The Relationship of Reliability and Validity of Personality Tests to Frame-of-Reference Instructions and Within-Person Inconsistency

Craig M. Reddock*, Michael D. Biderman** and Nhung T. Nguyen***

*Department of Psychology, Old Dominion University, Norfolk, VA 23529-0267, USA. credd005@odu.edu
**University of Tennessee, Chattanooga, TN
***Towson University, Towson, MD

The efficacy of both frame-of-reference (FOR) instructions and a measure of within-person inconsistency in predicting grade point average was investigated. The IPIP Big Five personality questionnaire was given to 329 students with generic instructions and 'at school' FOR instructions. The Wonderlic Personnel Test was also administered. A measure of within-person inconsistency was created based on the standard deviations of responses to items within the same Big Five dimension. The validity of conscientiousness was greater when FOR instructions were given. The measure of within-person inconsistency provided incremental validity over that of conscientiousness and cognitive ability. Additionally, within-person inconsistency moderated the relationship between conscientiousness and performance for the participants without the FOR instructions. Practical implications are discussed.

1. Introduction

In recent years, there has been considerable research on how personality can be measured and used to predict performance, focusing recently on the Big Five personality factors (Dudley, Orvis, Lebiecki, & Cortina, 2006; Murphy & Dzieweczynski, 2005; Ones, Dilchert, Viswesvaran, & Judge, 2007; Smith, Hanges & Dickson, 2001). Whereas these studies have done much to illuminate the value of personality tests as predictors of performance, as recently as 2007 Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007) cautioned that the issue is the very low validity of personality tests for predicting job performance. This caution is not without merit. While Ones et al. (2007) assert and provide evidence to show that personality measures can be robust predictors of performance, Morgeson et al. (2007) counter with several meta-analytic estimates of personality as a predictor of job performance that are quite low. Given the volume of research on personality and the perception held by some that these tests are limited in their usefulness (Guion & Gottier, 1965), there is certainly a need for ways to improve the validity of such tests.

Very recently, two factors that might enhance the validity of personality assessment have been investigated.

The first is the frame-of-reference (FOR) provided by the instructions given with personality questionnaires (Schmit, Ryan, Stierwalt, & Powell, 1995). The second is intraindividual or within-person variability (Edwards & Woehr, 2007). The purpose of the present study was to investigate the relationship between these two factors and to extend previous research on the relationship of each to the reliability and validity of personality tests.

1.1. FOR Instructions

Several recent studies have focused on the impact of including statements providing an explicit FOR to personality measures (Bing, Whanger, Davison, & VanHook, 2004; Holtz, Ployhart, & Dominguez, 2005; Hunthausen, Truxillo, Bauer, & Hammer, 2003; Lievens, De Corte, & Schollaert, 2008; Schmit et al., 1995). FOR instructions are context-specific tags that are added to personality questionnaires, either as a statement or paragraph presented with the personality items or as a modification of each item. For example, changing 'I am always prepared.' to 'I am always prepared at school' creates an 'at school' FOR.

It has been proposed that under generic instructions different respondents may use a different FOR when

responding to questionnaire items (Schmit *et al.*, 1995). If the FOR held by a respondent, even though consistently applied, is not related to the context in which the criterion is measured, it is possible that individual differences in the personality dimension in the inappropriate context will not correlate highly with individual differences in the criterion context, reducing validity.

Studies manipulating FOR instructions have yielded generally promising results regarding their use to increase the validity of personality measures. Schmit *et al.*'s (1995) nascent examination of FOR instructions showed that scores of participants instructed to answer conscientiousness items with an 'at school' FOR had higher criterion-related validity than scores of participants given generic conscientiousness items. Hunthausen *et al.* (2003), in a validation study of customer service managers, found that FOR instructions moderated the validity of the extraversion and openness scales in predicting supervisory ratings of job performance after controlling for cognitive ability.

Bing *et al.* (2004) replicated and extended the Schmit *et al.* (1995) study and found the validity of conscientiousness for predicting grade point average (GPA) in a condition in which participants responded to items with an 'at school' FOR to be larger than when the same participants responded to generic conscientiousness items. Lievens *et al.* (2008) found higher criterion-related validity in predicting self-reported GPA associated with 'at school' FOR items compared with both the generic items and 'at work' FOR conscientiousness items. That the validity of the 'at school' items was higher than the 'at work' items is important because it suggests that the validity increases noted in previous studies is due to the provision of a criterion-relevant FOR rather than to simply providing any specific FOR, whether correct or not.

