American Journal of Orthopsychiatry
2009, Vol. 79, No. 4, 452–460


0002-9432/09/$12.00 DOI: 10.1037/a0016508

# Revising the Personality Disorder Diagnostic Criteria for the *Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition (DSM-V):* Consider the Later Life Context

Steve Balsis, PhD
Texas A&M University

Daniel L. Segal, PhD
University of Colorado at Colorado Springs

Cailin Donahue, BA
Washington University in St. Louis

The categorical measurement approach implemented by the *Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition* (*DSM–IV*) personality disorder (PD) diagnostic system is theoretically and pragmatically limited. As a result, many prominent psychologists now advocate for a shift away from this approach in favor of more conceptually sound dimensional measurement. This shift is expected to improve the psychometric properties of the personality disorder (PD) diagnostic system and make it more useful for clinicians and researchers. The current article suggests that despite the probable benefits of such a change, several limitations will remain if the new diagnostic system does not closely consider the context of later life. A failure to address the unique challenges associated with the assessment of personality in older adults likely will result in the continued limited validity, reliability, and utility of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) system for this growing population. This article discusses these limitations and their possible implications.

*Keywords:* diagnosis, elderly, personality disorder, psychometric

There is solid consensus among clinicians and researchers that the diagnostic category of personality disorders (PDs) as conceptualized in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association, 2000) is severely flawed. Indeed, it has been argued that the PD criteria have limited reliability, validity, and utility (Widiger & Trull, 2007). Many believe that these limitations stem from the categorical measurement system currently implemented by the *DSM* (e.g., Widiger & Samuel, 2005). This categorical system is thought to be problematic because it cannot reflect the actual nature of PD pathology, which exists along several associated dimensions (e.g., Krueger, 2006). One might consider this classification system to be poorly conceived, in this way, because it cannot accurately capture the phenomena it was designed to measure.

Before discussing some of the problems of using a categorical system, it is important to highlight some of its strengths and review the initial justification for its use (for a more complete discussion, see Ruscio, 2008; Widiger & Trull, 2007). In its inception, the categorical model for measuring all forms of psychopathology, including PDs, had converging support. Specifically, a categorical diagnosis was thought to be easily understandable, allowing for clear boundaries between disorder and nondisorder, in line with the medical model underpinnings of the DSM system. It also was thought to be conceptually consistent with many decisions that clinicians make. Clinicians are often asked to decide whether a particular person should enter treatment or not, to refer a patient to a specialist or not, and to hospitalize a suicidal patient or not. A categorical system that could demarcate clear cutoffs held promise for being helpful for clinicians who needed to make such decisions "in the trenches." In addition to these issues of clarity and decision making, in a clinical setting, it was thought (and still is) that more refined appraisals of pathology may not be needed. A categorical approach may be sufficient because it can allow for simple and user-friendly measurement scales, features of scales that are becoming more necessary in time-pressured professional settings. When considered together, all of these benefits of the categorical approach provide a compelling rationale for its implementation. Although there may be some real potential benefits of using a categorical approach, there may also be some serious psychometric issues that arise when a categorical approach is used to measure personality pathology. The first goal of this article is to analyze this latter point—to fully consider the psychometric issues that arise from using the categorical approach.

Given the potential psychometric limitations associated with the current categorical system, it is not surprising that many psychologists advocate for dramatic revisions. Almost uniformly, conver-

Steve Balsis, PhD, Department of Psychology, Texas A&M University; Daniel L. Segal, PhD, Department of Psychology, University of Colorado at Colorado Springs; Cailin Donahue, BA, Department of Psychology, Washington University in St. Louis.

This research was supported in part by a grant from the National Institute of Mental Health, 1F31-MH075336-01A1, awarded to Steve Balsis. We thank Lisa Geraci and Matthew Wyrick for helpful comments on a draft of this article.

For reprints and correspondence: Steve Balsis, Texas A&M University, Department of Psychology, 4235 TAMU, College Station, TX 77843. E-mail: balsis@tamu.edu

sations regarding these revisions revolve around a transition to a more conceptually consistent dimensional system (e.g., Hankin, Fraley, Lahey, & Waldman, 2005; Krueger, 1999; Widiger & Clark, 2000). Most believe that such a change will improve the measurement properties of the diagnostic criteria (Widiger & Simonsen, 2005). These thoughts have been expressed in a variety of publications, most notably special issues devoted to the topic in the *Journal of Abnormal Psychology* (2005, Volume 114) and the *Journal of Personality Disorders* (2005, Volume 19). Indeed, these debates and suggestions for diagnostic improvements are timely and important given the active development of the *Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition* (*DSM-V*).

One major issue not adequately considered in the discussion surrounding the *DSM* revisions is the current mischaracterization of PDs in later life. The lack of attention to later life is remarkable, given that it likely influences clinicians' and researchers' abilities to conduct reliable and valid assessments on a large (and growing) segment of the population (Rosowsky, Abrams, & Zweig, 1999; Segal, Coolidge, & Rosowsky, 2006). The implications extend even beyond the utility of the criteria for clinicians and researchers. For instance, serious questions can be raised about the viability of theories that rest on data linked to the current criteria. The second aim of this article is to illustrate the potential measurement problems in any PD classification system (categorical or dimensional) that does not closely consider the context of later life.

## Validity Considered

### *Face Validity*

An item has good face validity if it measures what it intuitively appears to measure. For example, an item like, "I feel sad," that is scored on a scale from 1 (*not at all*) to 10 (*extremely*) has high face validity for measuring sadness. This item not only explicitly addresses the concept of sadness, but it also implements a dimensional (or incremental) rating scale for measuring the concept; this dimensional rating scale probably closely mirrors the actual dimensional nature of the concept. Face validity, then, has two parts. One part reflects the content of the item (what we refer to as content face validity), which must be intuitive and clear so that each respondent's interpretation of the item's meaning is similar. A second part reflects the scale of the item (what we refer to as scaling face validity) and requires that the format of the particular scale allow for the expression of all various possible incarnations of the phenomenon. For an item to have good face validity, it must be both intuitive and clear and must also be scored on a scale that approximates the true nature of the phenomenon of interest.

