Journal of Applied Psychology
2014, Vol. 99, No. 4, 564–586

© 2013 American Psychological Association
0021-9010/14/$12.00 DOI: 10.1037/a0034688

Uncovering Curvilinear Relationships Between Conscientiousness and Job Performance: How Theoretically Appropriate Measurement Makes an Empirical Difference

Nathan T. Carter
University of Georgia

Dev K. Dalal
University of Connecticut

Anthony S. Boyce
Aon Hewitt Consulting, New York, New York

Matthew S. O'Connell, Mei-Chuan Kung, and Kristin M. Delgado
Select International, Pittsburgh, Pennsylvania

The personality trait of conscientiousness has seen considerable attention from applied psychologists due to its efficacy for predicting job performance across performance dimensions and occupations. However, recent theoretical and empirical developments have questioned the assumption that more conscientiousness always results in better job performance, suggesting a curvilinear link between the 2. Despite these developments, the results of studies directly testing the idea have been mixed. Here, we propose this link has been obscured by another pervasive assumption known as the dominance model of measurement: that higher scores on traditional personality measures always indicate higher levels of conscientiousness. Recent research suggests dominance models show inferior fit to personality test scores as compared to ideal point models that allow for curvilinear relationships between traits and scores. Using data from 2 different samples of job incumbents, we show the rank-order changes that result from using an ideal point model expose a curvilinear link between conscientiousness and job performance 100% of the time, whereas results using dominance models show mixed results, similar to the current state of the literature. Finally, with an independent cross-validation sample, we show that selection based on predicted performance using ideal point scores results in more favorable objective hiring outcomes. Implications for practice and future research are discussed.

Keywords: personality, job performance, unfolding, ideal point, item response theory

The personality trait of conscientiousness has seen considerable attention from applied psychologists interested in understanding and predicting various dimensions of job performance. Conscientious individuals tend to plan ahead, be organized, show high self-control, follow rules, and be less impulsive (B. W. Roberts, Jackson, Fayard, Edmonds, & Meints, 2009). Therefore, it is not surprising meta-analyses have shown conscientiousness is the most efficacious personality predictor of performance across criterion dimensions and occupations (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001). But is more conscientiousness always better? That is, is there a point where too much concern for planning, organization, and rules can hinder performance?

Indeed, recent theoretical work in organizational (Pierce & Aguinis, 2013) and psychological (Grant & Schwartz, 2011) science suggests that too much of a seemingly desirable trait can be suboptimal. Personality theory suggests that those with moderate to high conscientiousness are generally adaptive and therefore productive. Excessive conscientiousness, however, can imply less positive behavioral outcomes. In fact, behaviors such as stalled task completion, overthinking, and preoccupation with order and detail can be expected when conscientiousness is excessively high (see American Psychiatric Association, 2000; Widiger, Trull, Clarkin, Sanderson, & Costa, 2002), implying the potential for a curvilinear relationship between conscientiousness and performance. In spite of these theoretical developments, empirical support for this curvilinear link has been mixed, with some studies finding and others failing to find the effect (e.g., Le et al., 2011).

The inconsistencies in these studies are problematic for applied psychologists. Personality testing is gaining in popularity and ease, with companies spending $3.8 billion in 2011 on talent management software that often uses algorithms involving personality test scoring (Walker, 2012), and large-scale educational testing firms, such as the Educational Testing Service, exploring options for using personality indicators for graduate school admissions tests (de Vise, 2009). Given the increasing popularity and use of personality tests in high-stakes environments, it is critical that we gain a full understanding of the internal and external functioning of personality measures.

This article was published Online First November 4, 2013.

Nathan T. Carter, Department of Psychology, University of Georgia; Dev K. Dalal, Department of Psychology, University of Connecticut; Anthony S. Boyce, Aon Hewitt Consulting, New York, New York; Matthew S. O'Connell, Mei-Chuan Kung, and Kristin M. Delgado, Select International, Pittsburgh, Pennsylvania.

Correspondence concerning this article should be addressed to Nathan T. Carter, Department of Psychology, 323 Psychology Building, University of Georgia, Athens, GA 30602. E-mail: carternt1981@gmail.com

In this article, we suggest curvilinear personality–performance trends may have been obfuscated in past studies by the application of the pervasive *dominance model* of measurement (Likert, 1932) implicit in classical test theory (CTT) statistics, conventional item response theory (IRT), and factor analytic (FA) models. The dominance perspective makes the assumption that higher CTT total scores (i.e., sum or average-item scores) on personality inventories always indicate a higher level of the measured trait. On the other hand, recent evidence has suggested that *ideal point* measurement models (Coombs, 1964; Thurstone, 1928) are a more theoretically and empirically appropriate method for scaling personality variables (Stark, Chernyshenko, Drasgow, & Williams, 2006). In contrast to the dominance perspective, ideal point models imply a curvilinear relationship between personality test scores and the measured personality trait, such that two similar observed scores could indicate quite different standings on the trait being measured.

We begin by presenting the theoretical reasons why conscientiousness should be curvilinearly related to performance and summarizing past research investigating the phenomenon. Next, we outline the theoretical and empirical rationale for why an ideal point model is most appropriate for responses to personality items. Drawing on past theoretical work, we show that conceptual differences in dominance and ideal point models suggest a rank-ordering of respondents that is more consistent with the idea that conscientiousness can, in fact, be “too much of a good thing” (Pierce & Aguinis, 2013). We then present the results of three studies exploring the interplay between the elusive curvilinear personality–performance relationship and the use of ideal point response modeling for the estimation of personality traits. In particular, we investigate whether using a curvilinear (i.e., ideal point) measurement model results in more consistently finding curvilinear relationships between personality and performance. Finally, we explore whether more accurate employee-selection decisions are made using this approach.

The Conscientiousness–Performance Relationship: Linear or Curvilinear?

The personality–performance relationship has been debated for decades, and relatively low criterion-related correlations have been a source of frustration for applied psychological researchers and practitioners (see Morgeson et al., 2007; Tett & Christiansen, 2007). It has been suggested, however, that applied psychology may have concluded too hastily that these relationships must be linear (e.g., Ones, Dilchert, Viswesveran, & Judge, 2007). Recently, researchers have begun to explore this functional relationship, finding mixed results, most frequently with regard to conscientiousness.

With the goal of providing a more theoretically enriched discussion, we chose to focus on conscientiousness here due to its established reputation as the five-factor model trait most predictive of job performance (see Barrick et al., 2001; Schmidt, Shaffer, & Oh, 2008), recent speculation about the functional form of its relation with performance (Le et al., 2011; Pierce & Aguinis, 2013), and the fact that its maladaptive extremes have been relatively well explored (e.g., B. W. Roberts et al., 2009). Of importance, we believe our discussion is applicable to other personality traits that researchers have also suggested have downsides at extreme levels (e.g., extraversion; Grant & Schwartz, 2011; Judge & LePine, 2007). Our discussion also applies to other self-report measures of personality, which have generally shown to be better fit by ideal point models (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006).

From the perspective of personality theory, moderate to high conscientiousness is adaptive and often beneficial (see B. W. Roberts et al., 2009). However, excessively conscientious individuals have been noted to deliberate for too long; be perfectionistic to the point that work is not completed; and be overly focused on organization, orderliness, rules, and details (Samuel & Widiger, 2011, p. 162). Moreover, highly conscientious individuals often are higher in self-critical perfectionism (Dunkley, Blankstein, Zuroff, Lecce, & Hui, 2006; Hill, McIntire, & Bacharach, 1997), exhibit behaviors associated with obsessive compulsion (B. W. Roberts et al., 2009), and have more adverse performance and stress reactions to negative feedback (Cianci, Klein, & Seijts, 2010) than those lower in conscientiousness.

Lower performance ratings for these extremely high-conscientiousness people might result, for example, because their excessive conscientiousness will not allow them to submit completed projects unless they felt the work was completely perfect, and thus task performance (TP) might suffer. Preoccupation with orderliness and details of tasks might inhibit prosocial and helping behavior at work, thus limiting organizational citizenship behavior (OCB). Highly adverse reactions to negative events at work might result in higher counterproductive work behavior (CWB). Therefore, it stands to reason that there might be a “sweet spot” of conscientiousness. For example, one needs to be conscientious enough to catch errors in one’s own work, keep tasks and deadlines well organized, and follow relevant rules and guidelines. However, too much conscientiousness can lead to a paralysis of sorts wherein the worker is overly concerned with minor errors, maintaining organization, and rigidly following guidelines that may be irrelevant in certain contexts—resulting in less positive work outcomes compared to those with moderately high, adaptive levels of conscientiousness.

It is important to note that high levels of conscientiousness are not synonymous with other maladaptive personality traits like neuroticism. In fact, these two traits have been shown to be quite distinct. First, although extremely high levels of conscientiousness are positively associated with the anxiety (Samuel, Lynam, Widiger, & Ball, 2012) and self-conscious emotions (Fayard, Roberts, Robins, & Watson, 2012) aspects of neuroticism, extreme conscientiousness is also negatively related to the impulsivity facet and unrelated to other facets (see Samuel et al., 2012). That is, although some similar outcomes might be expected for extreme conscientiousness and high neuroticism, the mechanisms that produce these outcomes are quite different. For example, it has been shown that the positive association between conscientiousness and feelings of guilt are not due to the overlap between conscientiousness and neuroticism (Fayard et al., 2012).

Second, extreme high-conscientious individuals would be expected to have higher performance than persons low in conscientiousness or high in neuroticism. Although there is some overlap of conscientiousness with regard to anxiety and self-consciousness, the neuroticism facets that are orthogonal to, or negatively related to, conscientiousness (i.e., angry hostility, vulnerability, impulsivity), are likely to have more severe negative work outcomes than would be expected of extreme conscientiousness (e.g., delays in task completion due to high perfectionism). Thus, those with extreme conscientiousness would be expected to have higher performance than those with high levels of neuroticism. This is in line with the curvilinear effects found in Study 1 by Le et al. (2011) in that task performance levels of excessively high-conscientious persons were still about one standard deviation higher than persons high in neuroticism (i.e., low in emotional stability).[1] Further, Bowling, Burns, Stewart, and Gruys (2011) showed that high conscientiousness acts as an attenuating factor for the relationship between neuroticism and CWB. In multiple samples they showed that only those low in conscientiousness and high in neuroticism showed elevated levels of CWB, concluding that high conscientiousness restricts expressions of neurotic behavior. Thus, the goal of understanding excessive conscientiousness and its relation to performance is to make distinctions between moderate-performing, extremely high-conscientiousness individuals and high-performing, moderately conscientiousness individuals, as opposed to separating low-performing (i.e., those low in conscientiousness or high in neuroticism) from high-performing individuals.

In spite of the potential downsides to extreme conscientiousness, only a limited amount of published research has considered curvilinear relationships between personality traits and job performance. As noted above, the results have been equivocal. Table 1 summarizes seven published articles that explored the curvilinear link between conscientiousness and various performance outcomes, with a total of 34 regression analyses testing such relationships. For each regression analysis we provide information concerning the sample studied, the predictor and criteria, and whether linear and curvilinear trends were significant. Additionally, where possible, we include the R^2 for the linear effect and change in R^2 for the curvilinear effect over the linear effect.

Day and Silverman (1989) found a measure of impulse expression to show curvilinear relationships with outcome measures of timeliness of work and cooperation, but five other outcomes investigated showed no curvilinear trend. Robie and Ryan (1999) tested for curvilinear relationships between measures of conscientiousness and supervisor ratings of job performance in four concurrent and one predictive validity studies; evidence of the trend was found in only one of these samples. LaHuis, Martin, and Avis (2005) found a significant quadratic relationship for a measure of conscientiousness with a one-item rating of job performance in a sample of clerical workers. In their second study using a similar sample and controlling for cognitive ability, these researchers again found a significant curvilinear relationship between conscientiousness and job performance.

Cucina and Vasilopoulos (2005) examined the relationship among all Big Five personality traits and academic performance (i.e., GPA) and found curvilinear relationships only for measures of openness and conscientiousness. Additionally, Vasilopoulos, Cucina, and Hunter (2007) found that measures of conscientiousness showed curvilinear relationships with performance on final exams in two training courses.

Across two studies, Whetzel, McDaniel, Yost, and Kim (2010) found that the conscientiousness dimension of the 32 Occupation Personality Questionnaire (OPQ) scales showed meaningful curvilinearity when using their most liberal "significance" rule (i.e., $\Delta R > .01$) for predicting supervisor performance ratings of TP. However, other scales identified by the current authors as facets of conscientiousness showed only one significant curvilinear relationship in one study (i.e., Conventional; see Table 1). Their results suggest that only a little over half (18.5 across the two studies) of the 32 scales showed $\Delta R > .01$ for the curvilinear effect (see Whetzel et al.'s Table 1).

Most recently, Le et al. (2011) showed mixed results regarding the curvilinear relationship between measures of conscientiousness and emotional stability and job performance dimensions of TP, OCB, and CWB. In the first study, significant quadratic effects were found for all three outcomes. However, in Study 2, no quadratic effects were significant for their measure of conscientiousness predicting the same outcomes, and the quadratic effect of their measure of emotional stability was significant only for predicting OCB; it was also significant for CWB after controlling for job complexity.

Table 1 shows that only 15 (44.1%) of 34 regressions showed significant curvilinearity; the same number showed only a significant linear effect (with a nonsignificant curvilinear effect). Despite several high-quality studies with large samples, different measures of personality, and various performance measures, results regarding the curvilinear personality–performance relationship are mixed, both within and between studies. We noted previously that this is highly problematic for research aiming to understand the scientific link between personality and work behavior, and perhaps even more frustrating to practitioners whose expensive and high-stakes employee selection programs are affected by such uncertainty. Of importance, each of the aforementioned studies scored their personality measures under the assumptions of dominance responding. If, however, a different personality scoring approach yielded more consistent findings, researchers and practitioners could proceed more confidently. Below, we discuss recent developments in the literature suggesting measurement models that assume the CTT total score is a viable proxy for personality traits may not be as appropriate as ideal point models for scaling personality measures.

Modeling Conscientiousness Scores: Dominance or Ideal Point?