The FOR effect has been observed in the relationships of personality to other variables important in organizational contexts. Heller, Ferris, Brown, and Watson (2009) found that work personality was a better predictor of job satisfaction than global personality or home personality. Bowling and Burns (2010) found that work-specific personality yielded stronger relationships with job satisfaction, work frustration, turnover intention, and absenteeism than did general personality.

In addition to increasing the likelihood that all respondents employ the appropriate FOR, Lievens *et al.* (2008) proposed that using FOR instructions would have the added effect of reducing within-person variability of responses. They pointed out that reducing such inconsistency would affect the reliability and thus the validity of those measures. As a demonstration of the effect of within-person differences in FOR, they had participants complete two questionnaires, one with an 'at work' and the other with an 'at school' FOR. They then mixed the items from the two questionnaires to form 10-item pseudo-questionnaires as if participants had filled out some of the items in the pseudo-questionnaire with 'at work' instructions and other items with 'at school' instructions. They found that the reliability of a 10-item mixture was highest when the proportion of similar items was the greatest (either zero or 100%), and that the validity for predicting GPA was the highest when all the items were 'at school'. This powerful result suggests that within-person inconsistency when using FOR instructions can affect validity through reliability and also that an inappropriate FOR can affect validity directly. Interestingly, however, Lievens *et al.* (2008) did not find consistent reliability differences between groups of participants who completed the generic instruction questionnaires and those who filling out the same questionnaires with either an 'at school' or 'at work' FOR. Bowling and Burns (2010) replicated this finding when they reported that work-specific and general personality measures had similar reliabilities.

1.2. Within-person inconsistency

Although the Lievens *et al.*'s (2008) article was the first to propose that FOR instructions might affect within-person inconsistency, such inconsistency has itself been the focus of research in other contexts. In such research, intraindividual variability has been considered within the context of metatraits and traitedness (Britt, 1993; Dwight, Wolf, & Golden, 2002), extreme response styles (Greenleaf, 1992), and the stability of affect across time (e.g., Eid & Diener, 1999; Kernis, 2005). In the traitedness research, the term metatrait is used to represent the quality of possessing versus not possessing a particular trait operationalized as the interitem variability on a personality measure. Persons exhibiting low interitem variability on a particular personality trait are traited while those exhibiting higher variability are untraited (Britt, 1993). Britt found higher correlations between pairs of constructs on which respondents had smaller interitem variability than between constructs for which respondents exhibited large interitem variability on one or both constructs. Dwight *et al.* (2002) also found higher predictor–criterion relationships among persons with smaller interitem variability on the predictor.

In an early study of within-person variability, Fiske and Rice (1955) suggested that such variability might be viewed as a separate personality characteristic or factor 'analogous to well-known factors of level scores in mental abilities, interests, and personality' (Fiske & Rice (1955, p. 217). Research has supported such a view. Fleeson (2001), using experience sampling techniques with repeated administrations of the same questionnaire items, provided evidence that variability was a stable individual difference trait across time. Baird, Kimdy, and Lucas (2006) measured variability as standard deviations of responses across roles controlling for mean level

differences and concluded that 'intraindividual personality variability is a broad, global trait.' (p.525). Importantly for the present study, Baird *et al.* using Big Five scales found that individuals who were variable on one trait also tended to be variable on the others. This result is in contrast to the suggestion from the literature on meta-traits that traitedness is a dimension specific phenomenon. Fleisher and Woehr (2008) provided evidence suggesting that consistency across the Big Five personality dimensions is a unidimensional construct and that self-reported consistency moderated personality-performance relationships, with less variability being associated with stronger relationships. These results suggest that there may be a global inconsistency construct, distinct from the dimension-related construct as conceptualized by traitedness literature.

Longitudinal research on within-person variability has provided evidence for the temporal consistency of intraindividual variability. For example, Eid and Diener (1999) measured variability in affect across 51 days. Fleeson (2001) computed the standard deviations of scale scores across time periods over a several-day span. Baird *et al.* (2006) based their measurement of variability on standard deviations of responses to questionnaire items administered six times, each one assuming a different role. Edwards and Woehr (2007) and Fleisher and Woehr (2008) used a response format that required respondents to categorize how each item was descriptive of their behavior, indicating the percentage of time the item was either very inaccurate, neither inaccurate nor accurate, or very accurate. They used the percentages to compute both an estimate of the level of descriptive accuracy for each item and an estimate of variability of the extent to which the item was descriptive. The results of these studies suggested that in addition to generalizing across personality traits at a particular time period, inconsistency is a stable characteristic.