A categorical system that is used to measure dimensional phenomena necessarily has poor scaling face validity. Consider that people vary in the degree of sadness that they experience at a particular time. One person may feel rather sad, but not extremely sad. Another person may feel just a little bit sad. Although a 10-point dimensional (incremental) scale can roughly capture the amount of sadness felt by both of these individuals (7 may indicate rather sad, 3 may indicate a little bit sad), a categorical (binary) rating scale that requires a categorical yes/no decision cannot even roughly capture the degrees of sadness felt by these individuals. Both people in this example may have responded, "yes" to the item "I feel sad," because both do genuinely feel some degree of sadness. This categorical scale, accordingly, would not detect the difference in severity of sadness experienced by these two individuals.

Notice, however, that the item, "I feel sad" has good content face validity regardless of its associated scale. The item is intuitive and clear with little or no ambiguity. Like this item, most *DSM–IV* PD criteria contain relatively good content face validity. These criteria purportedly measure PD pathology, and on the surface the criteria do indeed seem to fulfill this ambition. Take for example, the DSM item for dependent PD that states, "Urgently seeks another relationship as a source of care and support when a close relationship ends." This item is intuitively related to dependent PD pathology, and there is no immediate reason to believe that it does not measure some aspect of dependency.

The fundamental problem with face validity in the DSM PD items, according to most previous recommendations for their revision, centers on scaling face validity not content face validity (e.g., Widiger & Trull, 2007). Indeed, at present each PD criterion must be scored on a yes/no scale. Either a client does or does not meet the criterion. Consider the PD criterion, "*Urgently* seeks another relationship as a source of care and support when a close relationship ends." A client who actively but *not urgently* seeks other relationships must score either a "yes" or a "no" on this item. A score of "yes" will overestimate this dependent PD feature in this client, and a score of "no" will grossly underestimate it. On a yes/no scale, there is no room to indicate that this individual possesses only a degree of the PD feature.

The poor scaling face validity in the DSM has implications not only at the item level but also at the diagnostic level. For example, a person who meets 4, 5, 6, or 7 of the diagnostic criteria for avoidant PD can be diagnosed with the disorder. Yet a person who meets 0, 1, 2, or 3 of the criteria cannot be diagnosed with the disorder. This categorical differentiation, where a person can or cannot be diagnosed does not accurately reflect the underlying nature of the latent variable, which exists on a continuum (for a discussion on the dimensional nature of latent constructs, see Barrett, Petrides, Eysenck, & Eysenck, 1998; Cloninger, Przybeck, Svrakic, & Wetzel, 1994; Durrett & Westen, 2005; Kernberg, 1996; Livesley, 2005; Thomas, Turkheimer, & Oltmanns, 2003). Many people may have some degree of avoidant PD pathology, and a present/not present differentiation is a poor reflection of the underlying level of the latent PD pathology. The problems with poor scaling face validity in the *DSM* are problems for people of any age or population, but they can be largely repaired by the mere transition to a dimensional classification system.

Although making a shift from a categorical system to a dimensional one will solve problems with scaling face validity, a unique problem with content face validity emerges when the DSM items are applied to older adults. For younger adults, the schizoid item, "Almost always chooses solitary activities," seems intuitively related to schizoid PD pathology. However, when this item (and many others) is considered in a later life context, it becomes apparent that there is, indeed, poor content face validity in the DSM system. Many older adults may choose solitary activities for reasons unrelated to the underlying level of their schizoid PD pathology. For example, they may choose solitary activities because of medical illnesses that limit mobility, lack of adequate transportation to social activities, or a diminishing number of close social networks. This item, when applied to the context of later

life, may reflect issues beyond those associated with schizoid PD pathology.

Take for another example the criterion for schizoid PD, "Neither enjoys nor experiences sexual relations." Responses to this item may have little to do with schizoid PD pathology but may be closely linked to age-related physiological changes that make it unlikely or unenjoyable for some older adults to have sex or it simply may reflect the lack of suitable sexual partners for many in the older cohort, especially older women (Segal et al., 2006). Given the gross lack of face validity in just these two items, it is not surprising that some researchers have called for a revision to nearly 25% of the PD criteria for use with older adults (Agronin & Maletta, 2000). Thus, we see that the content face validity of the DSM is potentially very limited for application to many individuals in later life (whether the items are scaled categorically or dimensionally).

### *Content Validity*

Content validity indicates how well a test or measure samples all aspects of a concept, in this case, PD pathology. For a measure to have high content validity, it must measure the PD pathology broadly and give appropriate weight to different features of the pathology. Although the current DSM system has a moderate amount of content validity, there is clear room for improvement (Verheul & Widiger, 2004; Widiger & Trull, 2007). Consider that a formal diagnosis of "PD Not Otherwise Specified" is applied to individuals whose unique blend of personality pathology does not fit neatly into one of the 10 PD categories, which represent "pure" types or prototypical versions of the disorders. In their review, Widiger and Trull (2007) suggest that this "wastebasket" category is used often in clinical practice, suggesting that PDs come in many forms, and that the current diagnostic system does not adequately capture the full range of PD pathology. Under the current system, a client may have a particular constellation of PD symptoms, but the client may not receive one of the 10 PD diagnoses. Essentially, that client is placed into a nearly meaningless category that signifies many things all at once. A dimensional system, on the other hand, would be organized so that a client's unique personality profile would not be lost in the diagnostic process (Widiger & Samuel, 2005). Each profile would be unique and would capture subtleties of a client's personality.

