us to estimate the power to detect a moderating condition that eliminates the reliability and validity advantage for continuous measures of psychopathology. We ran Monte Carlo simulations in which observed conditions (e.g., study design characteristics) were exactly the same as in the actual meta-analyses but the population effect sizes were set equal to values that would nullify the main effect of continuousness, as just described.

These power analyses (summarized in Table 9) suggested that our moderation analyses were more than adequately powered to detect a moderating effect, for both the reliability and validity analyses, with one important exception: moderating effects of miscellaneous constructs not falling uniquely within the four a priori higher order forms of psychopathology (i.e., internalizing, externalizing, pathological introversion, or thought disorder). For all other moderators, including sample type and the four primary higher order forms of psychopathology, power was more than adequate for at least one of the tests, even though any given test might have been underpowered under any given condition. The lack of power to detect a

Table 9
Power Analyses

Effect	Binomial χ^2	Binomial *lnL*	Difference χ^2	Difference *lnL*
Reliability				
Type of construct				
Internalizing	0.9982	0.9272	0.9773	1.0000
Pathological introversion	0.9983	0.9612	0.9631	1.0000
Thought disorder	1.0000	0.9928	0.9947	1.0000
Other	0.6559	0.4013	0.4035	0.3363
Type of sample				
Clinical	0.4167	0.0769	0.9971	0.4912
Validity				
Type of construct				
Internalizing	0.9697	0.2877	0.8496	1.0000
Pathological introversion	0.9813	0.3574	0.7031	0.9216
Thought disorder	0.8907	0.4513	0.9983	1.0000
Multiple	0.4829	0.0615	0.5366	0.9455
Type of sample				
Clinical	0.9591	0.0505	0.9998	0.8589

Note. Values are the proportion of 10,000 simulation replications in which a moderator was correctly identified as significant at $\alpha = .05$, when the population moderator effect countervails the main effect of continuousness. Simulations are described in the text; conditions were designed to emulate observed conditions in every way, but with the moderator effect set to be exactly equal in magnitude but opposite in direction to the main effect of continuousness.

moderating effect for various miscellaneous constructs probably reflects the small number of studies examining forms of psychopathology that cannot be classified into one of the higher order forms of psychopathology. The results of power analyses suggest that some caution is warranted in interpreting results of the significance tests involving specific constructs not easily classified into one of the four higher order categories.

Other forms of psychopathology. With this consideration in mind, it is also important to note that the current analyses were not exhaustive in examining all forms of psychopathology—if such a thing is even possible. Although we believe that the current results are broadly representative of most general forms of psychopathology, particular forms of psychopathology might prove to be important exceptions. Many of the studies included in the meta-analysis, for example, were of personality disorder, reflecting an interest in that literature on comparing discrete and continuous measures of psychopathology (Trull & Durrett, 2005). Conversely, particular forms of psychopathology, such as eating disorder, were represented very little, and others, such as sexual disorders, were not represented at all. Although a diversity of forms of psychopathology was represented and the empirical results we obtained here were strongly consistent with results of methodological research, it is possible that the reliability and validity of continuous and discrete measures might not differ for certain specific forms of psychopathology. As our analyses suggested that greater reliability and validity tend to be observed in continuous measures as constructs are studied more, it is important that constructs not well represented be examined more in future studies.

Other considerations in evaluating the scale of observed measures. Finally, it is important to note that reliability and validity, although important, are not the only criteria for the utility of a scale of measurement. Discrete measurement, for example, is arguably critical in certain clinical decision-making settings, such as in settings where discrete outcomes are of primary importance (e.g., recidivism). Certain applications of discrete classifications (e.g., extreme-groups designs), moreover, are also extremely useful. The most appropriate interpretation of our results is that, in the absence of any other specific reason to assume otherwise, the weight of evidence suggests that psychopathology constructs are likely to be more reliably and validly assessed with a continuous scale of measurement. In particular settings, however, this generalization might not be appropriate.

It is important to acknowledge that in clinical settings, the primary purpose of an assessment is not necessarily to maximize reliability—which is a property of sample, not an individual—but to maximize the amount of information obtained or conveyed about an individual. Many have noted that the informativeness of an assessment depends on considerations such as base rates (e.g., Dawes, 1962; Meehl & Rosen, 1955). It is conceivable that by virtue of base rates, the particular measure, and other considerations, different assessment approaches might be differentially useful in different settings (cf. Kamphuis & Noordhof, 2009). In particular, if a latent construct is distributed as a mixture, this raises the interesting possibility that a single construct might possess discrete and continuous aspects simultaneously, with assessment of each aspect optimally informative in different settings, depending on characteristics of the scenario. Even more broadly, measurement goals might differ between clinical and research applications, given the generally idiographic versus nomothetic goals of the two endeavors, respectively (for a relatively recent review of clinical assessment from an idiographic perspective, see Haynes, Mumma, & Pinson, 2009).