Dominance and ideal point response processes make fundamentally different assumptions about response behavior to psychological scales. Dominance models assume the more of the attribute (e.g., personality, attitude) a respondent has, the higher the respondent's endorsement rating will be (e.g., *Strongly Agree* vs. *Agree*; see Figure 1). On the other hand, ideal point models do not assume monotonically increasing relationships. Instead, individuals are more likely to endorse items that are located near their standing on the latent attribute continuum. If an item is too extreme or not extreme enough to describe an individual, the individual is less

[1] Using the Le et al. (2011) regression results for low-complexity jobs (where curvilinearity was strongest; see their Table 3, p. 121) to calculate predicted values reveals that those low (−3 *SD*) in conscientiousness showed similar task performance, $\hat{Y} = 19.93$, to those with low neuroticism, $\hat{Y} = 19.33$; those with excessively high conscientiousness (+3 *SD*) showed task performance, $\hat{Y} = 21.31$, one full standard deviation ($SD = 3.99$) higher than those with high neuroticism, $\hat{Y} = 17.35$, and 0.49 *SD* higher than those low in conscientiousness and those low in neuroticism.

Table 1
Summary of Results of Past Studies Investigating Curvilinearity in the Conscientiousness–Performance Relationship

Source	Sample no.	Sample	Predictor(s)	Criterion	L (R^2)	C (ΔR^2)
1. Day & Silverman (1989)	1 of 1	Accountants (N = 43)	Impulse expression[a]	Potential for success		
				Technical ability		
				Timeliness of work		X (n/a)
				Client relations		
				Cooperation		X (n/a)
				Work ethic		
				Global performance		
2. Robie & Ryan (1999)	1 of 5	Various jobs, federal government (N = 999)	NEO-PI–R Conscientiousness	Overall performance	X (.010)	
	2 of 5	Multi-organization private sector (N = 200)	NEO-PI–R Conscientiousness	Overall performance	X (.060)	
	3 of 5	Department of Defense managers (N = 146)	PCI Conscientiousness	Overall performance	X (.060)	
	4 of 5	Wholesale sales representatives (N = 206)	PCI Conscientiousness	Overall performance	X (.060)	
	5 of 5	Long-haul semitruck drivers (N = 256)	PCI Conscientiousness	Overall performance	X (.070)	
3. LaHuis et al. (2005)	1 of 2	Clerical–federal government (N = 192)	Conscientiousness[a]	Overall performance		X (.020)
	2 of 2	Clerical–state government (N = 203)	NEO-PI–R Conscientiousness	Overall performance	X (n/a)	X (.020)
4. Cucina & Vasilopoulos (2005)	1 of 1	Undergraduate psychology students (N = 262)	IPIP Conscientiousness	Grade point average	X (.033)	X (.022)
5. Vasilopoulos et al. (2007)	1 of 1	Federal law enforcement trainees (N = 1,010)	Conscientiousness[a]	Training Exam 1	X (.011)	X (.012)
				Training Exam 2	X (.004)	X (.009)
				Exam composite	X (.010)	X (.013)
			Achievement motive[a]	Training Exam 1	X (n/a)	
				Training Exam 2		
				Exam composite	X (n/a)	
			Dependability[a]	Training Exam 1	X (n/a)	X (.007)
				Training Exam 2	X (n/a)	X (.008)
				Exam composite	X (n/a)	X (.010)
6. Whetzel et al. (2010)[b]	1 of 1	Financial service professionals (N = 576)[b]	OPI/OPQ Conscientiousness	Overall performance	X (<.001)	X (.011)
			OPI/OPQ Achieving	Overall performance	X (.040)	
			OPI/OPQ Detail Conscious	Overall performance	X (.021)	
			OPI/OPQ Conventional	Overall performance	X (.029)	
	1 of 2	Financial service professionals (N = 576)[b]	OPI/OPQ Conscientiousness	Overall performance	X (<.001)	X (.005)
			OPI/OPQ Achieving	Overall performance	X (.058)	
			OPI/OPQ Detail Conscious	Overall performance	X (.023)	
			OPI/OPQ Conventional	Overall performance	X (.018)	X (.005)
7. Le et al. (2011)	1 of 2	Various jobs, single organization (N = 602)	Conscientiousness[a]	Counterproductive work behavior	X (n/a)	X (.021)
				Organizational citizenship behavior	X (n/a)	X (.010)
				Task performance	X (n/a)	X (.014)
	2 of 2	Various jobs, multiple organizations (N = 956)	Conscientiousness[a]	Counterproductive work behavior	X (n/a)	
				Organizational citizenship behavior	X (n/a)	
				Task performance	X (n/a)	

Note. The columns labeled "L" and "C" represent significance of the linear and curvilinear (i.e., quadratic) relationship, respectively. An X in a cell indicates a significant linear effect if in the "L" column and indicates a significant or a significant curvilinear effect if in the "C" column. With the exception of training exam scores and grade point average, all performance measures were based on supervisor ratings. n/a indicates the statistic was not reported and could not be calculated given the information provided in the article. All studies used different measures of job performance. NEO-PI–R= NEO Personality Inventory—Revised; PCI = Psychological Contract Inventory; IPIP = International Personality Item Pool; OPI/OPQ = Occupational Personality Inventory/Occupational Personality Questionnaire.

[a] Indicates that the measure was created by the researchers and is not a standard scale (e.g., the NEO-PI–R). [b] Whetzel et al. (2010) used rules involving change in variance explained to evaluate significance. Here, we present results as significant if they met either of their rules ($\Delta R > .01$ or .025).

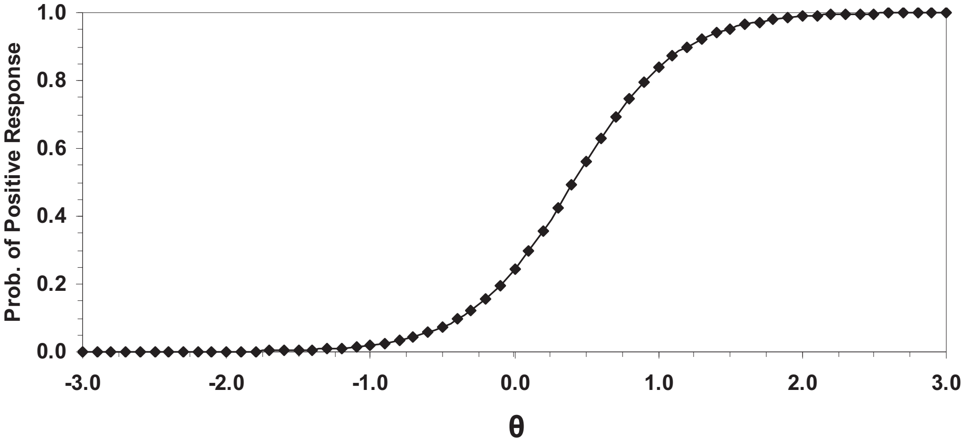

Figure 1. Example of a dominance response process model. Prob. = probability.

likely to fully endorse (i.e., strongly agree to) that item (see Figure 2).

These differences between dominance and ideal point models imply different scoring assumptions. Under dominance assumptions a total score created by summing or averaging across scale points endorsed for items (reverse scoring where appropriate) can be considered an accurate proxy for the latent variable. Conversely, scoring under ideal point assumptions requires considering observed responses relative to items' locations along the attribute continuum; that is, a simple total (i.e., sum or average) score is not a sufficient proxy for the attribute. Dominance approaches have received the bulk of attention from researchers primarily due to the ease of creating and scoring dominance scales. Indeed, the dominance approach to scaling facilitates the use of item–total correlations, factor analyses, and internal consistency reliability estimates (see Stark et al., 2006) as indicators of a scale's reliability and internal validity; these are tools with which scale developers have become quite comfortable (Zickar & Broadfoot, 2009).

Along with classical test theory and most factor analytic models, conventional item response theory (IRT) models such as the two-parameter logistic (2PL; Birnbaum, 1968) and graded response model (GRM; Samejima, 1969) carry dominance assumptions. That is, these models assume that as the latent trait (θ) increases, the probability of responding positively increases (see Figure 1). As noted above, the ideal point process implies that the probability of responding positively increases as the location of the item (δ) and the person (θ) is minimized (i.e., the peak of Figure 2). In other words, as (θ − δ) approaches zero, the individual is likely to endorse higher scale points; as θ − δ deviates from zero (in either direction) the individual is likely to endorse lower scale points. This implies that one can agree or disagree with an item "from above" or "from below" an item's location (J. S. Roberts, Laughlin, & Wedell, 1999). For example, people can disagree with the item "I have a daily planner, but struggle to keep it up to date" because they always keep their planner up to date (from above) or because they either do not have a planner, or never keep the one they have up to date (from below). In either case, the same response (i.e., disagree) is observed, but two very different levels of the trait are indicated. J. S. Roberts, Donoghue, and Laughlin (2000) developed the generalized graded unfolding IRT model (GGUM) as a manifestation of the ideal point response process for polytomous, graded survey items.

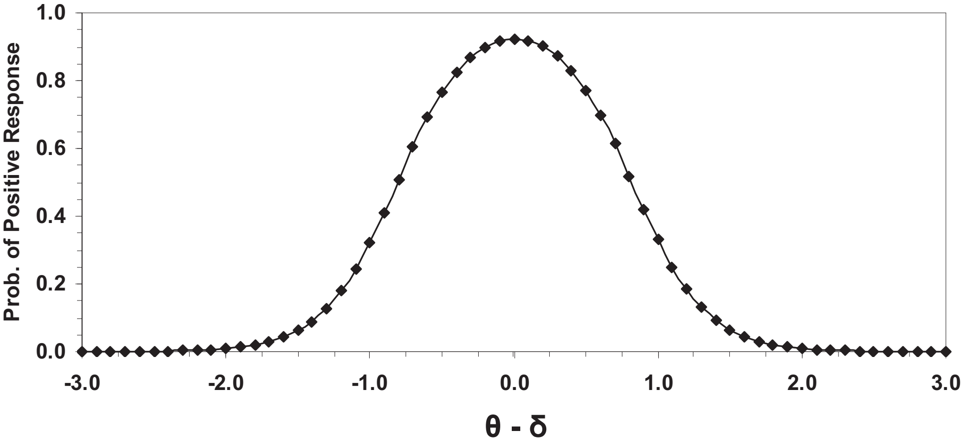

Figure 2. Example of an ideal point response process model. Prob. = probability.

Since J. S. Roberts et al. (2000) introduced the GGUM, applied researchers have begun to ask if the ideal point approach might be more appropriate for scaling noncognitive constructs (e.g., personality, attitudes, interests). One of the first major applications of the GGUM to personality data was by Stark et al. (2006). These researchers directly compared the fit of dominance and ideal point IRT models to data from the 16PF. They demonstrated that ideal point IRT (e.g., GGUM) models generally fit responses to the 16PF better than dominance models (e.g., 2-PL) and showed higher reliability across the trait continuum. Moreover, they showed the possibility of considerable differences in the top-down rank-order of individuals between dominance and ideal-point IRT scores.

Weekers and Meijer (2008) demonstrated that ideal point models showed better fit than dominance models when applied to responses from a Dutch personality inventory as well as responses from a Dutch translation of Chernyshenko et al.'s (2007) Orderliness scale. In addition, Weekers and Meijer found the correlation between dominance model-estimated person standing on the latent trait continuum (i.e., θ estimates) and GGUM based θ estimates to be .98 and .99 for the two scales, respectively. Inspection of their scatterplots (Weekers & Meijer, 2008, p. 75), however, would suggest differences in the top-down rank-order of individuals similar to Stark et al. (2006). Finally, Zampetakis (2010) fit the GGUM to responses to a creative personality inventory showing superior fit for the ideal point model compared to a dominance model. Like other researchers (e.g., Carter & Dalal, 2010), Zampetakis showed a correlation of only .81 between dominance and ideal point scoring.

In addition to fitting personality data, ideal point models have been shown to fit interest and attitude data better than dominance models. For instance, Tay, Drasgow, Rounds, and Williams (2009) showed that ideal point models fit responses to interest data better than dominance models. Furthermore, Tay et al. demonstrated that assuming the wrong response process (i.e., IRT model) can result in inappropriate estimates of respondents' interest levels (i.e., using the wrong IRT model can result in misestimating standing on the attribute continuum). Carter and Dalal (2010) showed that the GGUM fit responses to the Work Scale of the Job Descriptive Inventory better than the GRM or nominal response model (NRM; both dominance models). Like Tay et al., Carter and Dalal demonstrated that the CTT total score was less empirically and theoretically appropriate for indexing the attitude and correlated only .85 with the GGUM.

From this review, it is clear that the ideal point models fit data from these noncognitive constructs better than more traditional dominance models. Drasgow, Chernyshenko, and Stark (2010) argued that ideal point models show better fit to personality responses because such items involve a process wherein the respondent compares the extremity of an item to their own extremity on the personality trait when deciding whether or not to affirm the item; a process they refer to as *introspection.* As this review shows, this process holds even for scales that were developed under dominance assumptions.

Interestingly, under ideal point assumptions, persons with the highest CTT total score are not necessarily the highest in the trait. This follows from the idea that persons are most likely to endorse or agree with an item if the statement reflects their true standing on the measured trait. One implication of this model is that those with extreme levels of a trait would be less likely, rather than more likely (as implied by the dominance model) to endorse positively keyed items because the content does not reflect their own extremity (J. S. Roberts et al., 1999). For example, those with excessive levels of conscientiousness may not fully endorse an item like "I like to follow the rules;" this is true not because they do not follow the rules but because they always follow rules whether they like it or not, signaling the rigidity that is associated with very high conscientiousness (Judge & LePine, 2007). Therefore, an extremely conscientious individual would not affirm this item because he or she is higher on conscientiousness than this item implies (i.e., the item is not extreme enough).

An important consequence of ideal point models fitting personality response data better than dominance models is that dominance-based scores will result in incorrect rank-order inferences regarding individuals with extreme levels of the attribute. In particular, dominance scoring will incorrectly index those individuals as relatively *moderate* on the trait compared to their extreme true standing (J. S. Roberts et al., 1999). Stated differently, individuals who should be scored as extremely conscientious would (incorrectly) be scored as moderately high in conscientious with dominance scoring. Thus, the rank-ordering of respondents would differ between scaling approaches. Moreover, due to the fairly extreme-worded items typical of Likert-type personality items, rank-ordering changes would occur, as the studies reviewed above suggest, for a small but very important group of respondents in employee selection applications: those at the top of the rank-order.