Although this problem with content validity exists when the criteria are applied to members of any age group, a unique problem with content validity arises when the criteria are applied to individuals in later life. In fact, many of the DSM criteria seem to have been written to describe PD features as manifested in younger adults (see Segal et al., 2006 for a thorough critique of many individual PD criteria for older adults). This focus on younger adults leaves open the possibility that there may be many additional PD features that apply exclusively to later life, and it further remains likely that these features are yet to be fully understood and considered. As an example, consider the antisocial PD criterion "Irritability and aggressiveness, as indicated by repeated physical fights or assaults." This criterion may not apply to later life antisocial PD, because it might not be part of the typical symptom picture for older individuals with this disorder. Although the latent trait of aggression may manifest itself as a physical fight in younger adulthood, older adults may not have the sense of invulnerability needed to make physical fighting a viable option. Or getting into physical altercations may become a much less useful strategy for an older adult who is frail or lacks sheer physical stamina. Instead, aggression may manifest itself as an angry glare or as some other behavior that may be operationalized and behaviorally described. This likelihood that the same pathology presents itself differently at different life stages brings into question the content validity of criteria when they are applied to older adults. In other words, although some items may accurately assess PD pathology in younger adults, they may do so at the expense of inadequately assessing the full range of experiences common to PD pathology in later life.

Research supports the hypothesis that many items in the *DSM–IV* lack content validity when applied to the later life context. For example, a large-scale epidemiological study designed to examine the prevalence of, among other disorders and age groups, antisocial PD in people over age 55 (Narrow, Rae, Robins, & Regier, 2002) failed to find any cases of older adults with antisocial PD; in a subsample of 8,748 people over age 55 not a single person met this diagnosis. It is possible that these data reflect the true prevalence of antisocial PD in this population. Many people with antisocial PD are incarcerated, commit suicide, or simply do not participate in research (Robins, 1966). It also is possible, however, that the criteria are not appropriate for PD pathology assessment of older adults. A closer look at the specific criteria for antisocial PD shows that at least four of the seven criteria may be problematic when applied to older adults, specifically, (a) failure to conform to social norms with respect to lawful behaviors, as indicated by repeatedly performing acts that are grounds for arrest, (b) impulsivity or failure to plan ahead, (c) irritability and aggressiveness as indicated by repeated physical fights or assaults, and (d) consistent irresponsibility, as indicated by repeated failure to sustain consistent work behavior or honor financial obligations. These items may possess an inherent measurement bias that makes them unfit for the assessment of older adults (Agronin & Maletta, 2000). Because of the measurement bias in these items, they are unlikely to be endorsed by older adults. As a result, even an older adult with strong antisocial tendencies may not endorse enough items to receive a formal diagnosis.

Some researchers have suggested that the higher prevalence of PDs reported for younger adults compared with older adults simply indicates that PDs temper with age (Kenan et al., 2000; Paris, 2003). This hypothesis receives support from longitudinal data that show that PD pathology declines with age (see Paris, 2003, for a brief review). However, other studies have found that although some specific PD symptoms disappear with increasing age, significant personality problems nevertheless remain (Moffitt, Caspi, Harrington, & Milne, 2002). For example, a study that examined participants with PDs over a 33-year period found that, although specific behaviors required to meet a particular PD diagnosis declined with age, general social and interpersonal problems were still apparent (Drake & Vaillant, 1988). The conclusions of this study are consistent with the notion that the presentation of PD pathology may change with age. Thus, although it is possible that PD pathology may decrease, an alternate explanation for the apparent decrease in PD pathology with age is that the pathology simply presents itself in a different form, thereby remaining undetected by diagnostic criteria not designed for older people (Agronin & Maletta, 2000; Mroczek, Hurt, & Berman, 1999;

Segal, Hersen, Van Hasselt, Silberman, & Roth, 1996). In fact, it has been argued that although in many cases PD pathology becomes muted with age, there are also cases in which PD pathology remains the same across the adult life span, and yet other cases in which PD pathology becomes exacerbated in the late-life context, possibly reflecting an uncovering of pathology that was relatively quiescent during the adult years but becomes expressed in reaction to specific challenges and stressors associated with later life (e.g., increased dependency, sensory declines, loss of prestige and status, loss of significant others who may have minimized the expression or impact of the person's PD pathology; Sadavoy, 1987, 1996; Segal et al., 2006).

This idea that PD pathology may present differently in later life has been discussed elsewhere (e.g., Mroczek et al., 1999) and is consistent with many discussions of the heterotypic presentation of personality (e.g., Kagan, 1969). The important point to consider here is that the breadth of the current PD criteria or other measures based on the PD criteria may have limited content validity when applied to later life, because the criteria were written to most closely consider the breadth of PD pathology in younger adults. This focus on capturing the presentation of personality pathology in younger adulthood may have led to the neglect of some of the important presentations of the same personality features in later life. Even a shift to a dimensional classification system will not satisfactorily improve the content validity of the criteria for use with older adults if the dimensional system remains specific to a younger group.

### *Criterion Validity*

There are two types of criterion validity: concurrent validity and predictive validity (Segal & Coolidge, 2006a). Concurrent validity refers to the extent to which two different tests administered at the same time reflect a particular targeted phenomenon. Some degree of concurrent validity has been shown in the DSM classification of PDs. For example, many studies have shown that the DSM PD criteria share variance with several models of personality, including the Schedule for Nonadaptive and Adaptive Personality (Clark, McEwen, Collard, & Hickok, 1993), the Five Factor Model of Personality (see Morey & Zanarini, 2000), and the Millon Clinical Multiaxial Inventory (MCMI; Millon, Davis, & Millon, 1997). In contrast to concurrent validity, predictive validity refers to the extent to which a test administered at an earlier time predicts a subsequent outcome. Some degree of predictive validity also has been established for many DSM PD criteria. For younger people, PD pathology is associated positively with medical illness (e.g., Whiteman, Deary, & Fowkes, 2000), longer hospital stays (Spiessl, Hubner-Liebermann, Binder, & Cording, 2002), significantly greater later drug use and further psychiatric hospitalizations (e.g., Levy et al., 1999), and poorer treatment outcome (Gish et al., 2001). In addition, people with PDs are infamous for distressing hospital staff and other hospital residents. Although some degree of both concurrent validity and predictive validity have been established, it is expected that the associations between the DSM criteria and other concurrent or future outcome measures of PD pathology will improve when the DSM diagnostic system is scored dimensionally instead of categorically.