Conclusions

Despite these caveats, the current results indicate that continuous measures of psychopathology generally produce greater reliabilities and validities than do their discrete counterparts. Researchers and clinicians switching from a discrete measure to a continuous measure will typically see increases in reliability and validity on the order of 15–37% in a correlation metric. This increase, consistent with methodological literature, is observed in all domains of psychopathology and in different sample types. Future research and theory are likely to benefit from increased use of continuous measures of psychopathology and from an increased focus on better explicating how to represent discreteness and continuousness within psychopathology measures.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (2006). *Practice guideline for the psychiatric evaluation of adults* (2nd ed.). Washington, DC: Author.

Baca-Garcia, E., Perez-Rodriguez, M. M., Basurte-Villamor, I., Fernandez Del Moral, A. L., Jimenez-Arriero, M. A., Gonzalez De Rivera, J. L., . . . Oquendo, M. A. (2007). Diagnostic stability of psychiatric disorders in clinical practice. *British Journal of Psychiatry, 190,* 210–216. doi:10.1192/bjp.bp.106.024026

Bates, D., & Sarkar, D. (2006). The lme4 package [Computer software]. Retrieved from http://lme4.r-forge.r-project.org/

Beauchaine, T. P. (2007). A brief taxometrics primer. *Journal of Clinical Child and Adolescent Psychology, 36,* 654–676. doi:10.1080/15374410701662840

Blanchard, J. J., Horan, W. P., & Collins, L. M. (2005). Examining the latent structure of negative symptoms: Is there a distinct subtype of negative symptom schizophrenia? *Schizophrenia Research, 77,* 151–165. doi:10.1016/j.schres.2005.03.022

Blashfield, R. K. (1982). Feighner et al., invisible colleges, and the Matthew effect. *Schizophrenia Bulletin, 8,* 1–6. doi:10.1093/schbul/8.1.1

Bloch, D. A., & Kraemer, H. C. (1989). 2 × 2 kappa coefficients: Measures of agreement or association. *Biometrics, 45,* 269–287. doi:10.2307/2532052

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine, 21,* 1331–1335. doi:10.1002/sim.1108

Broadhead, W. E., Blazer, D. G., George, L. K., & Tse, C. K. (1990). Depression, disability days, and days lost from work in a prospective epidemiologic survey. *JAMA, 264,* 2524–2528. doi:10.1001/jama.1990.03450190056028

Carey, G., & Gottesman, I. I. (1978). Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry, 35,* 1454–1459. doi:10.1001/archpsyc.1978.01770360058007

Chanen, A. M., Jackson, H. J., McGorry, P. D., Allot, K. A., Clarkson, V., & Yuen, H. P. (2004). Two-year stability of personality disorder in older adolescent outpatients. *Journal of Personality Disorders, 18,* 526–541. doi:10.1521/pedi.18.6.526.54798

Clark, L. A. (1999). Dimensional approaches to personality disorder assessment and diagnosis. In C. R. Cloninger (Ed.), *Personality and psychopathology* (pp. 219–244). Arlington, VA: American Psychiatric Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. doi:10.1177/001316446002000104

Compton, W. M., & Guze, S. B. (1995). The neo-Kraepelinian revolution in psychiatric diagnosis. *European Archives of Psychiatry and Clinical Neuroscience, 245,* 196–201. doi:10.1007/BF02191797

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. doi:10.1037/h0040957

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology, 26,* 422–424. doi:10.1037/h0044612

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112,* 129–158. doi:10.1037/0033-295X.112.1.129

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14,* 349–366. doi:10.1037/a0016956

DSM–5 Neurocognitive Disorders Workgroup. (2010). *Neurocognitive disorders: A proposal from the* DSM–5 *Neurocognitive Disorders Work Group.* Retrieved from http://www.dsm5.org/ProposedRevisionAttachments/APANeurocognitiveDisordersProposalforDSM-5.pdf

Eysenck, H. J. (1970). The classification of depressive illnesses. *British Journal of Psychiatry, 117,* 241–250. doi:10.1192/bjp.117.538.241

Faraone, S. V., & Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry, 151,* 650–657.

Felsenstein, K., & Pötzelberger, K. (1998). The asymptotic loss of information for grouped data. *Journal of Multivariate Analysis, 67,* 99–127. doi:10.1006/jmva.1998.1759

Fergusson, D. M., Horwood, J., Ridder, E. M., & Beautrais, A. L. (2005). Subthreshold depression in adolescence and mental health outcomes in adulthood. *Archives of General Psychiatry, 62,* 66–72. doi:10.1001/archpsyc.62.1.66

Ferro, T., Klein, D., Schwartz, J. E., Kasch, K. L., & Leader, J. B. (1998). Thirty-month stability of personality disorder diagnoses in depressed outpatients. *American Journal of Psychiatry, 155,* 653–659.