Because of this misestimation of the trait, we propose the curvilinear form of the conscientiousness–performance relationship is misrepresented as linear. It was argued earlier that extreme conscientiousness should be associated with lower performance scores than moderately high conscientiousness. However, when dominance scoring is incorrectly applied, individuals with more extreme conscientiousness would be incorrectly indexed as moderately conscientious, whereas individuals who are moderately conscientious are correctly indexed as moderately conscientious. That is, dominance scoring confuses these individuals as being similar in conscientiousness. On the other hand, ideal point scoring differentiates between them, resulting in a more appropriate rank-order (and therein better model–data fit).

One result of the use of dominance scoring is that when regressing performance onto conscientiousness estimates, lower performance scores that should be associated with excessively conscientious individuals are instead associated with the (incorrect) moderate trait estimate, pulling predictions of performance downward for those correctly scored as moderate. As a result, what should be a curvilinear trend appears to be a simple linear trend because, as (incorrect) conscientiousness scores increase, performance scores appear to increase. The inflection point that should theoretically be seen when too much conscientiousness impairs performance is buried in the middle of the dominance-based attribute estimates, producing a seemingly linear trend. The inconsistent findings of a curvilinear trend in past research can likely be attributed to whether or not the dominance scoring used resulted in all or just some of the relatively extreme individuals being indexed as more moderate.

Although ideal point models show measurement advantages, limited attention has been given to the criterion-related validity implications of using ideal point scoring (Dalal, Withrow, Gibby,

& Zickar, 2010), in spite of early speculation that validity may change (e.g., Stark et al., 2006). The only study of which we are aware that investigated this issue (Chernyshenko et al., 2007) directly compared criterion-related validity of ideal point and dominance-based scales. These researchers created three scales of orderliness (a facet of conscientiousness) using CTT, dominance IRT, and ideal point IRT scale development techniques. No appreciable differences in criterion-related validity estimates were found among the three versions of the scale on several outcome measures. It is important to note, however, that these approaches resulted in different items within each scale wherein the scoring was appropriate for the scale constructed (i.e., the average item score was appropriate for the dominance scale). Further, only linear assessments of criterion-related validity were considered. What has not been investigated, however, are the implications of incorrect scoring of personality scales for the functional form of the personality–performance relationship.

The Current Studies

Two things are apparent from the reviews above. First, findings regarding the personality–performance link are currently unclear regarding the appropriate functional form (i.e., linear or curvilinear) for the regression of performance onto conscientiousness. Second, it appears that the ideal point model of item responding can be considered more appropriate than dominance models for application to personality responses. Notably, these two areas of applied personality research have not yet merged. Prior work investigating the curvilinear personality–performance link used CTT scoring of personality items (i.e., the average item score), a method that carries the assumptions of the dominance model. Further, the one study investigating the criterion-related validity of ideal point scores of personality measures considered only linear assessments of criterion-related validity (i.e., correlation).

We make two basic propositions based on the observations above. First, the inconsistency of findings regarding the curvilinear personality–performance link may be due to the use of a less theoretically appropriate scoring approach (i.e., using dominance-based CTT scoring), resulting in greater errors in measurement. Generally, we suggest this greater measurement error obscures the true relationship between personality and performance, resulting in the inconsistent findings observed in the literature. Therefore, we pose the following research question (RQ):

> *RQ1:* Do estimates of conscientiousness derived from an ideal point IRT model show a curvilinear personality–performance relationship more consistently than dominance-based CTT, FA, and IRT conscientiousness scores?

Note that we include dominance FA and IRT scoring to discern the influence of the generally higher reliability of latent trait estimates over the CTT scores from the type of response model (i.e., dominance versus ideal point). The estimate derived by the CTT total score does not attempt to partition "true" and "error" variance, whereas the FA and IRT estimates are more "purified" in that they take account of item features (e.g., factor loadings and IRT item location and discrimination, respectively) to better estimate trait standing. Therefore, we did expect that FA and IRT model-based estimates would uncover more curvilinear relationships as a predictor than the CTT score. However, we did not expect them to uncover curvilinear relationships as consistently as the ideal point model, because of their adherence to dominance assumptions (i.e., that higher CTT scores always indicate higher trait standing).

Our second RQ addresses both a theoretical and practical outcome of the current research. As suggested above, the downwardly biased dominance-based score received by excessively conscientious persons should change the rank-ordering for a small group of respondents at the top of the trait continuum. The different (and more accurate) ordering achieved by ideal point modeling will correctly order excessively conscientious persons as higher than their more moderate counterparts. Because the changes in rank-order apply to only a small proportion of the individuals (and therefore a small amount of covariance), we did not expect large increases in R^2. More specifically, rather than seeing a large increase in R^2, we would expect to see more accurate selection decisions for individuals at the top end of the predictor distribution:

> *RQ2:* Do better predictions result from combined use of the ideal point model at the level of measurement and a curvilinear predictive model?

Such a finding would imply more consistency between theoretical and empirical views of the conscientiousness–performance relationship and would also have significant implications for selection decisions at the individual level.

Below we consider these RQs in a series of three studies. The first two studies examine the relationship between conscientiousness and a variety of performance dimensions. We compare results of curvilinear regression analyses resulting from the use of CTT scores and factor analyses, a dominance IRT model (i.e., the GRM) and an ideal point IRT model (i.e., the GGUM), using a variety of important and commonly used performance outcomes as criteria. In the third study, we apply the findings of Study 2 to a new data set to evaluate model accuracy in predicting reasons for turnover (TO) and the occurrence of corrective actions (CAs) taken against the employee.

Study 1

Method

Sample. Data were collected as part of a large-scale validity study by a large international consulting firm. The data set consisted of 1,258 participants' responses to a 322-item survey that included both Likert-type self-report and situational judgment items. The supervisors of these employees also completed a 36-item set of performance ratings made on a 10-point scale. Participants were mostly male (61.5%), and approximately half were non-White (49.9%). All participants were informed that their responses were collected for research purposes and that the responses would be kept confidential.

Conscientiousness measure and scoring. Fifteen personality items from the firm's in-house measures were selected by the researchers to reflect a unidimensional conscientiousness scale (part of a broader compound measure used by the firm). All items were self-report Likert-type items on a 5-point scale ranging from *Strongly Agree* to *Strongly Disagree*. An example item is "I am often late for scheduled appointments." Four scoring approaches were used. First, a CTT approach was taken by using the average item score across the 15 items (with appropriate reverse-coding for negatively worded items) as an indicator of conscientiousness. The same CTT coding was used for FA and GRM scoring.

FA scores were obtained by estimating a one-factor principal axis solution and taking the regression-based factors scores in the SPSS v20 software package. Factor loadings ranged from .22 to .59 (M = .39, SD = .11). GRM item parameters were estimated using marginal maximum likelihood (MML) and person parameters using maximum a posteriori (MAP) scoring with default settings for Multilog v7.0 (Thissen, 2003). CTT scores, regression-based factor scores, and GRM person parameter estimation represent three dominance approaches to scoring. Finally, an ideal point scoring method was conducted. For this approach, raw codes for all items were used with no reverse coding. Item parameters were estimated using MML and person parameters were estimated using expected a posteriori (EAP) via the GGUM2004 (J. S. Roberts, Fang, Cui, & Wang, 2006) software program. Table 2 shows coefficient alpha for the CTT measure, and means, standard deviations, and intercorrelations of the CTT, FA, GRM, and GGUM scores.

As with all IRT analyses, it is important to ensure that scores resulting from the GRM and GGUM IRT models fit the data at hand. To address model–data fit, we calculated adjusted (to N = 3,000) χ^2/df ratios for each model using the MODFIT v2.0 (Stark, 2007) program with latent trait density estimation via an expectation-minimization (EM) algorithm for item singles, doubles, and triples (see Table 3). Item singles are a measure of the difference between the observed scores in the data and the scores that would be expected by the IRT model. Item singles suggested good fit for both the GRM and GGUM, with all values falling below the suggested cutoff of 3 (see Drasgow, Levine, Tsien, Williams, & Mead, 1995). In fact, all 15 items showed χ^2/df ratios < 1 for both the GRM and GGUM (see Table 3). Doubles and triples showed means larger than the suggested cutoff of 3, suggesting some potential problems with local dependence, an issue addressed in more detail later in this section. However, it has been recently suggested (Tay, Ali, Drasgow, & Williams, 2011) that the χ^2/df ratios < 3 criterion is inappropriate for doubles and triples, and these results are consistent with past IRT model–data fit analyses of personality measures (e.g., see Stark et al., 2006). Given these results, we concluded that person parameters would be well-interpreted and continued with our substantive analyses.

A second assumption of the IRT models utilized for scoring the conscientiousness measure is that a single dimension underlies scores. To address this assumption, we fit a second-order confirmatory factor analysis (CFA) to the responses with one higher order factor (i.e., conscientiousness) and seven facet-level latent variables. The model showed marginally acceptable fit to the data, with root-mean-square error of approximation (RMSEA) of .085, 90% confidence interval (CI) of [081, .091], standardized root-mean-square residual (SRMSR) of .081, non-normed fit index (NNFI) of .84, and $\chi^2(82) = 843.77$, $p < .001$. We compared this to a model with two higher order factors corresponding to the firms two in-house scales utilized to create this measure. The two-factor model showed slightly better fit to the data, $\chi^2(89) = 769.29$, $p < .001$; $NNFI$ = .87; $SRMSR$ = .079; $RMSEA$ = .078, 90% CI [.077, .088]. Comparing the two models they appear essentially equivalent, given their overlapping RMSE confidence intervals and very small differences in other fit indexes (i.e., NNFI, SRMSR). Further, the intercorrelation of factors for the two higher order factors was high, at .64. Therefore, we concluded that a single, latent variable underlies the items utilized here to measure conscientiousness.

As one reviewer pointed out, the hierarchical structure is indicative of the problems with local dependence mentioned above in the IRT model–data fit analyses. Notably, simple one- and two-factor models did not show adequate model–data fit, with $RMSEA$ of .106 and .102, respectively. Although the hierarchical structure indicates a violation of the assumption of local dependence, the violation is small. The degree of violation of the assumption of local independence can be assessed by considering the prepotency of the general factor. This involves comparing the size of the effect of the higher order factor onto the lower order factors to the size of the effect of the lower order factors on the observed variables (see Stark, Chernyshenko, & Drasgow, 2002). For all but one of the seven lower order traits, all item loadings were smaller than the loading of its respective trait onto the general factor. For the standardized solution, loadings of lower order onto the higher order factor ranged from .49 to 1.26 with a mean of .98 (SD = .26).[2] Item loadings ranged from .22 to .84 (M = .52, SD = .21). In other words, the relative effect of the general factor was greater than the effect of the specific factor on the observed variables with the exception of 3 of 15 items, all associated with the same specific factor. This relatively small violation of local independence is reflected in the somewhat high χ^2/df doubles and triples in Table 3. More detailed results can be obtained by contacting the authors.

In addition to model–data fit and dimensionality, another consideration here was the test information function (TIF) which reflects the reliability of conscientiousness scores across the trait continuum. As can be seen in Figure 3, the information function for the GRM showed slightly higher reliability at the low end of the continuum versus the high end, whereas the GGUM showed the highest reliability at the low and high ends of the trait continuum but relatively low reliability at more moderate levels. Importantly, the GGUM showed more information than the GRM at the high end.

Questions regarding construct validity may arise with respect to the post hoc construction of the conscientiousness measure. A series of analyses supported the construct validity of these measures; details of these analyses may be found in the Appendix of this article.

Performance measures. The first two authors examined the 36 performance rating items to identify appropriate items for TP, OCB, and CWB. This resulted in a five-item TP measure that reflected learning and following rules and procedures, a four-item OCB measure that included items concerning helping coworkers and productive use of downtime, and a two-item measure of CWB

[2] As one reviewer noted, a common misunderstanding in interpreting factor loadings in a completely standardized CFA solution is that loadings should not be greater than one. This misunderstanding has been attributed by Jöreskog (1999) to the fact that loadings in a standardized exploratory factor analysis with uncorrelated factors, loadings are analogous to correlation coefficients. However, when factors are correlated (or regressed onto a higher order factor), the loadings are analogous to regression coefficients and therefore may be larger than one. See Jöreskog (1999) for technical details on this issue.

Table 2
Intercorrelations Between Study Variables and Internal Consistency Estimates of Observed Scores for Study 1

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Conscientiousness (CTT)	3.83	0.41	*.72*						
2. Conscientiousness (FA)	0.014	0.86	.98**	—					
3. Conscientiousness (GRM)	0.022	0.82	.99**	.97**	—				
4. Conscientiousness (GGUM)	0.015	0.85	.82**	.86**	.82**	—			
5. Task performance	5.41	1.62	.11**	.11**	.10**	.09**	*.92*		
6. Organizational citizenship behavior	5.21	1.40	.09**	.09**	.09**	.06*	.82**	*.82*	
7. Counterproductive work behavior	6.44	1.10	−.12**	−.10**	−.10**	−.05*	−.48**	−.38**	*.77*

Note. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model. Values on the diagonal in italics represent coefficient alpha for the specified measure.
* $p < .05$. ** $p < .01$.

that concerned breaking rules and stealing the property of the employer. Coefficient alpha, mean and standard deviations for these scales are included in Table 2.

As with the measure of conscientiousness, some questions may arise regarding the construct validity of the performance measures. Again, the Appendix provides details regarding the analyses that show these measures are construct valid.

Data analysis. To examine the relationship between the four conscientiousness scores and the three performance scores, we conducted 12 hierarchical polynomial regressions from each combination of performance regressed onto the conscientiousness score. All conscientiousness scores were standardized, and the polynomial (i.e., curvilinear) term was calculated from that standardized value to avoid multicollinearity (Aiken & West, 1991; Dalal & Zickar, 2012). In the first step, the standardized conscientiousness scores were entered as a predictor of the performance dimension (e.g., TP), and the change in R^2 was evaluated for significance. In the second step, the squared value of the standardized conscientiousness scored was entered as an additional predictor, and the change in R^2 was evaluated for significance.