For an illustration of how criterion (concurrent or predictive) validity will improve, consider if the same participant completed a categorical measure of PD pathology (possible scores are 0 or 1), a dimensional measure of PD pathology (possible scores range from 1 to 10), and a target measure that represents either a redundant measure of personality used to establish concurrent validity or an outcome measure used to establish predictive validity (possible scores range from 1 to 10). Notice that scores of 0 on the categorical scale correspond to scores on the lower half (1 to 5) of the dimensional scale. Scores of 1 on the categorical scale correspond to scores on the upper half (6–10) of the dimensional scale. In this way, then, the categorical scale simply provides much less precise measure of the same personality.

Criterion validity can be established for the categorical scale by running a simple correlation between scores on it (column 1 of Table 1) and scores on the target measure (column 3). The resultant correlation ($r = .55$) indicates that there is moderate agreement between the categorical scale and the target measure. A more graded view of the same PD pathology, as measured with the dimensional scale, is depicted in column 2. The correlation between the dimensional scores and the target measure reflects a higher degree of concurrent validity. In this case the score improves substantially ($r = .84$) representing strong agreement between the two measures.

It is perhaps important to add that there are many possible patterns of scores that we could have represented in this table. However, we were unable to find a pattern where the categorical measure had greater criterion validity than the dimensional measure. We conclude that a categorical measure that artificially describes PD pathology in a binary fashion will tend to agree less with an external measure than a dimensional measure that more naturally reflects the construct at hand.

A different and more fundamental problem with criterion validity in the current DSM system arises when the PD items are applied to later life. Because the PD items were generally not designed to measure PD pathology in older adults (Balsis, Gleason, Woods, & Oltmanns, 2007; Segal et al., 2006), even a dimensional system may have less than optimal criterion validity because the PD items may reflect aspects related to aging, not aspects related to PD pathology. As noted earlier, an older adult who

Table 1
*Concurrent Validity Measuring Personality Disorder Pathology Categorically and Dimensionally*

| Item | Categorical measure | Dimensional measure | Concurrent measure |
|---|---|---|---|
| 1 | 0 | 3 | 4 |
| 2 | 0 | 5 | 6 |
| 3 | 1 | 8 | 9 |
| 4 | 0 | 2 | 1 |
| 5 | 0 | 5 | 6 |
| 6 | 1 | 7 | 7 |
| 7 | 0 | 4 | 6 |
| 8 | 1 | 6 | 5 |
| 9 | 1 | 7 | 5 |
| 10 | 1 | 9 | 8 |
| 11 | 0 | 5 | 6 |
| 12 | 1 | 9 | 10 |

*Note.* Correlation between categorical measure and concurrent measure, $r = .55$. Correlation between dimensional measure and concurrent measure, $r = .84$.

endorses the item, "Has little, if any, interest in having sexual experiences with another person," may do so for reasons associated with aging rather than reasons associated with schizoid PD pathology. The implication is that the diagnostic criteria may contain systematic measurement error when they are applied to later life. When establishing concurrent validity, this systematic error may reduce associations with other more appropriate measures of PD pathology. Correspondingly, when establishing predictive validity, this systematic error may also be the cause of weakened associations with outcome measures.

Predictive validity may be especially difficult to establish when the PD criteria are applied to later life, because there are a host of unique negative consequences of PD pathology in later life, some of which remain poorly understood (Rosowsky & Smyer, 1999). One easily can generate several hypotheses about how PD pathology in later life may have cascading effects for families and afflicted individuals. Consider that as most people age they rely more on immediate family members to meet their needs. At the same time, they tend to reduce contact with more distant relatives and friends (Antonucci & Akiyama, 1987). For people with PDs, later life may be especially difficult because they are likely to have chronically strained and poor relationships with family members, on whom they are sometimes forced to rely on during this time (for a relevant case study, see Siegel & Small, 1986). Consider, for example, a person who is egocentric, has difficulty seeing others' points of view, and acts selfishly, thus disrupting trust within the family. This type of broken trust, no matter when it happens in life, can plague a family for generations (Hargrave & Anderson, 1992), and may become especially impairing as families prepare to negotiate later life health care, financial challenges, and new living situations. These potential unique outcomes for PDs in later life, at both the family level and the individual level await future testing. To establish predictive validity in later life, however, we need psychometrically sound measures of PD pathology and further knowledge about the possible negative outcomes of PD pathology in this population.

## Reliability Considered

### *Alpha Reliability*

Not only are dimensional models more valid than categorical models, they also can be more reliable than categorical models. Internal consistency, often indicated by the level of alpha reliability, indicates how well items in a scale "hang together." Research has shown that the alpha reliability of each PD scale is moderate (Grilo et al., 2001); hence, the items for particular scales measure the same phenomena to some extent. The nine items for narcissistic PD, for instance, measure some aspect of narcissism. Nonetheless, a shift to a dimensional system should help improve the alpha reliability of the PD scales, because a dimensional system should contain less measurement error than a categorical system and allow items to share greater systematic (nonerror) variance.

A unique problem with alpha reliability arises when the PD criteria are applied to individuals in later life. Even dimensionally scaled items might be less internally consistent in an older adult sample for a variety of reasons. Most notably, some items that measure PD features in younger adults may measure different constructs in older adults. On the one hand, as one's life stage changes, an item like "Lacks empathy . . ." may continue to capture PD pathology well. On the other hand, as one grows older an item like, "Is unable to discard worn-out or worthless objects," may begin to capture something about the financial context of later life. In this particular case, the item would likely measure obsessive–compulsive PD pathology in younger adults but the financial context of aging among some older adults. When items on a particular PD scale measure different entities (here, genuine PD pathology vs. the behavioral manifestations or social contexts of aging), the measure by definition is not internally consistent, with this being statistically reflected by a lowered coefficient alpha.

In a recent empirical study of the reliability problem, Balsis & Cooper (2009) created hybrid diagnostic criteria by replacing underperforming DSM items (items that contained age-specific measurement bias) with items written specifically to measure PD pathology in later life. Findings indicated that the internal consistency (in this case represented by coefficient alpha) of the scales in this late life sample were higher for the hybrid criteria than the DSM criteria. In general, the alphas for the hybrid scales were as high or higher than the alphas of the DSM criteria. The alphas for the hybrid measures ranged from .76 to .90 with an average of .82. Meanwhile, the alphas for the DSM measures ranged from .54 to .87 with an average of .74. These analyses indicated that, on average, the hybrid measures had somewhat better internal consistency for use with this sample.