Fleiss, J. L., & Cohen, J. (1973). Equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619. doi:10.1177/001316447303300309

Flett, G. L., Vredenburg, K., & Krames, L. (1997). The continuity of depression in clinical and nonclinical samples. *Psychological Bulletin, 121,* 395–416. doi:10.1037/0033-2909.121.3.395

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349. doi:10.1037/0033-295X.92.3.317

Goedeker, K. C., & Tiffany, S. T. (2008). On the nature of nicotine addiction: A taxometric analysis. *Journal of Abnormal Psychology, 117,* 896–909. doi:10.1037/a0013296

Gotlib, I. H., Lewinsohn, P. M., & Seeley, J. R. (1995). Symptoms versus a diagnosis of depression: Differences in psychosocial functioning. *Journal of Consulting and Clinical Psychology, 63,* 90–100. doi:10.1037/0022-006X.63.1.90

Grant, B. F., Dawson, D. A., Stinson, F. S., Chou, S. P., Kay, W., & Pickering, R. P. (2003). The Alcohol Use Disorder and Associated Disabilities Interview Schedule–IV (AUDADIS-IV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug and Alcohol Dependence, 71,* 7–16. doi:10.1016/S0376-8716(03)00070-X

Gurland, J. (1968). A relatively simple form of the distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 30,* 276–283.

Gurland, J., & Milton, R. (1970). Further consideration of the distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 32,* 381–394.

Haslam, N. (2003a). Categorical versus dimensional models of mental disorder: The taxometric evidence. *Australian and New Zealand Journal of Psychiatry, 37,* 696–704. doi:10.1111/j.1440-1614.2003.01258.x

Haslam, N. (2003b). The dimensional view of personality disorders: A review of the taxometric evidence. *Clinical Psychology Review, 23,* 75–93. doi:10.1016/S0272-7358(02)00208-8

Haslam, N. (2007). The latent structure of mental disorders: A taxometric update on the categorical versus dimensional debate. *Current Psychiatry Reviews, 3,* 172–177.

Haynes, S. N., Mumma, G. H., & Pinson, C. (2009). Idiographic assessment: Conceptual and psychometric foundations of individualized behavioral assessment. *Clinical Psychology Review, 29,* 179–191. doi:10.1016/j.cpr.2008.12.003

Helzer, J., Kraemer, H., Krueger, R., Wittchen, H., Sirovatka, P., & Regier, D. (Eds.). (2008). *Dimensional approaches in diagnostic classification: Refining the research agenda for* DSM–5.Washington, DC: American Psychiatric Association.

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23,* 1663–1682. doi:10.1002/sim.1752

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology, 6,* 155–179. doi:10.1146/annurev.clinpsy.3.022806.091532

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Johnson, J., Weissman, M. M., & Klerman, G. L. (1992). Service utilization and social morbidity associated with depressive symptoms in the community. *JAMA, 267,* 1478–1483. doi:10.1001/jama.1992.03480110054033

Judd, L. L., Paulus, M. P., Wells, K. B., & Rapaport, M. H. (1996). Socioeconomic burden of subsyndromal depressive symptoms and major depression in a sample of the general population. *American Journal of Psychiatry, 153,* 1411–1417.

Kamphuis, J. H., & Noordhof, A. (2009). On categorical diagnoses in *DSM–V*: Cutting dimensions at useful points? *Psychological Assessment, 21,* 294–301. doi:10.1037/a0016697

Kendell, R., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry, 160,* 4–12. doi:10.1176/appi.ajp.160.1.4

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44,* 461–472. doi:10.1007/BF02296208

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2,* 111–133. doi:10.1146/annurev.clinpsy.2.022305.095213

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for *DSM–V. Journal of Abnormal Psychology, 114,* 537–550. doi:10.1037/0021-843X.114.4.537

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41,* 1183–1192. doi:10.1037/0003-066X.41.11.1183

La Spada, A. R., & Taylor, J. P. (2010). Repeat expansion disease: Progress and puzzles in disease pathogenesis. *Nature Reviews Genetics, 11,* 247–258. doi:10.1038/nrg2748

Lenzenweger, M. F., McLachlan, G., & Rubin, D. B. (2007). Resolving the latent structure of schizophrenia endophenotypes using expectation-maximization-based finite mixture modeling. *Journal of Abnormal Psychology, 116,* 16–29. doi:10.1037/0021-843X.116.1.16

Lewis, A. (1938). States of depression: Their clinical and aetiological differentiation. *BMJ, 2*(4060), 875–878. doi:10.1136/bmj.2.4060.875

Linscott, R. J., & van Os, J. (2010). Systematic reviews of categorical versus continuum models in psychosis: Evidence for discontinuous subpopulations underlying a psychometric continuum. Implications for *DSM–V, DSM–VI,* and *DSM–VII. Annual Review of Clinical Psychology, 6,* 391–419. doi:10.1146/annurev.clinpsy.032408.153506