Results

Table 4 shows the results for all 12 regression analyses at each step by the type of scoring used as a predictor. In Step 2, only the GGUM score showed a significant curvilinear effect with all three performance outcomes, confirming the expectations surrounding RQ1. The FA and GRM scores showed the same results as the CTT score, in that significant curvilinear relationships were found for TP and CWB but not for OCB. Notably, the change in R^2 for the addition of the curvilinear term was always largest for the GGUM predictor. Figure 4 shows all quadratic regression curves for the three performance variables regressed onto CTT, FA, GRM, and GGUM conscientiousness scores.

Table 5 shows the adjusted R^2 for the most complex model retained in each regression analysis based on scoring method. For TP, all scoring procedures showed a significant curvilinear effect, with the highest effect for FA, then a tie between CTT and GGUM, and finally GRM. For OCB and CWB, the GGUM score explained the most variance, though differences between predictors were small for all outcomes.

As shown in Table 4, the incremental R^2, though statistically significant, are generally small in magnitude. That is, it would appear that in most instances the incremental contribution of the curvilinear effect is practically insignificant. However, as noted with RQ2, we expected that the added value of utilizing the GGUM score would be realized in more accurate selection decisions, but not necessarily large increases in R^2. To assess the increase in selection decision accuracy, we conducted analyses similar to those presented by Bing et al. (2007). This technique involves three main steps: First, for each predictor score the best fit of regressions presented in Table 4 were used to calculate a predicted performance value, $\hat{Y}$. Second, all individuals were rank-ordered based on $\hat{Y}$, and the top 10 and top 20 "applicants" were selected. Third, we calculated the mean of the actual performance scores, Y, for those selected. To the extent that a particular scoring method results in more accurate decisions, the mean of actual performance should be higher (note, for CWB, the goal is to select the 10 and 20

Table 3
Model-Data Fit Adjusted (to N *= 3,000)* χ^2*/df Ratios From MODFIT 2.0 Program (Stark, 2007) for Study 1*

Model	Statistic	<1	1 < 2	2 < 3	3 < 4	4 < 5	5 < 7	>7	*M*	*SD*
GRM	Singlets	15	0	0	0	0	0	0	.034	.025
	Doublets	0	2	2	3	2	4	2	5.389	1.447
	Triplets	0	0	1	0	2	1	1	5.571	0.89
GGUM	Singlets	15	0	0	0	0	0	0	.002	.002
	Doublets	0	1	3	2	6	3	0	4.079	4.289
	Triplets	0	0	0	1	2	2	0	4.975	3.003

Note. Values in cells represent counts of the number of items within the range specified at the top of the column. Values less than 3 indicate good model–data fit (Drasgow et al., 1995). GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

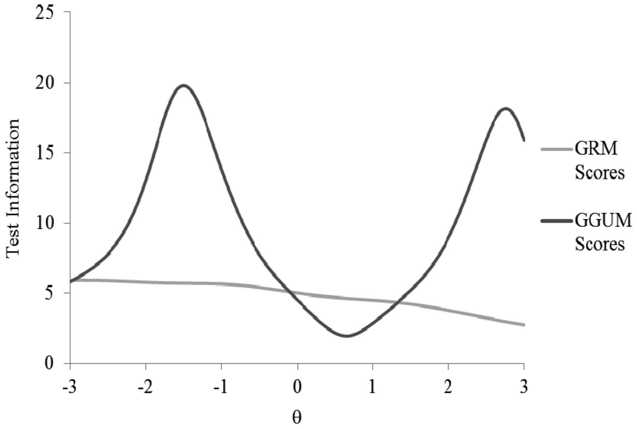

Figure 3. Test information functions for the Study 1 conscientiousness measure under the GRM and the GGUM. GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

highest scores, reflecting a “select out” strategy). The results in Table 6 show mean *Y* for the top 10 and top 20 persons (as determined by ranking on $\hat{Y}$). As can be seen, for all but one case (i.e., Top 20 on predicted TP), using the GGUM as the predictor in a curvilinear regression lead to the best performers being selected supporting our expectations, outlined in RQ2. For CWB, using the GGUM would more accurately identify the top 10 and top 20 persons who engaged in the most CWB facilitating screening out applicants at high risk for counterproductive behavior.

Discussion

Results of study 1 support the notion that ideal point scores consistently uncover curvilinear personality–performance relationships (i.e., RQ1). Whereas the CTT, FA, and GRM scores showed a curvilinear relationship only for the conscientiousness–TP and conscientiousness–CWB relationships, the GGUM showed significant curvilinear effects for all three criteria.

Although the variance explained overall by the GGUM was consistently similar to other scores overall, this does not necessarily indicate the GGUM scores are not more appropriate. In fact, the GGUM (in concert with past research) was shown to be the preferred model compared to the GRM in terms of model–data fit and showed high reliability, particularly at the extremes of the conscientiousness continuum, where the differences between curvilinear and linear models would likely occur (see Figure 3). We believe this difference in measurement precision drives the more consistent detection of curvilinear trends using GGUM scores as predictors.

These results support the idea that merging tests of the personality–performance relationship with ideal point personality scoring results in more consistent conclusions. Additionally, we found partial support for the idea that such a merger would shed light on the somewhat confusing finding of null differences in criterion-related validity for personality when using ideal point approaches to scoring in spite of increases in desirable psychometric properties. However, these results were limited to one measure of conscientiousness and one set of

Table 4
Results of Hierarchical Quadratic Regression Analysis for Each Outcome Variable by Type of Conscientiousness Estimate (CTT, FA, GRM, GGUM) for Study 1

		Conscientiousness estimate							
		CTT score		Factor scores		GRM θ		GGUM θ	
Outcome	Predictor	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$
Task performance	Step 1								
	Intercept	5.369**	.012** (.012**)	5.410**	.013** (.013**)	5.407**	.010** (.010**)	5.407**	.007** (.007**)
	Linear	.261**		.181**		.165**		.140**	
	Step 2								
	Intercept	**5.442****	**.017** (.005*)**	**5.505****	**.020** (.007**)**	**5.477****	**.015** (.005*)**	**5.517****	**.017** (.010**)**
	Linear	**.285****		**.161****		**.190****		**.192****	
	Quadratic	**−.153***		**−.096****		**−.072***		**−.115****	
Organizational citizenship behavior	Step 1								
	Intercept	5.178**	.009** (.009**)	5.208**	.009** (.009**)	5.206**	.008** (.008**)	5.205**	.003 (.003)
	Linear	.188**		.130**		.126**		.082	
	Step 2								
	Intercept	5.209**	.010** (.001)	5.260**	.012** (.003)	5.240**	.009** (.002)	**5.291****	**.011** (.008**)**
	Linear	.199**		.119**		.139**		**.122****	
	Quadratic	−.065		−.053		−.035		**−.089****	
Counterproductive work behavior	Step 1								
	Intercept	6.465**	.015** (.015**)	6.422**	.009** (.009**)	6.424**	.010** (.010**)	6.436**	.003 (.003)
	Linear	−.197**		−.106**		−.110**		−.060**	
	Step 2								
	Intercept	**6.410****	**.021* (.006*)**	**6.356****	**.017** (.007**)**	**6.379****	**.014** (.004*)**	**6.328****	**.024** (.021**)**
	Linear	**−.215****		**−.092****		**−.127****		**−.111****	
	Quadratic	**.115***		**.067****		**.046***		**.114****	

Note. $N = 1{,}030$ to 1,033 after listwise deletion. Boldface values indicate instances where the quadratic regression was significant. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.
* $p < .05$. ** $p < .01$.

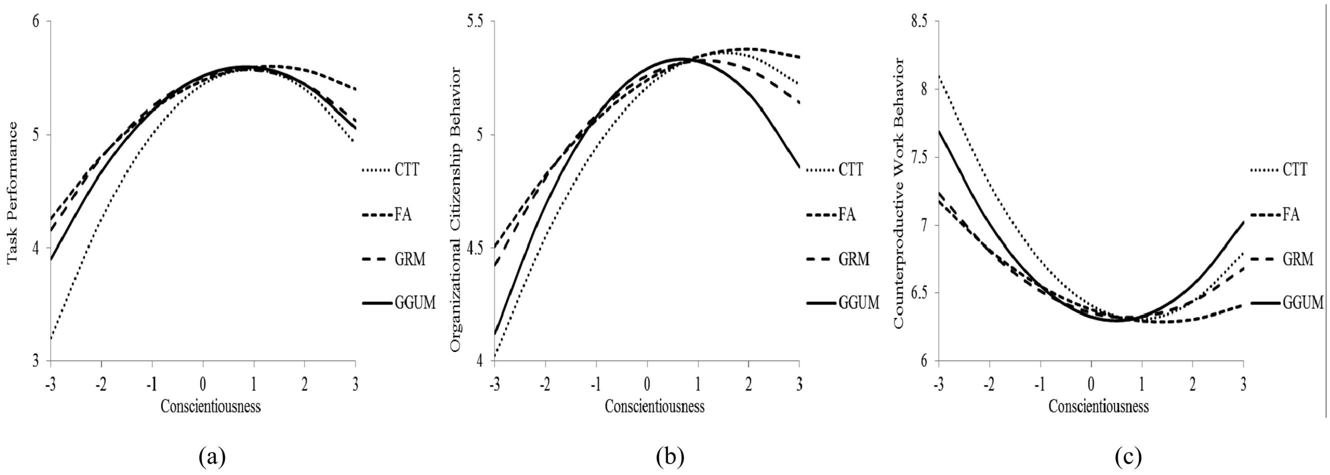

Figure 4. Quadratic regression lines for CTT, FA, GRM, and GGUM conscientiousness predicting (a) task performance; (b) organizational citizenship behavior; and (c) counterproductive work behavior in Study 1. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

performance measures in a single combined sample. Therefore, we sought to replicate these results in another sample using additional performance dimensions in Study 2.

Study 2

Method

Sample. The second set of data was collected as part of a large-scale validity study by another large international consulting firm. The data set consisted of 1,570 employees of a large retail chain. These participants responded to a 175-item survey that included a variety of item types (e.g., Likert-type self-report, situational judgment) and participated voluntarily, being told their responses would be kept confidential and that data were being collected for research purposes. Likert-type items were on a 6-point *Strongly Agree* to *Strongly Disagree* scale. The supervisors of these employees also completed a 57-item set of performance ratings. Twenty-nine of the performance items used an 8-point Likert-type scale of agreement, whereas 24 CWB items were answered with a 2-point “Yes” or “No” scale regarding whether the behavior had been observed. Participating employees were mostly male (60.8%) and White (61.5%).

Conscientiousness measure and scoring. Ten personality items from the firm’s in-house measures were selected by the first two authors to reflect the personality dimension of conscientiousness (a measure not currently used by the firm). All of the selected items were self-report items using a 6-point Likert-type scale ranging from *Strongly Disagree* to *Strongly Agree*. An example

Table 5
Adjusted R^2 *for the Most Complex Significant Regression Model of Performance Regressed on Conscientiousness in Study 1*

Performance dimension	Conscientiousness score			
	CTT	FA	GRM	GGUM
1. Task performance	.015 (C)	**.017** (C)	.013 (C)	.015 (C)
2. Organizational citizenship behavior	.008 (L)	.008 (L)	.008 (L)	**.009** (C)
3. Counterproductive work behavior	.019 (C)	.015 (C)	.012 (C)	**.022** (C)

Note. (L) indicates the linear model was the most complex significant model, whereas (C) indicates the curvilinear model was the most complex. Values in boldface indicate the model with the most variance explained with downward adjustment for the number of predictors in the model (i.e., adjusted R^2). CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

Table 6
Mean and Standard Deviation of True Criterion Variable for Persons Ranked in the Top 10 and Top 20 of the Respective Predicted Criterion Value for Study 1

Criterion	Conscientiousness score	Number selected, *n*			
		n = 10		*n* = 20	
		M	*SD*	*M*	*SD*
Task performance	CTT	5.94	1.81	**5.94**	**1.60**
	FA	5.42	1.86	4.96	1.63
	GRM	5.30	2.01	5.60	1.73
	GGUM	**6.32**	**1.59**	5.90	1.88
Organizational citizenship behavior	CTT	5.55	1.21	5.63	1.09
	FA	5.53	0.65	5.49	1.06
	GRM	5.43	0.68	5.49	1.06
	GGUM	**6.11**	**1.33**	**5.65**	**1.19**
Counterproductive work behavior[a]	CTT	7.20	1.29	7.20	1.11
	FA	7.10	1.22	6.93	1.00
	GRM	6.95	1.21	6.85	1.06
	GGUM	**7.35**	**0.69**	**7.35**	**0.87**

Note. Values in bold indicate the most desirable selection outcome. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.
[a] For counterproductive work behavior, the top 10 and 20 were selected out as opposed to selected in.

item is "I usually get my work done on time." Similar to the first study, four scoring approaches were used. First, a CTT approach was taken by calculating the average item score across the 10 items as an indicator of conscientiousness. Additionally, the FA, GRM, and GGUM were estimated using the same approaches discussed in Study 1. Means and standard deviations of all scoring procedures are included in Table 7.

For FA scores, a one-factor principal axis solution was obtained. Factor loadings ranged from .34 to .71 ($M = .56$, $SD = .13$). For the IRT analyses, we again calculated adjusted χ^2/df ratios for each model using the MODFIT v2.0 (Stark, 2007) program with latent trait density estimation via an expectation-minimization (EM) algorithm for item singles, doubles, and triples (see Table 8). Item singles suggested good fit for both the GRM and GGUM, with all values falling below the suggested cutoff of 3 (see Chernyshenko et al., 2001; Drasgow et al., 1995). However, both models showed doubles and triples greater than three. Surprisingly, mean ratios for doubles and triples were higher for the GGUM than the GRM, though neither suggested particularly good fit. This is consistent with past research that has shown slightly better fit for the 2-parameter logistic model (a special case of the GRM) compared to the GGUM for doubles and triples for some personality measures when item singles suggest better fit for the GGUM (see Stark et al., 2006). Moreover, as noted previously, the χ^2/df ratios < 3 criterion has been shown to be inappropriate for doubles and triples (Tay et al., 2011). As can be seen in Figure 5, the information function for the GRM and GGUM showed higher reliability at the low end of the continuum versus the high end but relatively low reliability at higher levels. Generally, the GGUM scores were more reliable than the GRM at low levels of Conscientiousness and similarly reliable at the high end.