### *Test–Retest Reliability*

According to the *DSM–IV*, PDs are defined as pervasive patterns of inner experience and outer behavior that deviate markedly from cultural expectations. They are assumed to be stable over time. A high test–retest reliability would reflect this stability. In a recent study, Weertman, Arntz, Dreessen, van Velzen, and Vertommen (2003) found that the current version of the Structured Clinical Interview for *DSM–IV* Axis II Personality Disorders (SCID-II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997) exhibited sufficient test–retest reliability for PDs when they were categorically measured. For reasons that are unclear, the overall kappa value ($k = .63$) in this study was higher than past reliability studies that used the Diagnostic and Statistical Manual of Mental Disorders–Third Edition–Revised (*DSM–III–R*)-based diagnostic criteria for PDs.

Whether the test–retest reliability is higher for categorical or dimensional models remains an open empirical question. One negative consequence of the categorical approach is that it is insensitive to slight changes in latent pathology that may be sufficient to cross the threshold for instigating the PD. This means that in cases where clients are near threshold for meeting a particular PD diagnosis, small changes in personality over time may dramatically and negatively influence the test–retest reliability. Suppose an individual scored a 6 on a 1 (*no pathology*) to 10 (*much pathology*) dimensional scale at Time 1 and a 5 on the same scale at Time 2. This person's data would have quite good test–retest reliability, because the score of 6 at Time 1 is similar to the score of 5 at Time 2. Now suppose this same person's personality was measured on a categorical yes (met diagnosis)/no (did not meet diagnosis) scale. The person may be coded a "yes" at Time 1 because he was just above the threshold for diagnosis, but the

person may be coded a "no" at Time 2 because he was just below the threshold. This categorical scale would grossly misrepresent this slight change in personality pathology, and the test–retest reliability would suffer accordingly.

Despite instances like this one in which the categorical model would fail to show high test–retest reliability, there does remain (as mentioned earlier) some degree of moderate test–retest reliability in PD pathology measured categorically in younger adults. This finding makes some sense because personality is relatively stable during younger adulthood, and the personality criteria may have been designed to reflect this stability during this period of life.

In contrast to younger adulthood, later life is a dynamic period that has shifting contexts. Rapid changes can be brought on by loss, maturation, therapeutic interventions, acute illness, cognitive impairment, and other social changes, many of which are commonly experienced by older adults. The PD diagnostic criteria were not designed to reflect personality stability during these shifting contexts. For example, just before retirement, a participant may respond "yes" to the item, "Avoids occupational activities that involve significant interpersonal contact with others." Just after retirement (after a change in the context but not in the latent pathology), the participant would respond "no" to the same item, because the context no longer applies. These types of context changes in later life can have artificial influences on personality data measured at two time points, negatively influencing the test–retest reliability for older adults.

### *Interrater Reliability*

Interrater reliability can be defined as the agreement between two independent raters of a given phenomenon (Segal & Coolidge, 2006b). Generally, interrater reliability is higher for dimensional ratings as compared to categorical ratings. In a recent study, Jane, Pagan, Turkheimer, Fiedler, and Oltmanns (2006) compared the interrater reliability of the same *DSM–IV* PDs when they were measured both categorically and dimensionally. Kappas for disorders measured categorically ranged from chance ($k = -.01$) for schizoid PD to very good ($k = .85$) for avoidant PD, with only moderate agreement ($k = .50$) across all PDs. When considering the data (from the same measurement tool) dimensionally, the kappas were larger and ranged from good ($k = .77$) for histrionic PD to very good ($k = .93$) for avoidant PD, with very good agreement ($k = .84$) across all PDs. In a similar study, Heumann and Morey (1990) reported that categorical measures fared much worse than dimensional measures of the same borderline PD pathology. Specifically, intraclass correlations (ICCs) were low for clinicians' categorical judgments of whether a client described in a vignette met borderline PD (average ICC = .20) whereas the ICCs were much higher on four-dimensional measures of borderline pathology when using these same clinicians' ratings on the same vignette (average ICC = .57). The findings across these studies seem to fit with intuition. Categorical measurement of personality (a dimensional phenomenon) should be less reliable than dimensional measurement of it.

Interrater reliability may be particularly difficult to establish when the current PD criteria are applied to older adults. To illustrate, consider the dilemma posed by Balsis, Woods, Gleason, and Oltmanns (2007). A clinician who wants to assess DSM avoidant PD pathology in an older, retired client is eventually faced with several items that may not be appropriate for the client. Consider the item, "Avoids occupational activities that involve significant interpersonal contact with others." There are perhaps three tacks a clinician can take when faced with this item. First, the clinician can try to adjust the item to fit the older adult context. For example, the clinician could ask the client whether he or she avoids *volunteer* activities that involve significant interpersonal contact. The problem with this approach is that volunteer activities are qualitatively different from occupational activities. How would the clinician know whether such an adjustment to the item is still measuring the same basic avoidant PD feature? A second tack the clinician could take would be to assess the item by considering if it applied in the client's distant past. The problem with this approach is that there is no way to know whether the client's past behavior reflects their personality in the present time. The client's personality may well have matured or regressed somewhat since the referential point in time. Or, the retrospective memory may itself have some distortions especially because individuals with PDs typically have inaccurate self-perceptions (Segal et al., 2006). A third approach is to apply the item at face value. This approach is perhaps the least desirable, because the item does not apply to many individuals in the later life context. Whatever approach the clinician takes, it will introduce some degree of error to the item. One clinician may take one approach and another clinician may take a different approach. Using these sorts of nonstandardized approaches may lead to error in measurement and disagreement between raters. It is perhaps important to note that this problem with interrater reliability will exist for any items not written for the later life context, regardless of the type of scale (categorical vs. dimensional) implemented for the items.