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499–532. doi:10.1207/s15327906mbr4104_4

Lubke, G., & Neale, M. C. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43,* 592–620. doi:10.1080/00273170802490673

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. doi:10.1037/1082-989X.7.1.19

Markon, K. E. (2010a). How things fall apart: Understanding the nature of internalizing through its relationship with impairment. *Journal of Abnormal Psychology, 119,* 447–458. doi:10.1037/a0019707

Markon, K. E. (2010b). Modeling psychopathology structure: A symptom-level analysis of Axis I and II disorders. *Psychological Medicine, 40,* 273–288. doi:10.1017/S0033291709990183

Markon, K. E., & Krueger, R. F. (2005). Categorical and continuous models of liability to externalizing disorders. *Archives of General Psychiatry, 62,* 1352–1359. doi:10.1001/archpsyc.62.12.1352

Markon, K. E., & Krueger, R. F. (2006). Information-theoretic latent distribution modeling: Distinguishing discrete and continuous latent variable models. *Psychological Methods, 11,* 228–243. doi:10.1037/1082-989X.11.3.228

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry, 176,* 249–252. doi:10.1192/bjp.176.3.249

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130,* 79–83. doi:10.1192/bjp.130.1.79

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241. doi:10.1146/annurev.psych.50.1.215

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley.

Meehl, P. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60,* 117–174. doi:10.1111/j.1467-6494.1992.tb00269.x

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). New York, NY: Wiley.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216. doi:10.1037/h0048070

Meehl, P. E., & Yonce, L. Y. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure) [Monograph]. *Psychological Reports, 74*(Suppl. 1-V74), 1059–1274.

Meehl, P. E., & Yonce, L. Y. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure) [Monograph]. *Psychological Reports, 78*(Suppl. 1-V78), 1091–1227.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. doi:10.1037/0003-066X.50.9.741

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia: Meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica, 119,* 252–265. doi:10.1111/j.1600-0447.2008.01326.x

Moreland, A. D., & Dumas, J. E. (2008). Categorical and dimensional approaches to the measurement of disruptive behavior in the preschool years: A meta-analysis. *Clinical Psychology Review, 28,* 1059–1070. doi:10.1016/j.cpr.2008.03.001

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and *DSM–V. Psychological Assessment, 21,* 302–312. doi:10.1037/a0016607

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic tests: Application of receiver operating characteristic analysis. *Archives of General Psychiatry, 44,* 550–555. doi:10.1001/archpsyc.1987.01800180068011

Nazikian, H., Rudd, R. P., Edwards, J., & Jackson, H. J. (1990). Personality disorder assessment for psychiatric inpatients. *Australian and New Zealand Journal of Psychiatry, 24,* 37–46.

Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and Psychopathology, 15,* 529–551. doi:10.1017/S0954579403000282

Pickles, A., Rowe, R., Simonoff, E., Foley, D., Rutter, M., & Silberg, J. (2001). Child psychiatric symptoms and psychosocial impairment: Relationship and prognostic significance. *British Journal of Psychiatry, 179,* 230–235.

Prisciandaro, J. J., & Roberts, J. E. (2009). A comparison of the predictive abilities of dimensional and categorical models of unipolar depression in the National Comorbidity Survey. *Psychological Medicine, 39,* 1087–1096. doi:10.1017/S0033291708004522

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Regier, D. A. (2007). Dimensional approaches to psychiatric classification: Refining the research agenda for *DSM–V*: An introduction. *International Journal of Methods in Psychiatric Research, 16,* S1–S5. doi:10.1002/mpr.209

Ritchie, K., & Touchon, J. (2000). Mild cognitive impairment: Conceptual basis and current nosological status. *Lancet, 355,* 225–228. doi:10.1016/S0140-6736(99)06155-3

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry, 126,* 983–987. doi:10.1176/appi.ajp.126.7.983

Ross, H. E., Gavin, D. R., & Skinner, H. A. (1990). Diagnostic validity of the MAST and the Alcohol Dependence Scale in the assessment of *DSM–III* alcohol disorders. *Journal of Studies on Alcohol, 51,* 506–513.

Rothery, P. (1979). A nonparametric measure of intraclass correlation. *Biometrika, 66,* 629–639. doi:10.1093/biomet/66.3.629

Ruscio, J., & Ruscio, A. M. (2002). A structure-based approach to psychological assessment: Matching measurement models to latent structure. *Assessment, 9,* 4–16. doi:10.1177/1073191102091002

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research, 42,* 349–386. doi:10.1080/00273170701360795

Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2006). Semi-nonparametric methods for detecting latent non-normality: A fusion of latent trait and latent class modeling. *Multivariate Behavioral Research, 41,* 427–443. doi:10.1207/s15327906mbr4104_1

Sherbourne, C. D., Wells, K. B., Hays, R. D., Rogers, W., Burnam, A., & Judd, L. L. (1994). Subthreshold depression and depressive disorder: Clinical characteristics of general medical and mental health specialty outpatients. *American Journal of Psychiatry, 151,* 1777–1784.