The studies above demonstrated that respondent inconsistency on one or two traits is associated with the strength of relationships involving those traits. However, further research is needed on the effect of such inconsistency defined as a general characteristic, as a characteristic that cuts across dimensions. In many testing situations, questionnaires comprised of multiple individual traits, for example, the Big Five, are administered. Thus, a general measure of inconsistency based on interitem variability of responses within dimensions is readily available whenever such questionnaires are administered. For this reason, if inconsistency is a general characteristic, it will be advantageous to take advantage of the information in multitrait questionnaires to measure that inconsistency. In addition to the investigation of the effects of using FOR instructions on validity, the second purpose of the present research was to investigate the reliability and validity of an inconsistency measure based on interitem variability within multiple personality dimensions.

The measure used here reflected variability occurring within a period of < 1 hr, often less than one-half hour, a much narrower time period than other investigators have used. We conceptualize such variability as being composed of two components, both of which have been proposed as factors affecting the reliability of psychological tests (Schmidt, Le, & Ilies, 2003). First, it involves what Schmidt *et al.* (2003) called random response error – error 'specific to a moment when subjects respond to an item of a measure' (Schmidt *et al.*, 2003, p. 208). Even though a respondent's position on whatever dimension a group of items represent remains unchanged, his or her choice of a response to indicate that position may vary from item to item because of fluctuations in mood, cognitive state, and external stimulation, for example. Second, it involves what Schmidt and colleagues called specific factor error, which at the item level is 'produced by respondent-specific interpretation of the wording of questionnaire items' (Schmidt *et al.*, 2003, p. 209). Respondents may interpret the correspondence of the content of an item to the dimension it represents differently from item to item, yielding differences in responses to items from the same dimension. Clearly one possible source of such inconsistency could be item to item variation in the FOR adopted by the respondent as suggested by Lievens *et al.* (2008). Although it may be possible to separate the two sources of inconsistency through studies designed differently than the present study, the purpose of this investigation was to explore the extent to which measures that could be computed from a typical single administration of a personality questionnaire were related to the validity of personality measures. For that reason, we focused on determining if individual differences in a general measure of inconsistency from a single administration of a personality questionnaire were reliable and useful in a selection situation. Based on the research discussed above, our belief was that if there were individual differences in inconsistency as measured within a single administration of a personality questionnaire (i.e., if some persons generally exhibit larger errors than others), then such differences would likely be related to scale reliability estimates and might be related, either through reliability or independently of it, to scale validity.

In this study, we focused on the use of conscientiousness as a predictor of academic performance as measured by GPA. Even though conscientiousness was the only Big Five dimension whose scale scores were treated as a predictor, we operationalized within-person inconsistency as the average variability within all dimensions of the Big Five, not just conscientiousness. We computed the standard deviation of item responses for each dimension within a single administration of a Big Five questionnaire and then computed the mean of those

standard deviations. That measure formed the basis for the operational definition of within-person inconsistency.

Based on our review of the literature on FOR effects and within-person inconsistency, we developed the following hypotheses related to the prediction of undergraduate academic performance by conscientiousness.

Virtually, all of the studies comparing the validity of personality tests administered with FOR instructions with those administered with generic instructions have shown stronger relationships with appropriate FOR instructions. Thus, we propose the following hypothesis:

H1: The validity of conscientiousness administered with an 'at school' FOR will be larger than the validity of conscientiousness administered using generic instructions with no context.

The Lievens *et al.*'s (2008) suggestion that FOR instructions would affect within-person inconsistency leads to the following hypothesis:

H2: The reliability of the Big Five scales administered with an 'at school' FOR will be larger than the reliability of the same scales administered using generic instructions with no context.

Based on the results reported above, it appears that response inconsistency over time is a substantive characteristic of respondent behavior. For our measure to be psychometrically appropriate, our measurement of inconsistency must exhibit acceptable indications of reliability. This leads to the following hypothesis:

H3: The within-person standard deviations for the Big Five dimensions will be positively correlated and the overall measure of within-person inconsistency will exhibit an acceptable level of reliability.