## Utility Considered

### *Implications for Clinical Practice*

Dimensional measures of PD pathology are expected to have more clinical utility than categorical measures. Whereas a clinician using a categorical approach is restricted to establish (or not establish) a discrete diagnosis, a clinician using a dimensional approach can more flexibly assess both the type and degree of PD pathology with all of its subtle shades and hues. This flexibility in assessment can and should lead to flexibility in treatment decisions. Take for example a man who is just below threshold for a diagnosis of narcissistic PD. If assessed with a categorical measure, the man would be mischaracterized as pathology free. The treating clinician subsequently may approach the client's treatment without much consideration of the client's narcissistic tendencies. On the other hand, if assessed by a dimensional measure, this client would be characterized accurately as having some narcissistic tendencies. The treating clinician may use this information by applying techniques that have been shown to work well with slightly narcissistic individuals, or the clinician may derive informed treatment strategies based upon the accurate graded description of the client's personality. This simple example illustrates how a client who falls just short of a PD diagnosis can receive an accurate personality assessment when a dimensional (but not a categorical) measurement system is implemented. This accurate assessment can in turn lead to better-informed treatments (e.g., Trull, 2005).

Although it would seem likely that a dimensional system would lead to better treatments than a categorical system, a recent review showed that little is known about the utility of either approach (Verheul, 2005). Categorical models were only shown to enhance clinicians' abilities to communicate clearly. However, if the clinicians' communication is based on flawed categorical measurement, their communication, although clear, is likely to be flawed. Although it is unknown whether the application of dimensional models will enhance the clarity of clinical communication, one would assume that dimensional models will at least make clinical communication more accurate.

Although making a shift to a dimensional system may improve treatment decisions and clinical communication, a unique clinical problem will remain if the criteria continue to neglect the later life context. Consider the possibility that criteria written for younger adults may have negative consequences for the clinician-client relationship when the client is older (Agronin & Maletta, 2000). For example, such criteria may leave a clinician working with an older client with limited information during the diagnostic process, at least in some cases. As a result, it is possible that the clinician may overlook important aspects of personality or overemphasize ageist stereotypes. Take for example an older man with narcissistic PD. This man may be noncompliant with treatment, unnecessarily challenge doctors' orders, feel entitled to special treatment by hospital staff resources, and become angered when his needs are not immediately gratified. These tendencies reflecting personality pathology could possibly be dismissed by a clinician who views older adults as generally grumpy people. Such a dismissal would prevent the clinician from considering this client's unique personality and keep the clinician from working within the hospital and family systems to help this client receive the most effective care. If the treating clinician, however, understood that the client's personality features were consistent with narcissistic PD, the clinician could take a different, more effective tack and decide to adjust interventions based on these narcissistic tendencies.

### *Implications for Research*

The difficulty of the DSM diagnostic criteria to accurately capture PD pathology in later life has significant implications for research conducted on PDs in older adults. As an example, consider the implications on one important area of research: epidemiology. Currently, the consensus from epidemiological studies of PDs is that prevalence rates for diagnoses are lower in later life. However, if the instruments used in these studies do not accurately test for PDs in older adults, then the reported prevalence rates may be inaccurate. Furthermore, recent studies have revealed that each PD may exhibit a different prevalence trend over the course of a lifetime. Balsis, Woods, et al. (2007) estimated the over and underdiagnosis of various PDs and found statistical evidence that the literature has likely overestimated the prevalence of obsessive–compulsive and schizoid PD pathology in older adults and underestimated the prevalence of avoidant and dependent PDs in older adults. Thus, at present, the true prevalence of PDs in later life likely remains unknown.

If the current prevalence rates for PDs among older adults are potentially inaccurate, the theories derived from them must also be placed under special scrutiny. For example, some researchers hypothesize that PDs become less severe with age (Kenan et al., 2000; Paris, 2003). This hypothesis may oversimplify the experience of many or at least some older adults with PDs. Simply put, if the PD criteria contain age-related measurement bias, they may have limited utility for deriving age-related theories. This is true, of course, not just for measuring personality pathology in later life. The same principle applies to measuring personality pathology in childhood. When instruments contain measurement bias across age groups, it becomes very difficult to compare scores across any age groups.

Solving the problem of this measurement bias is not easy. On the one hand, using age-relevant items can create problems with item equivalence across age groups. On the other hand, creating items without any age-associated information is difficult. Fortunately, new item response theory-based statistical techniques can help researchers meet the challenges presented by these longstanding measurement issues. If a researcher opts for a measurement tool that contains several age-specific items that vary across age groups, that researcher can use both the principles of "differential item functioning" (see Hambleton, Swaminathan, & Rogers, 1991) and "linking and equating" (see Dorans & Holland, 2000; Kolen & Brennan, 1995) to understand true relative age differences for the personality feature in question. If a researcher prefers an age neutral measure, that researcher can use similar item response theory techniques to identify and include only those items that function equivalently across age groups. Currently, it is only through these techniques (differential item functioning coupled with linking and equating or creating measures without age- or cohort-associated measurement artifact) that enable us to validly measure and compare personality features across age groups and over time.

### *Summary*

Given its past history as an improving although not perfect system, the future of the DSM appears promising. Support for a dimensional approach is growing, and the search for a new measurement tool is underway. Widiger and Simonsen (2005) summarized 18 alternative ways to measure PDs, with most of these proposals taking a dimensional approach. The implementation of any of these approaches in *DSM-V* would represent a dimensional shift and may solve many of the universal issues of reliability, validity, and utility. There is, as we have described, another problem that looms. The measurement system must account for PD pathology in the contexts of later life.

Currently, a healthy discussion and debate is taking place in the scientific literature to address the shortcomings of the *DSM–IV* PD classification system. These discussions take the view that the current system has limited validity, reliability, and utility because it is rooted in a categorical measurement approach that does not capture the true dimensional nature of PD pathology. Therefore, it is not surprising that proposals to improve the current system mostly center on a shift away from this categorical approach toward a dimensional approach that more accurately reflects the nature of PD pathology. Previously lacking from this discussion and debate, however, has been any serious consideration of the influence of age bias in the current PD criteria. This article suggests that a failure to address the unique challenges associated with the assessment of personality in older adults likely will result in the continued limited validity, reliability, and utility of the DSM

system for this growing population. Our hope is that future versions of the DSM will measure and capture the essence of PD pathology equivalently well across all adult age groups.