Skodol, A. E., Oldham, J. M., Rosnick, L., Kellman, H. D., & Hyler, S. E. (1991). Diagnosis of *DSM–III–R* personality disorders: A comparison of two structured interviews. *International Journal of Methods in Psychiatric Research, 1,* 13–26.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17,* 396–408. doi:10.1037/1040-3590.17.4.396

Swets, J. A. (1988, June 3). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293. doi:10.1126/science.3287615

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association, 82*(398), 559–567. doi:10.2307/2289465

Trull, T. J., & Durrett, C. A. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology, 1,* 355–380. doi:10.1146/annurev.clinpsy.1.102803.144009

Waldman, I. D., & Lilienfeld, S. O. (2001). Applications of taxometric methods to problems of comorbidity: Perspectives and challenges. *Clinical Psychology: Science and Practice, 8,* 520–527. doi:10.1093/clipsy.8.4.520

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Thousand Oaks, CA: Sage.

Watson, D. (2003). Investigating the construct validity of the dissociative taxon: Stability analyses of normal and pathological dissociation. *Journal of Abnormal Psychology, 112,* 298–305. doi:10.1037/0021-843X.112.2.298

Widiger, T. (1992). Categorical versus dimensional classification: Implications from and for research. *Journal of Personality Disorders, 6,* 287–300. doi:10.1521/pedi.1992.6.4.287

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for the *Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition. Journal of Abnormal Psychology, 114,* 494–504. doi:10.1037/0021-843X.114.4.494

Zimmerman, M., & Coryell, W. (1989). The reliability of personality disorder diagnoses in a nonpatient sample. *Journal of Personality Disorders, 3,* 53–57. doi:10.1521/pedi.1989.3.1.10

Zimmerman, M., Pfohl, M., Coryell, W., Stangl, D., & Corenthal, C. (1988). Diagnosing personality disorder in depressed patients: A comparison of patient and informant interviews. *Archives of General Psychiatry, 45,* 733–737. doi:10.1001/archpsyc.1988.01800320045005

Appendix A

Studies Included in Reliability Meta-Analysis

Brown, T. A., Di Nardo, P. A., Lehman, C. L., & Campbell, L. A. (2001). Reliability of *DSM–IV* anxiety and mood disorders: Implications for classification of emotional disorders. *Journal of Abnormal Psychology, 110,* 49–58. doi:10.1037//0021-843X.110.1.49

Cacciola, J. S., Rutherford, M. J., Alterman, A. I., McKay, J. R., & Mulvaney, F. D. (1998). Long-term test–retest reliability of personality disorder diagnoses in opiate dependent patients. *Journal of Personality Disorders, 12,* 332–337. doi:10.1521/pedi.1998.12.4.332

Chanen, A. M., Jackson, H. J., McGorry, P. D., Allot, K. A., Clarkson, V., & Yuen, H. P. (2004). Two-year stability of personality disorder in older adolescent outpatients. *Journal of Personality Disorders, 18,* 526–541. doi:10.1521/pedi.18.6.526.54798

Clark, L. A. (1999). Dimensional approaches to personality disorder assessment and diagnosis. In C. R. Cloninger (Ed.), *Personality and psychopathology* (pp. 219–244). Arlington, VA: American Psychiatric Press.

Critchfield, K. L., Levy, K. N., & Clarkin, J. F. (2007). The Personality Disorders Institute/Borderline Personality Disorder Research Foundation randomized control trial for borderline personality disorder: Reliability of Axis I and II diagnoses. *Psychiatric Quarterly, 78,* 15–24. doi:10.1007/s11126-006-9023-x

Damen, K. F. M., De Jong, C. A. J., & Van der Kroft, P. J. A. (2004). Interrater reliability of the structured interview for *DSM–IV* personality in an opioid-dependent patient sample. *European Addiction Research, 10*, 99–104. doi:10.1159/000077697

Dreessen, L., & Arnoud, A. (1998). Short-interval test–retest interrater reliability of the Structured Clinical Interview for *DSM–III–R* Personality Disorders (SCID–II) in outpatients. *Journal of Personality Disorders, 12,* 138–148. doi:10.1521/pedi.1998.12.2.138

Ferro, T., Klein, D. N., Schwartz, J. E., Kasch, K. L., & Leader, M. A. (1998). Thirty-month stability of personality disorder diagnoses in depressed outpatients. *American Journal of Psychiatry, 155,* 653–659.