Because coefficient α, the most frequently used estimate of reliability, is based on the covariances of items, it is expected that for a given scale, smaller differences between item responses will be associated with larger covariances of those items and thus larger reliability estimates. This leads to the expectation that high variability of responses between items within the same dimension will be associated with low reliability of scale scores based on those responses. Although the inverse relationship of reliability to within-person variability within a single scale is essentially a mathematical tautology based on the definition of reliability, it is not necessarily the case that reliability will be inversely related to a measure based on the average of within-person variability across multiple dimensions, although we expect that it will. Thus we hypothesize the following:

H4: The reliability of the Big Five scales will be larger for those respondents with smaller average within-person variability estimates.

Although there is little evidence regarding the nomological net surrounding the measure of inconsistency

used in this study, Biderman (2007) found that a latent variable of inconsistency whose indicators were standard deviations of item responses was significantly and inversely related to cognitive ability as measured by the Wonderlic Personnel Test (WPT: Wonderlic, 1999). Because the predictive validity of cognitive ability measured as standardized test scores such as SAT or ACT for academic performance has been well established (e.g., Bridgeman, McCamley-Jenkins, & Ervin, 2000; Halpin, Halpin, & Schaer, 1981), it can be argued that a measure related to cognitive ability should also be related to academic performance. Thus, we hypothesize:

H5a: Within-person variability will be negatively related to academic performance.

Because cognitive ability and conscientiousness are well-known to be essentially orthogonal (e.g., Goff & Ackerman, 1992), colinearity between those two predictors will be essentially zero. Thus, we expect that any relationship of inconsistency to academic performance will remain significant after controlling for conscientiousness. Therefore we propose:

H5b: Within-person variability will be significantly and negatively related to academic performance after controlling for conscientiousness.

Finally, if there are individual differences in inconsistency, upholding H3 above, and if reliability is related to these individual differences, upholding H4, then we will expect the differences in reliability to have an effect on validity. Specifically, we expect the validity of conscientiousness for those persons with the greatest inconsistency, thereby having the lowest reliability, to be the smallest. This leads to the final hypothesis:

H6: Within-person variability will moderate the validity of conscientiousness, such that the conscientiousness-academic performance relationship will be weaker for respondents who exhibit greater within-person variability.

3. Method

3.1. Participants

Participants were 329 undergraduate students, 120 from a Mid-Atlantic University and 209 from a Southeastern University. Demographic breakdown of the total sample was 40% male, 73% White, 18% Black, and 9% other. The average age of participants was 21.0 with standard deviation of 4.0.

3.2. Measures

3.2.1. Big Five
The sample 50-item scale from the International Personality Item Pool web site was used for both the Generic

and FOR conditions. For the Generic condition, the items were used without modification, with the exception that each item was phrased as a sentence. For example, the first item on the web site is a positively worded item indicating Extraversion. On the web site it is listed as 'Am always prepared.' For the questionnaire used in this study, the item was changed to 'I am always prepared.' For the FOR version of the questionnaire, the phrase 'at school' was added to each item, either at the beginning of the item, for example, 'At school, I am always prepared' or at the end of the item, for example, 'I feel little concern for others when at school.' Participants were asked to indicate how accurately each statement described them on a 1, labeled 'Completely Inaccurate' to 7, labeled 'Completely Accurate' scale. We used a 7-point scale based on Finney and DiStefano's (2006) recommendation that a 7-point scale would make the individual responses correspond more nearly to a continuous distribution than would a 5-point scale. Moreover, we hoped that the use of a 7-point scale would weaken the relationship between within-person variability and central tendency discussed by Baird et al. (2006).

3.2.2. Within-person inconsistency

We operationalized this construct as the extent to which individual responses to personality items vary within a dimension across items designed to measure that dimension. The measure was based on the standard deviations of responses to Big Five items. Specifically, five standard deviations were computed for each participant – one from each of the Big Five dimensions – and their mean was used as the measure of inconsistency. In the following, inconsistency estimates are labeled *V*, for variability. Since no attempt was made to adjust this estimate it is labeled Unadjusted *V*. Two values of Unadjusted *V* were obtained for each respondent – one from the responses in the Generic condition and one from the responses in the FOR condition.