A final issue regarding the development of the *DSM-V* deserves brief mention. The *DSM–IV* has been criticized (correctly so in our opinion) for being too "Western-centric" and lacking validity for use in cultures that differ from those in the most developed Western nations (Li, Jenkins, & Sundsmo, 2007; Thakker & Ward, 1998). The category of PDs is particularly vulnerable to limited cross-cultural validity because the extent to which certain personality traits are judged to be pathological is clearly influenced by cultural standards. For example, cultures vary tremendously to the extent to which people are expected (and thus valued) to be individualistic versus communal focused, independent versus interdependent, dominant versus submissive, connected versus disconnected, conforming versus nonconforming, emotionally expressive versus emotional reserved, and active versus passive. The categorical nature of the DSM system conceptualizes PD pathology based on idea prototypes of PDs, or pure types of PDs (Segal & Coolidge, 2001). We would argue that these prototypes in a categorical system may have less cross-cultural validity than a dimensional approach to PD pathology, because in a dimensional system, individuals from diverse cultures would be described in terms of the PD features they actually exhibit regardless of whether the PD features coalesce or converge around the prototypes which would be necessary for a categorical diagnosis. The extent to which aging further influences the cross-cultural relevance of the PD category in the DSM system is an open empirical question, and studies should be conducted to examine the extent to which the age-associated measurement bias described in this article regarding certain PDs applies to other cultures.

## References


Agronin, M. E., & Maletta, G. (2000). Personality disorders in later life: Understanding and overcoming the gap in research. *American Journal of Geriatric Psychiatry, 8,* 4–18.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Antonucci, T. C., & Akiyama, H. (1987). Social networks in adult life and a preliminary examination of the convoy model. *Journal of Gerontology, 42,* 519–527.

Balsis, S., & Cooper, L. D. (2009). *Measuring personality disorders in later life: Hybrid criteria.* Manuscript submitted for publication.

Balsis, S., Gleason, M. E. J., Woods, C. M., & Oltmanns, T. F. (2007). An item response theory analysis of *DSM–IV* personality disorder criteria across younger and older age groups. *Psychology and Aging, 22,* 171–185.

Balsis, S., Woods, C. M., Gleason, M. E. J., & Oltmanns, T. F. (2007). Overdiagnosis and underdiagnosis of personality disorders in older adults. *American Journal of Geriatric Psychiatry, 15,* 742–753.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25,* 805–819.

Clark, L. A., McEwen, J. L., Collard, L. M., & Hickok, L. G. (1993). Symptoms and traits of personality disorder: Two new methods for their assessment. *Psychological Assessment, 5,* 81–91.

Cloninger, C. R., Przybeck, T. R., Svrakic, D. D., & Wetzel, R. (1994). *The Temperament and Character Inventory (TCI): A guide to its development and use.* St. Louis, MO: Washington University School of Medicine.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281–306.

Drake, R. I., & Vaillant, G. E. (1988). Longitudinal views of personality disorder. *Journal of Personality Disorders, 2,* 44–48.

Durrett, C., & Westen, D. (2005). The structure of axis II disorders in adolescents: A cluster- and factor-analytic investigation of *DSM–IV* categories and criteria. *Journal of Personality Disorders, 19,* 440–461.

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). *Structured Clinical Interview for DSM–IV Axis II Personality Disorders (SCID-II*). Washington, DC: American Psychiatric Press.

Gish, R. G., Lee, A., Brooks, L., Leung, J., Lau, J. Y., & Moore, D. H. (2001). Long-term follow-up of clients diagnosed with alcohol dependence or alcohol abuse who were evaluated for liver transplantation. *Liver Transplantation, 7,* 581–587.

Grilo, C. M., McGlashan, T. H., Morey, L. C., Gunderson, J. G., Skodol, A. E., & Shea, M. T. (2001). Internal consistency, intercriterion overlap and diagnostic efficiency of criteria sets for *DSM–IV* schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Acta Psychiatrica Scandinavica, 104,* 264–272.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hankin, B. L., Fraley, R. C., Lahey, B. B., & Waldman, I. D. (2005). Is depression best viewed as a continuum or discrete category? A taxometric analysis of childhood and adolescent depression in a population-based sample. *Journal of Abnormal Psychology, 114,* 96–110.

Hargrave, T. D., & Anderson, W. T. (1992). *Finishing well: Aging and reparation in the intergenerational family.* Philadelphia: Brunner/Mazel.

Heumann, K. A., & Morey, L. C. (1990). Reliability of categorical and dimensional judgments of personality disorder. *American Journal of Psychiatry, 147,* 498–500.

Jane, J. S., Pagan, J. L., Turkheimer, E., Fiedler, E. R., & Oltmanns, T. F. (2006). The interrater reliability of the Structured Interview for *DSM–IV* Personality. *Comprehensive Psychiatry, 47,* 368–375.

Kagan, J. (1969). The three faces of continuity in human development. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 53–65). Chicago: Rand McNally.

Kenan, M. M., Kendjelic, E. M., Molinari, V. A., Williams, W., Norris, M., & Kunik, M. E. (2000). Age-related differences in the frequency of personality disorders among inpatient veterans. *International Journal of Geriatric Psychiatry, 15,* 831–837.

Kernberg, O. F. (1996). A psychoanalytic theory of personality disorders. In J. F. Clarkin, & M. F. Lenzenweger (Eds.), *Major theories of personality disorder* (pp. 106–140). New York: Guilford Press.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Krueger, R. F. (1999). The structures of common mental disorders. *Archives of General Psychiatry, 56,* 921–926.

Levy, K. N., Becker, D. F., Grilo, C. M., Mattanah, J. J. F., Garnet, K. E., Quinlan, D. M., et al. (1999). Concurrent and predictive validity of the personality disorder diagnosis in adolescent inpatients. *American Journal of Psychiatry, 156,* 1522–1528.