Grant, B. F., Dawson, D. A., Stinson, F. S., Chou, S. P., Kay, W. & Pickering, R. P. (2003). The Alcohol Use Disorder and Associated Disabilities Interview Schedule–IV (AUDADIS-IV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug and Alcohol Dependence, 71,* 7–16. doi:10.1016/S0376-8716(03)00070-X

Hogg, B., Jackson, H. J., Rudd, R. P., & Edwards, J. (1990). Diagnosing personality disorders in recent-onset schizophrenia. *Journal of Nervous and Mental Disease, 178,* 194–199.

Hove, O., & Havik, O. E. (2008). Psychometric properties of Psychopathology checklists for Adults with Intellectual Disability (P-AID) on a community sample of adults with intellectual disability. *Research in Developmental Diseases, 29,* 467–482. doi:10.1016/j.ridd.2007.09.002

Jackson, H. J., Gazis, J., Rudd, R. P., & Edwards, J. (1991). Concordance between two personality disorder instruments with psychiatric inpatients. *Comprehensive Psychiatry, 32,* 252–260. doi:10.1016/0010-440X(91)90046-F

Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., . . . Regier, D. A. (1994). The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. *Archives of General Psychiatry, 51,* 215–224.

Loranger, A. W., Susman, V. L., Oldham, J. M., & Russakoff, L. M. (1987). The Personality Disorder Examination: A preliminary report. *Journal of Personality Disorders, 1,* 1–13. doi:10.1521/pedi.1987.1.1.1

Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., . . . Petrachi, M. (1997). Interrater reliability and internal consistency of the Structured Clinical Interview for *DSM–IV* Axis II Personality Disorders (SCID–II), version 2.0. *Journal of Personality Disorders, 11,* 279–284. doi:10.1521/pedi.1997.11.3.279

Molinari, V., Kunik, M. E., Mulsant, B., & Rifai, A. H. (1998). The relationship between patient, informant, social worker, and consensus diagnoses of personality disorder in elderly depressed inpatients. *American Journal of Geriatric Psychiatry, 6,* 136–144.

Nazikian, H., Rudd, R. P., Edwards, J., & Jackson, H. J. (1990). Personality disorder assessment for psychiatric inpatients. *Australian and New Zealand Journal of Psychiatry, 24,* 37–46.

Ottosson, H., Bodlund, O., Ekselius, L., Grann, M., von Knorring, L., Kullgren, G., . . . Söderberg, S. (1998). *DSM–IV* and *ICD–10* personality disorders: A comparison of a self-report questionnaire (DIP-Q) with a structured interview. *European Psychiatry, 13,* 246–253. doi:10.1016/S0924-9338(98)80013-8

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 3,* 46–54.

Rounsaville, B. J., Kranzler, H. R., Ball, S., Tennen, H., Poling, J., & Triffleman, E. (1998). Personality disorders in substance abusers: Relation to substance use. *Journal of Nervous and Mental Disease*, *186,* 87–95.

Schneider, B., Maurer, K., Sargk, D., Heiskel, H., Weber, B., Frölich, L., . . . Seidler, A. (2004). Concordance of *DSM–IV* Axis I and II diagnoses by personal and informant's interview. *Psychiatry Research, 127,* 121–136. doi:10.1016/j.psychres.2004.02.015

Schotte, C. K., De Doncker, D., Dmitruk, D., van Mulders, I., D'Haenen, H., & Cosyns, P. (2004). The ADP–IV Questionnaire: Differential validity and concordance with the semi-structured interview. *Journal of Personality Disorders, 18,* 405–419. doi:10.1521/pedi.18.4.405.40348

(*Appendices continue*)

Shankman, S. A., & Klein, D. N. (2002). Dimensional diagnosis of depression: Adding the dimension of course to severity, and comparison to the *DSM*. *Comprehensive Psychiatry, 43,* 420–426. doi:10.1053/comp.2002.35902

Skodol, A. E., Oldham, J. M., Bender, D. S., Dyck, I. R., Stout, R. L., Morey, L. C., . . . Gunderson, J. G. (2005). Dimensional representations of *DSM–IV* personality disorders: Relationships to functional impairment. *American Journal of Psychiatry, 162,* 1919–1925. doi:10.1176/appi.ajp.162.10.1919

Vandiver, T., & Sher, K. (1991). Temporal stability of the Diagnostic Interview Schedule. *Psychological Assessment, 3,* 277–281. doi:10.1037/1040-3590.3.2.277

Weertman, A., Arntz, A., Dreessen, L., van Velzen, C., & Vertommen, S. (2003). Short-interval test–restest interrater reliability of the Dutch version of the Structured Clinical Interview for *DSM–IV* Personality Disorders (SCID–II). *Journal of Personality Disorders, 17,* 562–567. doi:10.1521/pedi.17.6.562.25359

Widiger, T. A., Trull, T. J., Hurt, S. W., Clarkin, J., & Frances, A. (1987). A multidimensional scaling of the *DSM–III* personality disorders. *Archives of General Psychiatry, 44,* 557–563.