3.2.3. Adjusted within-person inconsistency

As previously noted, Baird et al. (2006) found that the nature of the relationships with inconsistency changed considerably after correcting for mean level differences. They provided convincing evidence that standard deviation based measures of inconsistency can strongly depend on the mean level of responses on a trait. For example, on a traditional Likert response scale with five categories, if the mean of a group of respondents is much larger than the midpoint of the scale, ceiling effects will force the correlation of within-person standard deviations with mean level to be negative, contaminating the estimates of the relationships of the within-person standard deviation to other measures. To control for such confounding, the present study used a 7-point response scale to effectively move the floor and ceiling of the scale away from the mean level. We also computed an adjusted

V measure that controlled for mean level in a manner suggested by Baird et al. (2006). First, the standard deviation of item responses within a dimension was regressed onto the 10-item scores representing the dimension and the squares of those item scores. The mean of the standard deviations of the residuals from these five regressions from each condition was used as the adjusted within-person inconsistency value for that condition. Two such variables were computed, one for the Generic condition and one for the FOR condition. They are labeled Adjusted *V* in what follows.

3.2.4. Latent inconsistency variable

As a second way of computing an adjusted measure of inconsistency, a model of inconsistency in which standard deviations of items within each scale served as indicators of a single latent inconsistency variable was developed in a manner suggested by Biderman (2007). The model is presented in Figure 1. In this model, the standard deviations that served as indicators of the latent inconsistency variable were regressed onto scale scores to remove the dependence on scale level. Factor scores of the latent inconsistency variable were computed using the regression approach, and those factor scores were analyzed in the same way that the within-person inconsistency measures defined above were analyzed. In the following, the factor score measures from each condition were labeled Latent *V*.

3.2.5. Cognitive ability

The WPT was administered to all participants before the personality questionnaires as a control measure. This test is a 12-min timed test, with each item increasing in difficulty compared with the one immediately preceding it.

3.2.6. Other measures

Three other measures were administered between the Generic and FOR versions of the Big Five questionnaires to reduce the memory effect as a threat to internal validity in within-subjects studies (Schwab, 2005). They were the BIDR Impression Management and Self Deception scales (Paulhus, 1984) and a situational judgment test of integrity (Becker, 2005). As these measures were used simply as filler tasks, they were not analyzed for this article.