As in Study 1, we sought to establish unidimensionality evidence for our post hoc measures. The measure of conscientiousness was composed of 10 items from the consulting firm's measures of work ethic, trustworthiness, responsibility, attention to detail, and integrity. As for Study 1, we provide internal (e.g., CFA) and external (e.g., correlational) construct validity evidence. To begin, we estimated a one-factor model CFA model.[3] Although some fit indices were higher than convention (i.e., $RMSEA = .099$, 90% CI [.093, 1.07]; $\chi^2(35) = 580.72$), other fit indices suggested overall strong fit (i.e., $SRMSR = .039$; $NNFI = .959$). Moreover, the mean standardized factor loading was .72 ($SD = .13$). Overall, this evidence suggests a single factor structure is most appropriate. We also found support for the construct validity of this measure; details are provided in the Appendix.

Performance measures. The first two authors examined the 57 performance rating items to identify appropriate items for TP and OCB. The final measure of TP included 16 items; the OCB measure included seven items. Evidence of the construct validity of these two performance measures is provided in the Appendix. Additionally, we used the consulting firm's in-house performance measures of CWB (24 items), safety performance (four items), and global performance (four items) ratings. Means, standard deviations, and coefficent alphas for performance measures are included in Table 7.

Data analysis. To examine the relationship between the four conscientiousness scores (CTT, FA, GRM, and GGUM) and the five performance scores (TP, OCB, CWB, global performance, and safety performance), we conducted 20 hierarchical polynomial regressions from each combination of performance regressed onto each conscientiousness score. As with Study 1, conscientiousness scores were standardized, the polynomial (i.e., curvilinear) term was calculated from that standardized value to avoid multicollinearity (Aiken & West, 1991; Dalal & Zickar, 2012), and the same hiearchical regression technique as Study 1 was conducted.

Results

Table 9 shows the results of the 20 regression analyses at each step by the type of scoring used as a predictor. With the exception of CWB regressed onto GRM and GGUM scores, the linear effects in Step 1 were significant when the performance dimensions were regressed onto the four scoring approaches. In Step 2, only the GGUM score showed a significant curvilinear effect for all five performance outcomes again showing support for RQ1. The CTT and FA scores showed a significant curvilinear relationship only with CWB, whereas GRM scores showed significant curvilinear relations with all outcomes except TP. Figure 6 shows the quadratic regression curves for the TP, FA, OCB, and CWB variables regressed onto CTT, FA, GRM and GGUM conscientiousness scores (i.e., measures of the same performance constructs used in Study 1). Figure 7 shows the quadratic curves for safety performance and global performance regressed onto CTT, FA, GRM, and GGUM scores.

Table 10 shows the adjusted R^2 for the most complex model retained in each regression analysis. Here, variance explained was slightly greater or equal to other models for the GGUM for all performance outcomes. The GGUM scores explained equal variance in OCBs as the CTT score and equal variance in CWBs as both the CTT and GRM scores. The generally small differences in criterion-related validity confirmed our expectations surrounding RQ2, that variance explained would not be greatly affected by the use of an unfolding measurement model.

As with Study 1, we hoped to show that the added value of utilizing the GGUM score would be realized in more accurate selection decisions. Thus, we conducted the Bing et al. (2007) analyses again. The results in Table 11 show mean Y for the top 10 and top 20 persons (as determined by ranking on $\hat{Y}$). As can be seen, for all but two cases (i.e., Top 10 and Top 20 on predicted safety performance), using the GGUM as the predictor in a curvilinear regression produced the best selection decisions. Again, for CWB, the GGUM was better able to identify the top 10 and top 20 individuals who engaged in the most CWB facilitating screening out. These findings confirmed our expectations regarding RQ2, that in spite of little or no change in variance-explained, selection outcomes were usually more favorable when using GGUM scores.

Discussion

Supporting our expectations regarding RQ1, GGUM scoring consistently uncovered more curvliniear relationships than did CTT, FA, or GRM scoring; this effect was extended to a situation with a measure of conscientiousness that was more similar to those typically used in research and practice. The GGUM scores showed all five regressions had significant curvilinear term, whereas CTT

[3] Due to the limited number of indicators per facet, a hierarchical factor structure could not be fit.

Table 7
Intercorrelations Among Study Variables and Internal Consistency Estimates of Observed Scores for Study 2

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Conscientiousness (CTT)	5.18	0.52	*.81*								
2. Conscientiousness (FA)	0.000	0.92	.98**	—							
3. Conscientiousness (GRM)	1.37	0.63	.92**	.94**	—						
4. Conscientiousness (GGUM)	−0.002	0.93	.95**	.97**	.94**	—					
5. Task performance	4.82	1.01	.09**	.07**	.05*	.06*	*.83*				
6. Organizational citizenship behavior	5.63	1.09	.14**	.13**	.12**	.12**	.60**	*.91*			
7. Counterproductive work behavior	1.04	0.07	−.07**	−.05*	−.02	−.03	−.25**	−.43**	*.76*		
8. Safety performance	6.14	0.94	.09**	.08**	.07**	.06**	.28**	.40**	−.33**	*.67*	
9. Global performance	4.89	1.40	.12**	.11**	.09**	.10**	.52**	.77**	−.42**	−.36**	*.92*

Note. Values on the diagonal in italics represent coefficient alpha for the specified measure. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.
* $p < .05$. ** $p < .01$.

and FA scores showed only one, and GRM scores showed four. However, differences in effect sizes were not large, providing partial support for our expectations regarding RQ2.

Potential limitations of the findings in Studies 1 and 2 include (a) all predictive validity evidence is dependent on the calibration data set and (b) all outcome measures were subjective supervisor ratings. Therefore, we conducted a third study to address some of these limitations. To reflect the reality of practice in employee selection, we applied the regression coefficients from the highest R^2 models (see Table 9) in Study 2 to forecast performance for a new sample. That is, predicted values for each criterion in Study 2 were calculated with CTT, FA, GRM, or GGUM scores as predictors. Using these predictions of performance, we made mock selections (i.e., the top 100) based on each scoring approach. We then compared the resulting selection decisions for CTT, FA, GRM, and GGUM predictors in terms of behavioral, objective outcomes: turnover rates (overall and by the reason for turnover), and percent of employees who had received a corrective action.

Study 3

Method

Sample and measures. The third set of data was collected by the same consulting firm as in Study 2. The data set consisted of 1,737 employees of a large retail chain. The participants responded to a survey that included the same 10-item measure of conscientiousness as in Study 2 but responded on a 5-point (as opposed to 6 in Study 2) *Strongly Agree* to *Strongly Disagree* scale, which showed similar results for model–data fit, and information.[4]

Information was available regarding whether that employee had left the company, the reason for turnover, and whether employees had received any corrective actions. Reasons for turnover included (a) attendance problems; (b) issues with person–environment fit; (c) performance problems; (d) behavioral problems; (e) for abandoning their job; or (f) leaving to take a new job.

Data analysis. To examine whether the use of GGUM predictor scores of conscientiousness in a curvilinear model would result in better selection outcomes as compared to CTT, FA, or GRM scores, we conducted a three-stage process. First, we applied the appropriate coefficients in Table 9 to CTT, FA, GRM, and GGUM predictor scores to calculate a predicted value for each of the five performance measures used in Study 2 (using the best fitting predictive model for each predictor). We then sorted each predicted value variable to select the top 100 "applicants." This was done for predicted values of each criterion (TP, OCB, CWB, safety performance, and global performance) using each predictor (CTT, FA, GRM, and GGUM). Finally, we computed rates of overall turnover, rates for various turnover reasons (e.g., poor performance, accepting a new job), and rates of corrective action for these selected individuals.

Results

Table 12 shows the turnover by reason (and overall), and corrective action rates that resulted from selection based on values of predicted TP, OCB, CWB, safety performance, and global performance. Most pertinent here is the percent rate difference when using GGUM predictor scores compared to the CTT, FA, and GRM scores; percent differences greater than an absolute value of 30% are highlighted to facilitate this discussion. Using the GGUM score as a predictor and selecting individuals with the top 100 predicted-TP scores resulted in a 33% decrease in turnover due to performance, a 100% decrease in turnover due to behavior, and a 42% decrease in turnover due to job abandonment. Additionally, use of the GGUM resulted in an increase in turnover due to leaving for a new job (128%). Similarly, the GGUM as a predictor and selecting individuals with the top 100 predicted-OCB scores showed considerable decreases in turnover due to performance (66%) and job abandonment (41%), and an increase in turnover due to leaving for a new job (42%). Selecting the top 100 individuals on predicted CWB from the GGUM score compared to the CTT score resulted in a 100% decrease in turnover due to performance (brining the rate to 0), and behavior (67%), whereas FA and GRM scores performed the same as the GGUM scores. The top 100 individuals on predicted safety performance when predicted with GGUM scores showed a decrease in turnover due to behavior (100%) compared to all other predictors, and decreases in turnover due to performance (33%) and job abandonment (50%) compared to CTT and FA scores (the GRM performed similarly well compared to the GGUM scores). A substantial increase in turnover due

[4] Full results of these psychometric analyses can be obtained by contacting the first author.

Table 8
Model–Data Fit Adjusted (to N = *3,000)* χ^2*/df Ratios From MODFIT 2.0 Program (Stark, 2007) for Study 2*

Model	Statistic	<1	1 < 2	2 < 3	3 < 4	4 < 5	5 < 7	>7	*M*	*SD*
GRM	Singlets	10	0	0	0	0	0	0	0.001	0.002
	Doublets	2	1	8	13	4	6	11	5.515	4.178
	Triplets	0	0	9	19	24	34	34	6.108	2.821
GGUM	Singlets	10	0	0	0	0	0	0	0.000	0.000
	Doublets	1	0	1	4	7	14	18	8.761	6.603
	Triplets	0	0	0	2	4	25	89	9.902	3.941

Note. Values in cells represent counts of the number of items within the range specified at the top of the column. Values less than 3 indicate good model–data fit (Drasgow et al., 1995). GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

to fit (43%) were observed compared to use of GRM scores, and an increase in turnover due to leaving for a new job were observed compared to all other predictors (128% compared to CTT and FA, and 60% compared to GRM). Finally, when the top 100 individuals based on predicted global performance (with GGUM scores as the predictor) were selected, decreases in turnover due to performance (66%) and job abandonment (33%) were observed compared to other predictors. Importantly, averaging across all predicted criteria (see Figure 8 and Table 12), the use of the GGUM predictive model led to substantial decreases in turnover due to performance, behavior problems, and job abandonment.

Discussion

The results of Study 3 further buttressed the results of Studies 1 and 2 by showing that GGUM predictive models selected applicants who were less likely to turn over for reasons that might be viewed as detrimental to the organization. We observed decreases in turnover due to fit, performance, behavior, and job abandonment when we averaged across criterion dimensions. Decreases in such behavior would be associated with the selection of those with an adaptive (rather than maladaptive) level conscientiousness, particularly in responsible behavior and impulse control.

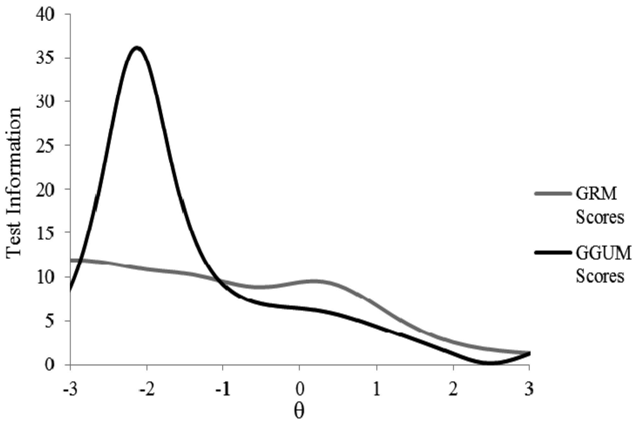

Figure 5. Test information functions for the Study 2 conscientiousness measure under the GRM and the GGUM. GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

Notably reflecting the achievement motive associated with the conscientiousness trait, those selected by the GGUM were more likely to turn over for a new job. One explanation for this finding is that these employees were the top performers who were highly attractive job candidates at other organizations. That is, these moderately conscientious, high-performing individuals are likely to be productive, and therein attractive to other employers, as well as motivated to move on to better or more fulfilling jobs. This suggests that those implementing this approach might need to consider retention efforts to avoid losing these adaptive, high-performing, achievement-oriented individuals.

Overall, the apparent lack of relation between conscientiousness and overall turnover is in line with past researchers whom have found no indication of a direct link (e.g., Orvis, Dudley, & Cortina, 2008). However, when considering reasons for turnover across predicted performance dimensions, our findings are in line with the construct domain of adaptive levels of conscientiousness.

General Discussion

The relationship between personality and performance is scientifically important for a complete understanding of workplace behavior in a variety of domains and is practically important to organizations utilizing personality measures for employee selection, placement, and promotion efforts. Recently emerging perspectives in applied personality research have suggested that the functional form of this relationship may be curvilinear, as opposed to the commonly assumed linear form (e.g., Pierce & Aguinis, 2013). Our results support this position.