Zanarini, M. C., & Frankenburg, F. R. (2001). Attainment and maintenance of reliability of Axis I and II disorders over the course of a longitudinal study. *Comprehensive Psychiatry, 42,* 369–374. doi:10.1053/comp.2001.24556

Zanarini, M. C., Frankenburg, F. R., & Vujanovic, A. (2002). Inter-rater and test–retest reliability of the Revised Diagnostic Interview for Borderlines. *Journal of Personality Disorders, 16,* 270–276. doi:10.1521/pedi.16.3.270.22538

Zanarini, M. C., Skodol, A. E., Bender, D., Dolan, R., Sanislow, C., Schaefer, E., . . . Gunderson, J. G. (2000). The collaborative longitudinal personality disorders study: Reliability of Axis I and II diagnoses. *Journal of Personality Disorders, 14,* 291–299. doi:10.1521/pedi.2000.14.4.291

Zimmerman, M., & Coryell, W. (1989). The reliability of personality disorder diagnoses in a nonpatient sample. *Journal of Personality Disorders, 3,* 53–57. doi:10.1521/pedi.1989.3.1.10

Appendix B

Studies Included in Validity Meta-Analysis

Angst, J., & Merikangas, K. (2001). Multi-dimensional criteria for the diagnosis of depression. *Journal of Affective Disorders, 62,* 7–15. doi:10.1016/S0165-0327(00)00346-3

Blanchard, J. J., Horan, W. P., & Collins, L. M. (2005). Examining the latent structure of negative symptoms: Is there a distinct subtype of negative symptom schizophrenia? *Schizophrenia Research, 77,* 151–165. doi:10.1016/j.schres.2005.03.022

Edens, J. F., Marcus, D. F., & Morey, L. C. (2009). Paranoid personality has a dimensional latent structure: Taxometric analysis of community and clinical samples. *Journal of Abnormal Psychology, 118,* 545–553. doi:10.1037/a0016313

Fergusson, D. M., & Horwood, L. J. (1995). Predictive validity of categorically and dimensionally scored measures of disruptive childhood behaviors. *Journal of the American Academy of Child & Adolescent Psychiatry, 34,* 477–487. doi:10.1097/00004583-199504000-00015

Fossati, A., Maffei, C., Bagnato, M., Donati, D., Donini, M., Fiorilli, M., . . . Ansoldi, M. (1998). Brief communication: Criterion validity of the Personality Diagnostic Questionnaire–4+ (PDQ–4+) in a mixed psychiatric sample. *Journal of Personality Disorders, 12,* 172–178. doi:10.1521/pedi.1998.12.2.172

Giesbrecht, T., Merckelbach, H., & Geraerts, E. (2007). The dissociative experiences taxon is related to fantasy proneness. *Journal of Nervous and Mental Disease, 195,* 769–772. doi:10.1097/NMD.0b013e318142ce55

Goedeker, K. C., & Tiffany, S. T. (2008). On the nature of nicotine addiction: A taxometric analysis. *Journal of Abnormal Psychology, 117,* 896–909. doi:10.1037/a0013296

Guy, L. S., Poythress, N. G., Douglas, K. S., Skeem, J. L., & Edens, J. F. (2008). Correspondence between self-report and interview-based assessments of antisocial personality disorder. *Psychological Assessment, 20,* 47–54. doi:10.1037/1040-3590.20.1.47

Hasin, D. S., Liu, X., Alderson, D., & Grant, B. F. (2006). *DSM–IV* alcohol dependence: A categorical or dimensional phenotype? *Psychological Medicine, 36,* 1695–1705. doi:10.1017/S0033291706009068

Heath, A. C., Whitfield, J. B., Madden, P. A. F., Bucholz, K. K., Dinwiddie, S. H., Slutske, W. S., . . . Martin, N. G. (2001). Towards a molecular epidemiology of alcohol dependence: Analysing the interplay of genetic and environmental risk factors. *British Journal of Psychiatry, 178*(Suppl. 40), s33–s40. doi:10.1192/bjp.178.40.s33

Hyler, S. E., Rieder, R. O., Williams, J., Sptizer, R. L., Lyons, M., & Hendler, J. (1989). A comparison of clinical and self-report diagnoses of *DSM–III* personality disorders in 552 patients. *Comprehensive Psychiatry, 30,* 170–178. doi:10.1016/0010-440X(89)90070-9

(*Appendices follow*)

Kavoussi, R. J., Coccaro, E. F., Klar, H. M., Bernstein, D., & Siever, L. J. (1990). Structured interviews for borderline personality disorder. *American Journal of Psychiatry, 147*, 1522–1525.