Li, S. T., Jenkins, S., & Sundsmo, A. (2007). Impact of race and ethnicity. In M. Hersen, S. M. Turner, & D. C. Beidel (Eds.), *Adult psychopathology and diagnosis* (5th ed.; pp. 101–121). New York: Wiley.

Livesley, W. J. (2005). Behavioral and molecular genetic contributions to a dimensional classification of personality disorder. *Journal of Personality Disorders, 19,* 131–155.

Millon, T., Davis, R., & Millon, C. (1997). *MCMI-III manual* (2nd ed.). Minneapolis, MN: National Computer Systems.

Moffitt, T. E., Caspi, A., Harrington, H., & Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology, 14,* 179–207.

Morey, L. C., & Zanarini, M. C. (2000). Borderline personality: Traits and disorder. *Journal of Abnormal Psychology, 109,* 733–737.

Mroczek, D. K., Hurt, S. W., & Berman, W. H. (1999). Conceptual and methodological issues in the assessment of personality disorders in older adults. In E. Rosowsky, R. C. Abrams, & R. A. Zweig (Eds.), *Personality disorders in older adults: Emerging issues in diagnosis and treatment* (pp. 135–150). Mahwah, NJ: Erlbaum Publishers.

Narrow, W. E., Rae, D. S., Robins, L. N., & Regier, D. A. (2002). Revised prevalence based estimates of mental disorders in the United States: Using a clinical significance criterion to reconcile 2 surveys' estimates. *Archives of General Psychiatry, 59,* 115–123.

Paris, J. (2003). Personality disorders over time: Precursors, course and outcome. *Journal of Personality Disorders, 17,* 479–488.

Robins, L. N. (1966). *Deviant children grown up: A sociological and psychiatric study of sociopathic personality.* Oxford, England: Williams & Wilkins.

Rosowsky, E., Abrams, R. C., & Zweig, R. A. (Eds.). (1999). *Personality disorders in older adults: Emerging issues in diagnosis and treatment.* Mahwah, NJ: Erlbaum Publishers.

Rosowsky, E., & Smyer, M. A. (1999). Personality disorders and the difficult nursing home resident. In E. Rosowsky, R. C. Abrams, & R. A. Zweig (Eds.), *Personality disorders in older adults: Emerging issues in diagnosis and treatment* (pp. 257–274). Mahwah, NJ: Erlbaum Publishers.

Ruscio, A. M. (2008). Important questions remain to be addressed before adopting a dimensional classification of mental disorders. *American Psychologist, 63,* 61–62.

Sadavoy, J. (1987). Character pathology in the elderly. *Journal of Geriatric Psychiatry, 20,* 165–178.

Sadavoy, J. (1996). Personality disorder in old age: Symptom expression. *Clinical Gerontologist, 16,* 19–36.

Segal, D. L., & Coolidge, F. L. (2001). Diagnosis and classification. In M. Hersen & V. B. Van Hasselt (Eds.). *Advanced abnormal psychology* (2nd ed., pp. 5–22). New York: Kluwer Academic/Plenum.

Segal, D. L., & Coolidge, F. L. (2006a). Validity. In N. J. Salkind (Ed.), *Encyclopedia of human development.* Thousand Oaks, CA: Sage.

Segal, D. L., & Coolidge, F. L. (2006b). Reliability. In N. J. Salkind (Ed.), *Encyclopedia of human development.* Thousand Oaks, CA: Sage.

Segal, D. L., Coolidge, F. L., & Rosowsky, E. (2006). *Personality disorders and older adults: Diagnosis, assessment, and treatment.* Hoboken, NJ: Wiley.

Segal, D. L., Hersen, M., Van Hasselt, V. B., Silberman, C. S., & Roth, L. (1996). Diagnosis and assessment of personality disorders in older adults: A critical review. *Journal of Personality Disorders, 10,* 384–399.

Siegel, D. J., & Small, G. W. (1986). Borderline personality disorder in the elderly: A case study. *Canadian Journal of Psychiatry, 31,* 859–860.

Spiessl, H., Hubner-Liebermann, B., Binder, H., & Cording, C. (2002). Heavy users in a psychiatric hospital: A cohort study on 1811 clients over five years. *Psychiatrische Praxis, 29,* 350–354.

Thakker, J., & Ward, T. (1998). Culture and classification: The cross-cultural application of the. *DSM–IV. Clinical Psychology Review, 18,* 501–529.

Thomas, C., Turkheimer, E., & Oltmanns, T. F. (2003). Factorial structure of pathological personality as evaluated by peers. *Journal of Abnormal Psychology, 112,* 81–91.

Trull, T. J. (2005). Dimensional models of personality disorder: Coverage and cutoffs. *Journal of Personality Disorders, 19,* 262–282.

Verheul, R. (2005). Clinical utility of dimensional models for personality pathology. *Journal of Personality Disorders, 19,* 283–302.

Verheul, R., & Widiger, T. A. (2004). A meta-analysis of the prevalence and usage of the personality disorder not otherwise specified (PDNOS) diagnosis. *Journal of Personality Disorders, 18,* 309–319.

Weertman, A., Arntz, A., Dreessen, L., van Velzen, C., & Vertommen, S. (2003). Short-interval test-retest 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.

Whiteman, M. C., Deary, I. J., & Fowkes, F. R. (2000). Personality and health: Cardiovascular disease. In S. E. Hampson (Ed.). *Advances in personality psychology* (pp. 157–198). New York, Psychology Press.

Widiger, T. A., & Clark, L. A. (2000). Toward DSM-V and the classification of psychopathology. *Psychological Bulletin, 126,* 946–963.

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions: A question for DSM-V. *Journal of Abnormal Psychology, 114,* 494–504.

Widiger, T. A., & Simonsen, E. (2005). Alternative dimensional models of personality disorder: Finding a common ground. *Journal of Personality Disorders, 19,* 110–130.

Widiger, T. A., & Trull, T. J. (2007). Plate tectonics in the classification of personality disorder: Shifting to a dimensional model. *American Psychologist, 62,* 71–83.

Received September 16, 2008
Revision received May 12, 2009
Accepted May 18, 2009 ■