Of importance, the results of these studies suggest theoretically appropriate scoring is critical for our understanding of the personality–performance relationship. In response to our first research question, when conscientiousness estimates were derived using the ideal point approach, 100% of the regressions showed a significant curvilinear trend. On the other hand, the CTT and FA approaches showed only 37.5%, and the GRM showed 75% of these regressions had a significant curvilinear term. Further, predicted outcomes were almost always higher for the model using the GGUM predictor suggesting better decision making would result. Finally, we found that application of Study 2 regression results in a mock selection procedure showed changes in behavioral outcomes consistent with adaptive levels of conscientious-

Table 9
Results of Hierarchical Quadratic Regression Analysis for Each Outcome Variable by Type of Conscientiousness Estimate (CTT, FA, GRM, GGUM) for Study 2

		Conscientiousness estimate							
		CTT score		Factor scores		GRM θ		GGUM θ	
Outcome	Predictor	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$	*B*	$R^2(\Delta R^2)$
Task performance	Step 1								
	Intercept	4.826**	.008** (.008**)	4.826**	.004* (.004*)	4.826**	.003* (.003*)	4.824**	.004* (.004*)
	Linear	.090**		.066*		.057*		.061*	
	Step 2								
	Intercept	4.832**	.008** (.000)	4.828**	.004* (.000)	4.864**	.006* (.003)	**4.907****	**.011* (.007*)**
	Linear	.086**		.064*		.069*		**.094****	
	Quadratic	−.006		−.002		−.038		**−.082****	
Organizational citizenship behavior	Step 1								
	Intercept	5.633**	.019** (.019**)	5.633**	.017** (.017**)	5.633**	.014** (.014**)	5.632**	.015** (.015**)
	Linear	.153**		.143**		.130**		.133**	
	Step 2								
	Intercept	5.643**	.020** (.001)	5.640**	.017* (.000)	**5.677****	**.017** (.003*)**	**5.712****	**.021** (.006**)**
	Linear	.147**		.137**		**.144****		**.165****	
	Quadratic	−.009		−.007		**−.044***		**−.080****	
Counterproductive work behavior	Step 1								
	Intercept	1.039**	.005* (.005*)	1.039**	.002 (.002)	1.039**	.002 (.002)	1.039**	.001 (.001)
	Linear	−.005*		−.004		−0.003		−.002	
	Step 2								
	Intercept	**1.036****	**.009** (.004**)**	**1.036****	**.008** (.006**)**	**1.034****	**.010** (.008**)**	**1.032****	**.010* (.009*)**
	Linear	**−.003**		**−.002**		**−.005***		**−.005***	
	Quadratic	**.003****		**.003****		**.005****		**.007***	
Safety performance	Step 1								
	Intercept	6.148*	.007* (.007*)	6.148**	.006** (.006**)	6.148**	.005** (.005**)	6.147*	.004* (.004*)
	Linear	.081*		.074**		.066**		.058*	
	Step 2								
	Intercept	6.171*	.009* (.002)	6.170**	.009** (.002)	**6.196****	**.010** (.005**)**	**6.208***	**.012* (.008*)**
	Linear	.066*		.057*		**.081****		**.072***	
	Quadratic	−.023		−.022		**−.048****		**−.060***	
Global performance	Step 1								
	Intercept	4.892*	.014* (.014*)	4.892**	.012** (.012**)	4.892**	.010** (.010**)	4.891*	.010* (.010*)
	Linear	.167*		.156**		.138**		.142*	
	Step 2								
	Intercept	4.907*	.014* (.000)	4.901**	.012** (.000)	**4.950****	**.013** (.003*)**	**4.993***	**.016* (.006*)**
	Linear	.157*		.148**		**.157****		**.182***	
	Quadratic	−.015		−.009		**−.058***		**−.102***	

Note. N = 1,426 to 1,429 after listwise deletion. Boldface values indicate instances where the quadratic regression was significant. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.
* $p < .05$. ** $p < .01$.

ness. These findings support the approach as beneficial for actuarial prediction in employee selection.

From these results it is clear that ideal point scoring can help detect curvilinear trends where they exist. It would also appear that simply utilizing more reliable scoring (i.e., IRT model scoring) is insufficient. If detection were simply a by-product of more reliable scoring, the GRM and GGUM scoring would have resulted in the same number of significant curvilinear relationships. As this was not the case, we suggest that IRT based scoring is not sufficient for accurate recovery of curvilinear personality–performance relationships. Adequately detecting curvilinear relationships between personality and performance requires attention to theoretically appropriate scoring. All past studies that have found (e.g., LaHuis et al., 2005; Le et al., 2011) and not found (e.g., Robie & Ryan, 1999) evidence of curvilinearity have utilized the CTT approach to scoring. Were these studies reanalyzed utilizing ideal point scoring, our results suggest they might very well find consistent evidence of the curvilinear trend.

As suspected by past theoretical work (Grant & Schwartz, 2011; Pierce & Aguinis, 2013), the inverted *U*-shaped relation between traits assumed to have positive outcomes no matter how high may, in fact, have a breaking point. Beyond this point, these traits are likely to be associated with (at best) limited gains in behavioral efficacy, and (at worst) maladaptive, extreme behavior. Our results suggest that when conscientiousness is scaled appropriately (i.e., using an ideal point framework), this curvilinear relationship is realized. On the other hand, the use of dominance measurement models can mask curvilinear effects and make it appear as though the commonly assumed linear, monotonic function is correct. Further, our results suggest that those with moderate total scores and moderate conscientiousness result in the most fruitful outcomes. Those with very high total scores (e.g., those marking *Strongly Agree* to all items) are neither the highest in conscientiousness; nor will they be the best performers. It is notable that whereas the CTT, FA, and GRM would, the GGUM would not select those responding *Strongly Agree* to all items.

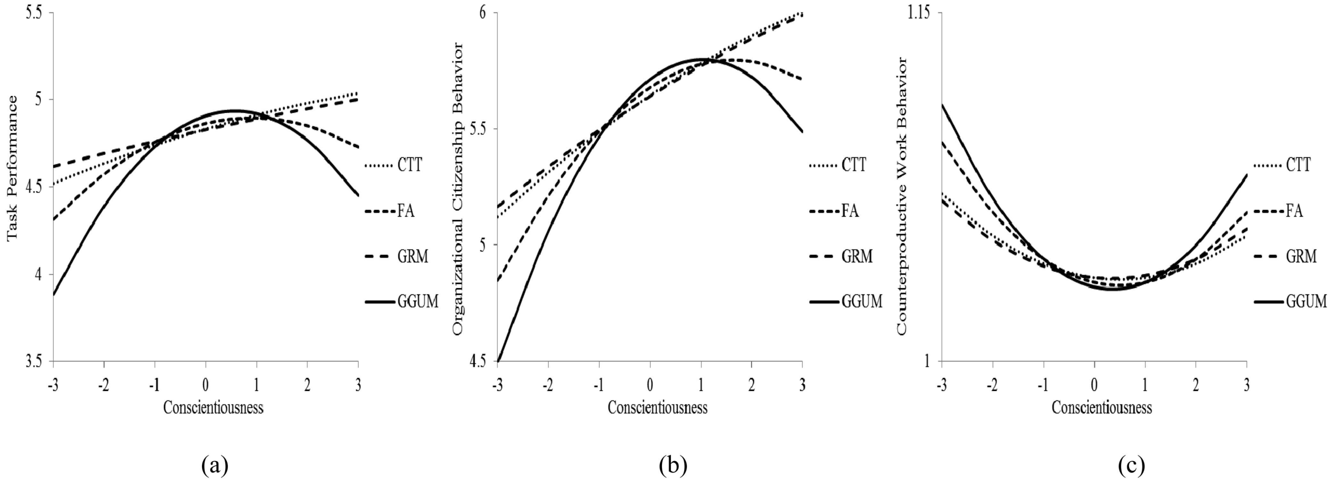

Figure 6. Quadratic regression lines for CTT, FA, GRM, and GGUM conscientiousness predicting (a) task performance; (b) organizational citizenship behavior; and (c) counterproductive work behavior in Study 2. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

Practically, our findings suggest that organizations using personality measures would benefit from utilizing ideal point modeling and considering curvilinear personality–performance links. We showed that the GGUM consistently outpredicted CTT, FA, and GRM in terms of generating superior performance levels among simulated selection groups; we also transported these results to a new data set showing portrayals of behavioral benefit for organizations that is consistent with the desirable aspects of conscientiousness, such as impulse control, higher performance, achievement motivation, and following rules. Across multiple dimensions of performance, using the GGUM with a curvilinear model nearly always produces the best selection decisions.

Limitations and Future Directions

Although our results have important real-world implications, our studies have some drawbacks that limit our generalizations. First, we considered only the use of conscientiousness as a predictor. This was the only trait considered because it is likely the most highly utilized personality trait for performance prediction and is the personality variable shown to predict performance across job types (e.g., Barrick & Mount, 1991), and recent empirical and theoretical work has suggested it specifically as a viable candidate for curvilinearity among the other five factor model traits. Future research should consider other personality measures

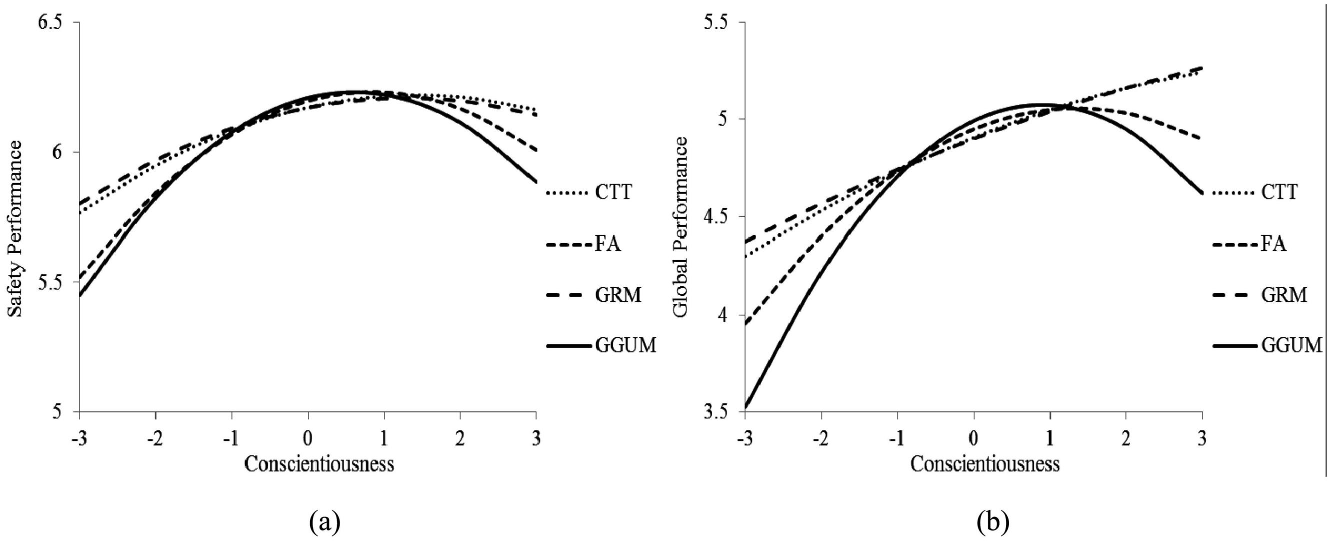

Figure 7. Quadratic regression lines for CTT, FA, GRM, and GGUM conscientiousness predicting (a) safety performance and (b) global performance in Study 2. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

Table 10
Adjusted R^2 *for the Most Complex Significant Regression Model of Performance Regressed on Conscientiousness in Study 2*

	Conscientiousness score			
Performance dimension	CTT	FA	GRM	GGUM
1. Task performance	.007 (L)	.004 (L)	.002 (L)	**.010** (C)
2. Organizational citizenship behavior	**.019** (L)	**.016** (L)	.015 (C)	**.019** (C)
3. Counterproductive work behavior	**.008** (C)	.007 (C)	**.008** (C)	**.008** (C)
4. Safety performance	.007 (L)	.006 (L)	.008 (C)	**.010** (C)
5. Global performance	.013 (L)	.012 (L)	.012 (C)	**.015** (C)

Note. (L) indicates the linear model was the most complex significant model, whereas (C) indicates the curvilinear model was the most complex. Values in boldface indicate the model with the most variance explained with downward adjustment for the number of predictors in the model (i.e., adjusted R^2). CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

scored according to ideal point specifications. We suspect that other traits would show curvilinear relationships with performance, and that such studies may show higher criterion-related validity than previously seen if scored using ideal point IRT models. It may even be the case that, when using an ideal point model, personality characteristics that have historically showed near zero validity with performance criteria, may prove to have significant and meaningful relationships with dimensions of performance.

Another consideration is that the conscientiousness measures used here were built under dominance assumptions. That is, items were developed and retained by these firms based on analyses such as item–total correlations, factor analysis, and/or conventional IRT modeling, all of which carry dominance assumptions. Although the studies reviewed above (e.g., Carter & Dalal, 2010; Stark et al., 2006; Tay et al., 2009; Weekers & Meijer, 2008) show that the ideal point model fits data from dominance scales better than dominance models, measures specifically developed under ideal point assumptions (e.g., Chernyshenko et al., 2007) may show even greater gains in variance explained for curvilinear relations that those found here. Future research should explore this possibility as well.

We also believe our study was limited by the fact that our samples reflected relatively low complexity jobs in fast food and retail chains. Le et al. (2011) found mixed evidence of a moderating role of job complexity in the curvilinear personality–performance relationship for measures of conscientiousness and emotional stability on TP, OCB, and CWB. These results may have been more clear had ideal point scoring been utilized. Moreover, alternative moderators of curvilinear relationships should be considered. For example, curvilinear trends in strong situations may be less pronounced than in weak situations (Mischel, 1977).

The issue of sample size will certainly concern researchers and practitioners (Dalal et al., 2010). Proper estimation of GGUM requires a sample size of 750 or greater for measures with 15 to 20 items (see J. S. Roberts et al., 2000). Further, there is currently no fully appropriate and accessible method for ensuring that measures are unidimensional unfolding. One potential avenue for future research is to determine whether the original Thurstonian scoring approach, which uses raters to calibrate item weights, are successful in achieving the desired properties of GGUM estimates. The standard Thurstonian approach requires somewhat large calibration samples (usually around 300; see Guilford, 1954), but these calibrations can be done on more convenient samples than workers. For example, college students and data sourcing (e.g., M-Turk) could be used for calibration purposes. Indeed, Stark, Chernyshenko, and Guenole (2011) found that rationally derived item weights can approximate ideal point IRT model weights.