Morey, L. C., Hopwood, C. J., Gunderson, J. G., Skodol, A. E., Shea, M. T., Yen, S., . . . McGlashan, T. H. (2007). Comparison of alternative models for personality disorders. *Psychological Medicine, 37,* 983–994. doi:10.1017/S0033291706009482

Morey, L. C., Warner, M. B., Shea, M. T., Gunderson, J. G., Sanislow, C. A., Grilo, C., . . . McGlashan, T. H. (2003). The representation of four personality disorders by the Schedule for Nonadaptive and Adaptive Personality dimensional model of personality. *Psychological Assessment, 15,* 326–332. doi:10.1037/1040-3590.15.3.326

Morrison, C. H. (2002). *Toward a clinical–empirical classification of eating disorders* (Unpublished doctoral dissertation). Boston University.

Peralta, V., Cuesta, M., Giraldo, C., Cardenas, A., & Gonzalez, F. (2002). Classifying psychotic disorders: Issues regarding categorial vs. dimensional approaches and time frame to assess symptoms. *European Archives of Psychiatry and Clinical Neuroscience, 252,* 12–18. doi:10.1007/s004060200002

Prisciandaro, J. J., & Roberts, J. E. (2009). A comparison of the predictive abilities of dimensional and categorical models of unipolar depression in the National Comorbidity Survey. *Psychological Medicine, 39*, 1087–1096. doi:10.1017/S0033291708004522

Rosenman, S., Korten, A., Medway, J., & Evans, M. (2003). Dimensional vs. categorical diagnosis in psychosis. *Acta Psychiatrica Scandinavica, 107,* 378–384. doi:10.1034/j.1600-0447.2003.00059.x

Ross, H. E., Gavin, D. R., & Skinner, H. A. (1990). Diagnostic validity of the MAST and the Alcohol Dependence Scale in the assessment of *DSM–III* alcohol disorders. *Journal of Studies on Alcohol, 51,* 506–513.

Skodol, A. E., Oldham, J. M., Rosnick, L., Kellman, H. D., & Hyler, S. E. (1991). Diagnosis of *DSM–III–R* personality disorders: A comparison of two structured interviews. *International Journal of Methods in Psychiatric Research, 1*, 13–26.

Smith, T., Klein, M., & Benjamin, L. (2003). Validation of the Wisconsin Personality Disorders Inventory–IV with the SCID–II. *Journal of Personality Disorders, 17,* 173–187. doi:10.1521/pedi.17.3.173.22150

Van Os, J., Fahy, T., Jones, P., Harvey, I., Sham, P., Lewis, S., . . . Murray, R. (1996). Psychopathological syndromes in the functional psychoses: Associations with course and outcome. *Psychological Medicine, 26,* 163–176. doi:10.1017/S0033291700033808

Van Os, J., Gilvarry, C., Bale, R., Van Horn, E., Tattan, T., White, I., & Murray, R. (1999). A comparison of the utility of dimensional and categorical representations of psychosis. *Psychological Medicine, 29,* 595–606.

Waller, N. G., & Ross, C. A. (1997). The prevalence and biometric structure of pathological dissociation in the general population: Taxometric and behavior genetic findings. *Journal of Abnormal Psychology, 106,* 499–510. doi:10.1037/0021-843X.106.4.499

Yang, J., McCrae, R. R., Costa, P. T., Jr., Yao, S., Dai, X., Cai, T., & Gao, B. (2000). The cross-cultural generalizability of Axis-II constructs: An evaluation of two personality disorder assessment instruments in the People's Republic of China. *Journal of Personality Disorders, 14,* 249–263. doi:10.1521/pedi.2000.14.3.249

Zimmerman, M., & Coryell, W. (1990). Diagnosing personality disorders in the community: A comparison of self-report and interview measures. *Archives of General Psychiatry, 47*, 527–531. doi:10.1001/archpsyc.1990.01810180027005

Zimmerman, M., Pfohl, B., Coryell, W., Stangl, D., & Corenthal, D. (1988). Diagnosing personality disorder in depressed patients: A comparison of patient and informant interviews. *Archives of General Psychiatry, 45*, 733–737. doi:10.1001/archpsyc.1988.01800320045005

Received April 21, 2010
Revision received April 16, 2011
Accepted April 18, 2011 ■

Correction to Markon et al. (2011)

The article "The Reliability and Validity of Discrete and Continuous Measures of Psychopathology: A Quantitative Review" by Kristian E. Markon, Michael Chmielewski, and Christopher J. Miller (*Psychological Bulletin,* 2011, Vol. 137, No. 5, pp. 856–879) contained a production-related error.

In the Samples section of Meta-Analysis 1: Reliability, third paragraph, the number of studies reporting data on clinical samples is incorrect. The sentence "Four studies included clinical samples, and eight studies included nonclinical samples" should read "Twenty-four studies included clinical samples, and eight studies included nonclinical samples."

DOI: 10.1037/a0025727