Finally, future research should seek to further establish and verify theoretical rationales for both the curvilinear personality–performance relationship as well as the ideal point model for personality responses. We believe two avenues would be potentially fruitful. First, researchers could identify whether individuals who are past or directly on the inflection point of the personality–performance and ideal point curves exhibit signs of response distortion (e.g., using the covariance index; see Burns & Christiansen, 2011). We found that those with the highest average item-score were not often selected using GGUM predictors. In other words, these results suggest that a response pattern that indiscriminately uses *Strongly Agree* would be unlikely to be selected using the GGUM. It follows that application of personality test scoring from an ideal point, curvilinear perspective may alleviate many of the concerns and issues that arise from applicant faking behavior (cf. O'Connell, Kung, & Tristan, 2011; Peterson, Griffith, Isaacson, O'Connell, & Mangos, 2011).

Table 11
Mean and Standard Deviation of True Criterion Variable for Persons Ranked in the Top 10 and Top 20 of the Respective Predicted Criterion Value for Study 2

		Number selected, *n*			
		n = 10		*n* = 20	
Criterion	Conscientiousness score	*M*	*SD*	*M*	*SD*
Task performance	CTT	4.54	0.89	4.30	0.98
	FA	2.71	0.68	3.10	0.62
	GRM	2.71	0.68	3.10	0.62
	GGUM	**4.73**	**1.35**	**4.74**	**1.22**
Organizational citizenship behavior	CTT	3.91	1.38	4.44	1.37
	FA	3.91	1.38	4.44	1.37
	GRM	5.70	0.81	5.82	0.78
	GGUM	**6.10**	**1.07**	**6.09**	**1.06**
Counterproductive work behavior[a]	CTT	1.07	0.10	1.07	0.12
	FA	1.08	0.11	**1.08**	**0.11**
	GRM	1.07	0.11	1.07	0.12
	GGUM	**1.09**	**0.10**	**1.08**	**0.11**
Safety performance	CTT	5.77	1.56	5.98	1.26
	FA	5.77	1.56	5.98	1.26
	GRM	**6.20**	**0.80**	**6.35**	**0.77**
	GGUM	5.97	1.13	6.20	0.97
Global performance	CTT	2.83	1.13	3.45	1.63
	FA	2.83	1.13	3.45	1.63
	GRM	4.97	1.25	4.85	1.48
	GGUM	**5.20**	**1.44**	**4.95**	**1.32**

Note. Values in bold indicate the most desirable selection outcome. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.
[a] For counterproductive work behavior, the top 10 and 20 were selected out as opposed to selected in.

Table 12
Reason for Turnover Rates, Corrective Action, and Worker Compensation Claims of the Top 100 Predicted Performance Scores by Predicted Criteria and Predictor Score Used for Study 3

		Incident rate by predictor score, %				% difference in rate compared to GGUM		
Predicted criterion	Objective outcome	CTT	GRM	FA	GGUM	ΔCTT	ΔFA	ΔGRM
Task performance	TO (overall)	43	43	43	45	5	5	5
	TO (attendance)	7	7	7	9	29	29	29
	TO (fit)	9	9	9	10	11	11	11
	TO (performance)	3	3	3	2	**−33**	**−33**	**−33**
	TO (behavior)	4	4	4	0	**−100**	**−100**	**−100**
	TO (abandon)	12	12	12	7	**−42**	**−42**	**−42**
	TO (new job)	7	7	7	16	**129**	**129**	**129**
	Corrective action	46	46	46	48	4	4	4
Organizational citizenship behavior	TO (overall)	43	43	43	37	−14	−14	−14
	TO (attendance)	7	7	7	9	29	29	29
	TO (fit)	9	9	9	7	−22	−22	−22
	TO (performance)	3	3	3	1	**−67**	**−67**	**−67**
	TO (behavior)	4	4	4	3	−25	−25	−25
	TO (abandon)	12	12	12	7	**−42**	**−42**	**−42**
	TO (new job)	7	7	7	10	**43**	**43**	**43**
	Corrective action	46	46	46	51	11	11	11
Counterproductive work behavior	TO (overall)	44	41	42	40	−9	−2	−5
	TO (attendance)	10	9	9	9	−10	0	0
	TO (fit)	9	10	11	10	11	0	−9
	TO (performance)	1	0	0	0	**−100**	0	0
	TO (behavior)	3	1	1	1	**−67**	0	0
	TO (abandon)	8	9	9	9	13	0	0
	TO (new job)	10	10	10	9	−10	−10	−10
	Corrective action	47	39	40	39	−17	0	−3
Safety performance	TO (overall)	43	43	38	44	2	2	16
	TO (attendance)	7	7	10	9	29	29	−10
	TO (fit)	9	9	7	10	11	11	**43**
	TO (performance)	3	3	2	2	**−33**	**−33**	0
	TO (behavior)	4	4	3	0	**−100**	**−100**	**−100**
	TO (abandon)	12	12	6	6%	**−50**	**−50**	0
	TO (new job)	7	7	10	16	**129**	**129**	**60**
	Corrective action	46	46	52	46	0	0	−12
Global performance	TO (overall)	43	43	43	36	−16	−16	−16
	TO (attendance)	7	7	7	7	0	0	0
	TO (fit)	9	9	9	10	11	11	11
	TO (performance)	3	3	3	1	**−67**	**−67**	**−67**
	TO (behavior)	4	4	4	3	−25	−25	−25
	TO (abandon)	12	12	12	8	**−33**	**−33**	**−33**
	TO (new job)	7	7	7	10	0	0	0
	Corrective action	46	46	52	46	−13	−13	−13
Average across predicted criteria	TO (overall)	43.2	42.6	41.8	40.4	−6	−5	−3
	TO (attendance)	7.6	7.4	8.0	8.6	15	17	9
	TO (fit)	9.0	9.2	9.0	9.4	4	2	7
	TO (performance)	2.6	2.4	2.2	1.2	**−60**	**−40**	**−33**
	TO (behavior)	3.8	3.4	3.2	1.4	**−63**	**−50**	**−50**
	TO (abandon)	11.2	11.4	10.2	7.4	**−31**	**−33**	−23
	TO (new job)	7.6	7.6	8.2	11.6	**58**	**58**	**44**
	Corrective action	46.2	44.6	46	44.8	−3	0	−2

Note. Values in bold represent a greater than 30% difference in rate compared to GGUM. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model; TO = turnover.

Second, research should look to see if subclinical psychopathology could explain the curvilinearity in functions of substantive and response forms. If ideal point models truly are more appropriate, they may tap into extreme variants of normal personality. For example, extremely high GGUM trait estimates may be indicative of low standing on an obsessive-compulsive dimension, whereas CTT scores do not reflect this extended part of the conscientiousness dimension. As noted earlier, research on psychopathology has shown some support for the idea that obsessive-compulsive disorder is an extension of normal conscientiousness (e.g., Samuel & Widiger, 2004). Further strengthening the promise of this line of research, some of the strongest curvilinear trends between personality and performance were found using dark-side personality traits (see Benson and Campbell, 2007).

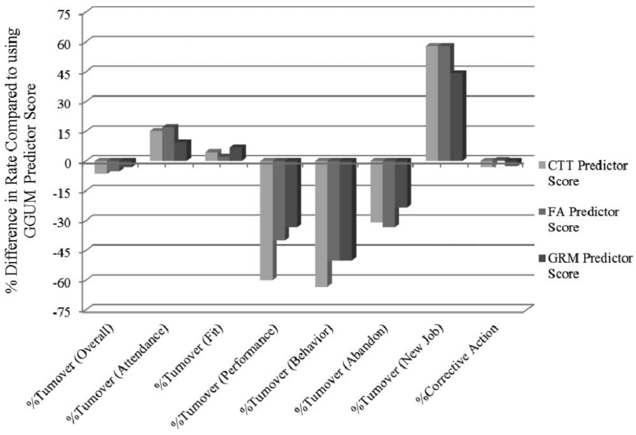

Figure 8. Average percent difference in behavioral outcomes when using the GGUM over CTT, FA, and GRM predictor scores across predicted criteria for Study 3. CTT = classical test theory; FA = factor analytic; GRM = graded response model; GGUM = generalized graded unfolding item response theory model.

Conclusion

We believe these results have important implications for applied personality research as well as practice. Our findings suggest that ideal point modeling of personality measures leads to more consistent substantive findings—with significant curvilinear relationships between conscientiousness and multiple dimensions of job performance observed 100% of the time. Moreover, use of ideal point modeling combined with curvilinear models produces substantially improved selection decisions in comparison to dominance scoring. From a practical standpoint, we suggest that greater accuracy in selection decisions would result from the paired use of ideal point scoring and curvilinear predictive models. We encourage more research that enables a more complete understanding of curvilinearity in the personality–performance link and personality item response process and the interactions among these. To obtain a clearer understanding of the role of personality in work behavior and performance, it is critical that we understand the measurement models and functional forms that define them.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

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

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). The FFM personality dimensions and job performance: Meta-analysis of meta-analyses. *International Journal of Selection and Assessment, 9,* 9–30. doi:10.1111/1468-2389.00160

Benson, M. J., & Campbell, J. P. (2007). To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Selection and Assessment, 15,* 232–249. doi:10.1111/j.1468-2389.2007.00384.x

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92,* 410–424.

Bing, M. N., Stewart, S. M., Davison, H. K., Green, P. D., McIntyre, M. D., & James, L. R. (2007). An integrative typology of personality assessment for aggression: Implications for predicting counterproductive workplace behavior. *Journal of Applied Psychology, 92,* 722–744. doi:10.1037/0021-9010.92.3.722

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's latent ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bowling, N. A., Burns, G. N., Stewart, S. M., & Gruys, M. L. (2011). Conscientiousness and agreeableness as moderators of the relationship between neuroticism and counterproductive work behaviors: A constructive replication. *International Journal of Selection and Assessment, 19,* 320–330. doi:10.1111/j.1468-2389.2011.00561.x

Burns, G. N., & Christiansen, N. D. (2011). Methods of measuring faking behavior. *Human Performance, 24,* 358–372. doi:10.1080/08959285.2011.597473

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work Satisfaction Scale. *Personality and Individual Differences, 49,* 743–748. doi:10.1016/j.paid.2010.06.019

Casillas, A., Robbins, S., McKinniss, T., Postlethwaite, B., & Oh, I. S. (2009). Using narrow facets of an integrity test to predict safety: A test validation study. *International Journal of Selection and Assessment, 17,* 119–125. doi:10.1111/j.1468-2389.2009.00456.x

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. A. (2001). Fitting item response theory model to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523–562.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88–106. doi:10.1037/1040-3590.19.1.88

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96,* 1140–1166. doi:10.1037/a0024004

Cianci, A. M., Klein, H. J., & Seijts, G. H. (2010). The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. *Journal of Applied Psychology, 95,* 618–630. doi:10.1037/a0019130

Clarke, S., & Robertson, I. T. (2005). A meta-analytic review of the Big Five personality factors and accident involvement in occupational and non-occupational settings. *Journal of Occupational and Organizational Psychology, 78,* 355–376.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85,* 678–707. doi:10.1037/0021-9010.85.5.678

Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.

Cucina, J. M., & Vasilopoulos, N. L. (2005). Nonlinear personality–performance relationships and the spurious moderating effects of traitedness. *Journal of Personality, 73,* 227–260. doi:10.1111/j.1467-6494.2004.00309.x

Dalal, D. K., Withrow, S., Gibby, R. E., & Zickar, M. J. (2010). Six questions that practitioners (might) have about ideal point response process items. *Industrial and Organizational Psychology, 3,* 498–501. doi:10.1111/j.1754-9434.2010.01279.x

Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods, 15,* 339–362. doi:10.1177/1094428111430540

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior.

Journal of Applied Psychology, 90, 1241–1255. doi:10.1037/0021-9010.90.6.1241

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42,* 25–36. doi:10.1111/j.1744-6570.1989.tb01549.x

de Vise, D. (2009, July 10). ETS creates rating system for graduate school applicants' personality traits. *The Washington Post*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/07/09/AR2009070902085.html

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3,* 465–476.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement, 19,* 143–166. doi:10.1177/014662169501900203

Dunkley, D. M., Blankstein, K. R., Zuroff, D. C., Lecce, S., & Hui, D. (2006). Self-critical and personal standards factors of perfectionism located within the five-factor model of personality. *Personality and Individual Differences, 40,* 409–420. doi:10.1016/j.paid.2005.07.020

Fayard, J. V., Roberts, B. W., Robins, R. W., & Watson, D. (2012). Uncovering the affective core of conscientiousness: The role of self-conscious emotions. *Journal of Personality, 80,* 1–32. doi:10.1111/j.1467-6494.2011.00720.x

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456. doi:10.1037/0033-2909.116.3.429

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61,* 579–616. doi:10.1111/j.1744-6570.2008.00123.x

Grant, A. M., & Schwartz, B. (2011). Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on Psychological Science, 6,* 61–76. doi:10.1177/1745691610393523

Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.

Hill, R. W., McIntire, K., & Bacharach, V. R. (1997). Perfectionism and the Big Five factors. *Journal of Social Behavior and Personality, 12,* 257–270.

Hoffman, B. J., & Woehr, D. J. (2009). Disentangling the meaning of multisource feedback: An examination of the nomological network surrounding source and dimension factors. *Personnel Psychology, 62,* 735–765. doi:10.1111/j.1744-6570.2009.01156.x

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869–879.

Jackson, J. J., Bogg, T., Walton, K. E., Wood, D., Harms, P. D., Lodi-Smith, J., . . . Roberts, B. W. (2009). Not all conscientiousness scales change alike: A multimethod, multisample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology, 96,* 446–459. doi:10.1037/a0014156

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35,* 1–19.

Johnson, J. W. (2001). Determining the relative importance of predictors in multiple regression: Practical applications of relative weights. In F. Columbus (Ed.), *Advances in psychology research* (Vol. 5, pp. 231–251). Huntington, NY: Nova Science.

Jöreskog, K. G. (1999). *How large can a standardized coefficient be?* [Technical documentation for the LISREL program]. Retrieved from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf

Judge, T. A., & LePine, J. A. (2007). The bright and dark sides of personality: Implications for personnel selection in individual and team contexts. In J. Langan-Fox, C. Cooper, & R. Klimoski (Eds.), *Research companion to the dysfunctional workplace: Management challenges and symptoms* (pp. 332–355). Cheltenham, England: Elgar.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2011). *The performance of RMSEA in models with small degrees of freedom*. Unpublished manuscript, University of Connecticut.

LaHuis, D. M., Martin, N. R., & Avis, J. M. (2005). Investigating nonlinear conscientiousness–job performance relations for clerical employees. *Human Performance, 18,* 199–212. doi:10.1207/s15327043hup1803_1

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96,* 113–133. doi:10.1037/a0021016

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–55.

Meyers, R. D., Dalal, R. S., & Bonaccio, S. (2009). A meta-analytic investigation into the moderating effects of situational strength on the conscientiousness–performance relationship. *Journal of Organizational Behavior, 30,* 1077–1102.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Hillsdale, NJ: Erlbaum.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729. doi:10.1111/j.1744-6570.2007.00089.x

O'Connell, M., Kung, M. C., & Tristan, E. (2011). Beyond impression management: Evaluating three measures of response distortion and their relationship to job performance. *International Journal of Selection and Assessment, 19,* 340–351. doi:10.1111/j.1468-2389.2011.00563.x

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027.

Orvis, K. A., Dudley, N. M., & Cortina, J. M. (2008). Conscientiousness and reactions to psychological contract breach: A longitudinal field study. *Journal of Applied Psychology, 93,* 1183–1193. doi:10.1037/0021-9010.93.5.1183

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance, 24,* 270–290. doi:10.1080/08959285.2011.580808

Pierce, J. R., & Aguinis, H. (2013). The too-much-of-a-good-thing effect in management. *Journal of Management, 39,* 313–338. doi:10.1177/0149206311410060

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135,* 322–338. doi:10.1037/a0014996

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In M. Learly & R. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 369–381). New York, NY: Guilford Press.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32. doi:10.1177/01466216000241001

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 30,* 64–65.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59,* 211–233. doi:10.1177/00131649921969811

Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *International Journal of Selection and Assessment, 7,* 157–169. doi:10.1111/1468-2389.00115

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87,* 66–80.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10,* 117–125.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (Psychometric Monograph No. 17). Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samuel, D. B., Lynam, D. R., Widiger, T. A., & Ball, S. A. (2012). An expert consensus approach to relating the proposed *DSM–5* types and traits. *Personality Disorders, 3,* 1–16. doi:10.1037/a0023787

Samuel, D. B., & Widiger, T. A. (2004). Clinicians' personality descriptions of prototypic personality disorders. *Journal of Personality Disorders, 18,* 286–308. doi:10.1521/pedi.18.3.286.35446

Samuel, D. B., & Widiger, T. A. (2011). Conscientiousness and obsessive-compulsive personality disorder. *Personality Disorders, 2,* 161–174. doi:10.1037/a0021216

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in jobs and training performance. *Personnel Psychology, 61,* 827–868. doi:10.1111/j.1744-6570.2008.00132.x

Stark, S. (2007). *Modfit version 2.0* [Computer software]. Retrieved from http://work.psych.uiuc.edu/irt/tutorial.asp

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2002). *Investigating the effects of local dependence on the accuracy of IRT ability estimation* (AICPA Tech. Rep., Series Two). New Orleans, LA: American Institute of Certified Public Accountants.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91,* 25–39. doi:10.1037/0021-9010.91.1.25

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts' ratings be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods, 14,* 256–278. doi:10.1177/1094428109356712

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model-data fit of ideal point and dominance models. *Applied Psychological Measurement, 35,* 280–295. doi:10.1177/0146621610390674

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94,* 1287–1304. doi:10.1037/a0015899

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60,* 967–993. doi:10.1111/j.1744-6570.2007.00098.x

Thissen, D. (2003). *Multilog 7: Multiple categorical item analysis and test scoring using item response theory* [Computer program]. Chicago, IL: Scientific Software.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554. doi:10.1086/214483

Vasilopoulos, N. L., Cucina, J. M., & Hunter, A. E. (2007). Personality and training proficiency: Issues of bandwidth-fidelity and curvilinearity. *Journal of Occupational and Organizational Psychology, 80,* 109–131. doi:10.1348/096317906X102114

Walker, J. (2012, September 20). Meet the new boss: Big data. *The Wall Street Journal*. Retrieved from http://finance.yahoo.com/news/meet-boss-big-data-000000184.html

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment, 24,* 65–77. doi:10.1027/1015-5759.24.1.65

Whetzel, D. L., McDaniel, M. A., Yost, A. P., & Kim, N. (2010). Linearity of personality–performance relationships: A large-scale examination. *International Journal of Selection and Assessment, 18,* 310–320. doi:10.1111/j.1468-2389.2010.00514.x

Widiger, T. A., Trull, T. J., Clarkin, J. F., Sanderson, C. J., & Costa, P. T., Jr. (2002). A description of the *DSM–IV* personality disorders with the five-factor model of personality. In P. T. Costa Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 89–99). Washington, DC: American Psychological Association. doi:10.1037/10423-006

Zampetakis, L. A. (2010). Unfolding the measurement of the creative personality. *Journal of Creative Behavior, 44,* 105–123. doi:10.1002/j.2162-6057.2010.tb01328.x

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 37–59). New York, NY: Routledge.

Appendix

Construct Validity Evidence for Study 1 and Study 2 Measures

This appendix provides construct validity evidence for the scales used in Studies 1 and 2. Whereas factor structure evidence is cited in text for the conscientiousness measures, factor structure evidence for the performance measures is presented here. Where possible, evidence for external relations for all scales is presented here.

Study 1: Conscientiousness Measure

To determine the construct-related validity of the 10-item conscientiousness measure constructed in Study 1, we estimated relationships between scores on our measure with external variables. On the basis of past research we expected the scores to be positively related to age (e.g., Jackson et al., 2009), academic performance (e.g., Poropat, 2009), and achievement motivation (e.g., Colquitt, LePine, & Noe, 2000). In addition, we expected scores to not differ significantly among racial subgroups (e.g., Foldes, Duehr, & Ones, 2008) or gender (e.g., Feingold, 1994). Results agreed with these expectations. First, CTT, FA, GRM, and GGUM estimates were significantly positively correlated with age (.14, .11, .14, and .10, respectively; all *p*s $< .001$). Second, scores were significantly related to responses to the biodata item asking “What were your grades in high school?” (.24, .26, .24, and .24, respectively; all *p*s $< .001$). Third, the estimates correlated .51, .54, .53, and .48 (all *p*s $< .001$), respectively, with the consulting firm’s 16-item measure of achievement motivation ($\alpha = .82$). The *t* tests conducted for gender differences showed statistically significant differences in conscientiousness, with women showing higher conscientiousness. However, variance explained by gender was always less than 1.8%, showing only partial support for our expectations. No significant differences were found between racial/ethnic groups. Given the above evidence in conjunction with the original, explicit intent of item-writing efforts (i.e., to tap skills involving: behaving with integrity, being dependable, being productive, following rules, valuing quality in one’s work, demonstrating work ethic, being focused on improvement in work, and following rules), we believe that the scale we constructed showed moderate but adequate evidence of construct validity as a measure of conscientiousness.

Study 1: Performance Measures

First, we fit a three-factor CFA to the data wherein individual items were modeled to load onto their specific performance dimension, $\chi^2(41) = 573.60$. Although the RMSEA was high ($RMSEA = .116$, 90% CI [.108, .124]), other fit indices suggested good model–data fit ($SRMSR = .038$; $NNFI = .97$). Moreover, evaluation of alternative models collapsing to two factors showed no improvement in fit, as indicated by comparison of Akaike information criterion. Standardized factor loadings ranged from .69 to .86.

In addition, we compared the factor intercorrelations among TP, OCB, and CWB to values found in past studies. We expected that, if these measures were assessing their intended constructs, the correlation between task performance and OCB would be large, but that the TP–CWB and OCB–CWB correlations would be moderate (e.g., Casillas, Robbins, McKinniss, Postlethwaite, & Oh, 2009; Le et al., 2011). As expected, the task performance and OCB factors were correlated highly at .84 ($p < .001$; Casillas et al. showed $r = .74$; Le et al. showed $r = .80$ and .68), whereas CWB was correlated relatively moderately with both task performance ($r = -.62$; $p < .001$; Casillas et al., 2009, showed $r = -.46$; Le et al., 2011, showed $r = -.63$ and $-.39$) and OCB ($r = -.57$, $p < .001$; Casillas et al., 2009, showed $r = -.54$; Dalal, 2005, showed meta-analytic estimates of $-.32$ with 90% credibility interval ranging from $-.82$ to .24; Le et al., 2009, showed $r = -.62$ and $-.48$).

Finally, correlations between the four conscientiousness estimates (CTT, FA, GRM, and GGUM) with these performance measures were consistent with past reviews. Correlations between conscientiousness and task performance ranged from .09 and .11, consistent with three past meta-analyses that have shown uncorrected mean *r*s of .10, .10, and .13 and unreliability-corrected correlations of .23, .16, and .15 (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Meyers, Dalal, & Bonaccio, 2009). Conscientiousness and OCB relationships ranged from .06 to .09 in this study, which is somewhat lower but consistent with Chiaburu’s (2011) uncorrected mean *r*s of .14 and .18. Correlations between conscientiousness and CWB here ranged from $-.05$ to $-.12$. Although this is on the low side, it is fairly consistent with uncorrected mean $r = -.16$ (Salgado, 2002) and $-.20$ (Berry, Ones, & Sackett, 2007). On the basis of the content of the items and the above analyses, we believe that the measures of TP, OCB, and CWB are successfully measuring their intended constructs.

Study 2: Conscientiousness Measure

We expected, consistent with Study 1 and past research (e.g., Colquitt et al., 2000), that our CTT, FA, GRM, and GGUM conscientiousness scores would correlate highly with this firm’s seven-item measure of achievement motivation ($\alpha = .72$); as expected, the scores correlated significantly ($r = .44$, $r = .46$, $r = .46$, and $r = .45$, respectively; $p < .001$ for all). Second, if our post hoc scale was measuring conscientiousness, we expected the CTT, FA, GRM, and GGUM scores to correlate significantly with the firm’s in house, seven-item self-report measure of workplace safety behavior ($\alpha = .81$; e.g., Clarke & Robertson, 2005).[A1] As expected, the CTT, FA, GRM, and GGUM scores correlated strongly ($r = .48$, $r = .45$, $r = .49$, $r = .47$, respectively; $p < .001$ for all) and consistent with Clarke and Robertson’s (2005) meta-analytically corrected $\rho = -.30$ between conscientiousness and accident involvement. Finally, significance tests showed no differences with regard to gender (e.g., Feingold, 1994) or race (e.g., Foldes et al., 2008). Taken together, we believe, these findings provide adequate evidence for the construct validity of the constructed conscientiousness measure.

[A1] Note that this is a self-report measure distinct from the supervisor ratings of safety behavior used as a criterion in the main analyses.

(*Appendix continues*)

Study 2: Performance Measures

To evaluate the construct validity of our constructed performance measures of TP and OCB, as well as the consulting firm's measures of CWB, safety performance, and global performance, we estimated a three-factor model with 47 supervisor ratings as indicators of TP (16 items), OCB (7 items), and CWB (24 items). Due to problems with multicollinearity in the indicators, however, the solution showed problems associated with overfitting (i.e., a considerable amount of covariance could not be explained by the latent variables). Thus, we used a strategy similar to the one suggested by Hoffman and Woehr (2009) wherein items were combined into parcel indicators. We created three parcels for each latent factor (i.e., nine parcels total).

The model specified with item parcels was more reasonable. Indeed, whereas some fit indices suggested less than adequate fit (e.g., $\chi^2(24) = 563.94$, $p < .001$, $RMSEA = .125$, 90% CI [.116, .134]), other fit indices suggested good fit (e.g., $NNFI = .94$, $SRMSR = .056$). We note that, although the RMSEA value was above convention, Kenny, Kaniskan, and McCoach (2011) have shown that RMSEA is not an appropriate fit index for relatively low-*df* models. In addition to the good fit suggested by the other indices, the average standardized factor loadings were .83 ($SD = .14$), .82 ($SD = .12$), and .84 ($SD = .04$) for TP, OCB, and CWB, respectively. Finally, interfactor correlations were in line with findings here and previous research outlined above (e.g., Casillas et al., 2009; Dalal, 2005; Le et al., 2011) in that TP and OCB were highly correlated ($r = .63$), relative to the relationship between OCB and CWB ($r = -.56$), and larger still compared to the relationship between task performance and CWB ($r = -.38$; all values significant $p < .001$).

We also sought to provide evidence of the construct validity for the global performance measure. To do this, we regressed global performance scores onto TP, OCB, and CWB scores (all were significant) and conducted relative weight analyses (Johnson, 2000) using the RWEIGHT program (Johnson, 2001). In their policy-capturing study of global performance judgments, Rotundo and Sackett (2002) showed that TP was weighted the most, followed by CWB, and finally OCB. Therefore, if the global performance scores are indeed measuring overall performance, we should expect the highest relative weight attributable to TP, followed by CWB, and finally OCB. Results confirmed expectations in that of the total 31.7% of variance in global performance explained, TP accounted for the highest proportion (59.4%), followed by CWB (38.4%) and then OCB (5.8%).

Finally, we looked to establish the construct validity of the firm's in-house safety performance ratings. We expected to see positive correlations between safety performance ratings and measures of responsibility and safety orientation. Results only somewhat confirmed these expectations as safety performance ratings showed significant but low correlations with scores on the firm's 12-item measure of responsibility ($\alpha = .85$, $r = .06$, $p = .039$) and the firm's 12-item self-report measure of safety orientation ($\alpha = .85$, $r = .07$, $p = .012$). Although little support for construct validity evidence could be surmised from the data at hand, we decided to consider these ratings for purposes of empirical replication of effects. In sum, we believe the factor analytic and correlational evidence provided above generally suggests adequate construct validity for the criterion variables used in Study 2, though conclusions regarded safety performance should be somewhat more conservative.

Received October 12, 2012
Revision received September 4, 2013
Accepted September 10, 2013 ■