3.2.7. Academic performance

Cumulative GPA was the criterion in the analyses that follow. With permission from students, GPAs were obtained from university records at the end of the semester in which participation occurred.
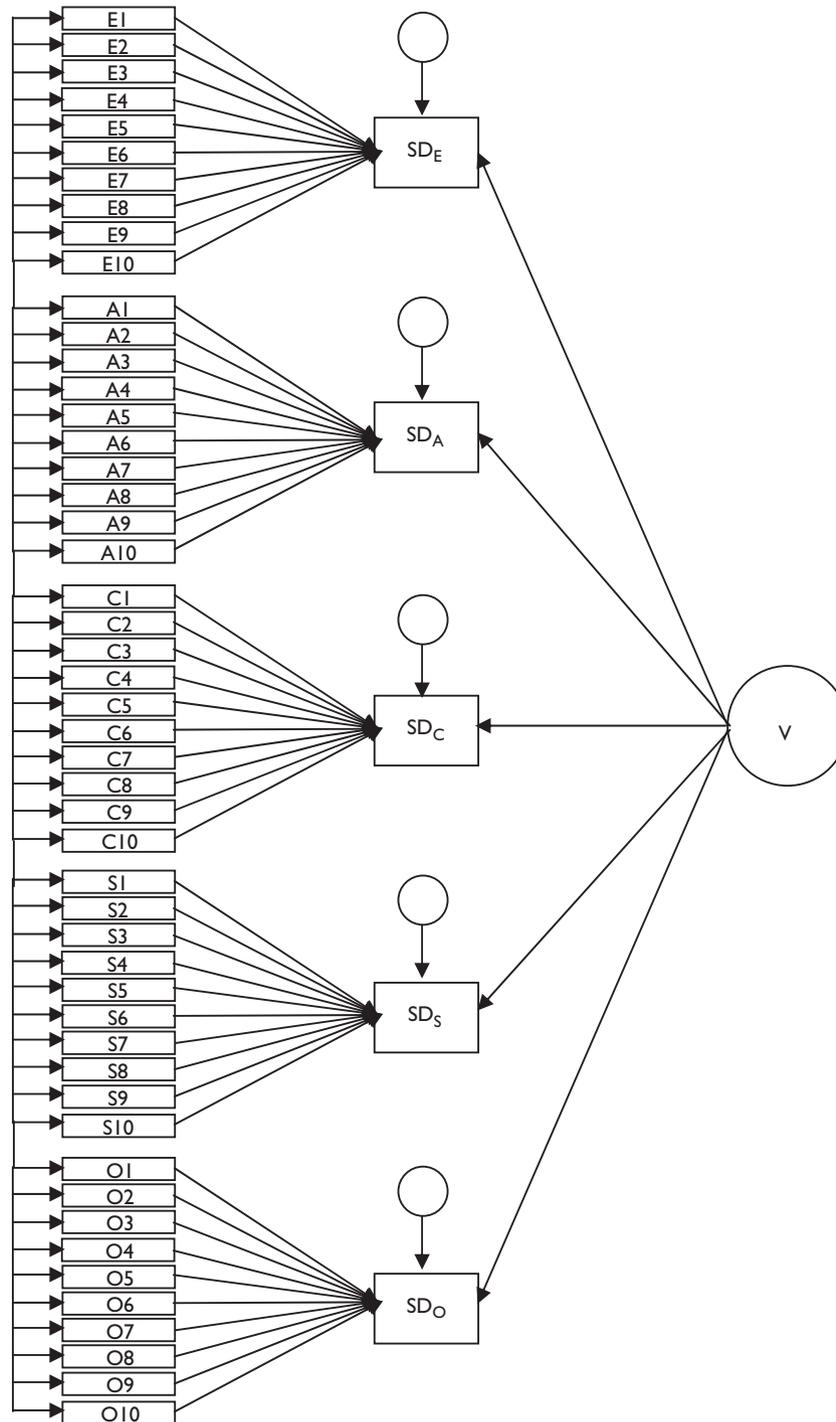
Figure 1. Confirmatory factor analysis model of within-dimension standard deviations.

3.3. Design and procedure

In this research, we compared a generic instruction condition with a FOR instruction condition using a within-subjects design. Participants were first given the WPT. After the WPT, participants were given a questionnaire packet that included all the scales described above with either the Generic instruction Big Five questionnaire preceding the FOR or vice versa. Measures were administered to participants in groups of 2–20 persons. Fifty-one percent received the Generic version of the Big Five scale before the FOR version.

4. Results

Application of the model in Figure 1 yielded χ^2 values of 223.009 and 261.097 ($p < .001$ for both). CFI values were

.940 and .897 and RMSEA values were .037 (90% CI: .025–.047) and .046 (90% CI: .036–.055) for the Generic and FOR conditions, respectively. Mean doubly standardized loadings of the standard deviation indicators on the variability latent variable were all larger than .5, and the factor determinacy values were .918 and .893 for the Generic and FOR condition latent variables, respectively. These results all supported the conclusion that the model of Figure 1 fit the data acceptably and that the latent inconsistency variable was an appropriate representation of within-person variability for each condition. For the analyses that follow, factor scores for the latent variable in each condition were computed using the regression procedure (Muthén, 1998–2004) and added to the data set containing the remaining data for the analyses that follow.

Table 1 presents means, standard deviations, reliabilities, and correlations of the Big Five scales from the Generic and FOR conditions, the WPT, the three inconsistency measures from each condition, and GPA. Correlations between the Big Five scale scores in both conditions were generally positive. Correlations of the measures of inconsistency with personality scale scores were small and not consistently positive or negative. Reliabilities of all variables, including the inconsistency variables, were larger than .70.[1]

The three measures of inconsistency – the unadjusted mean of standard deviations, the adjusted measure based on the mean of residuals, and the latent variable factor scores were very highly correlated, with all correlations between measures within a condition larger than .94 and all correlations between measures across conditions larger than .70. For this reason, although the analyses below were presented separately for all three variables, the focus of the descriptions of those analyses was on the unadjusted measures.[2]

Hypothesis 1 stated that the validity of conscientiousness would be larger when a FOR related to the situation in which the criterion was obtained was provided in the items. Using a test of significance of the difference between dependent correlations (Meng, Rosenthal, & Rubin, 1992) the validity from the FOR Condition of .27 was larger than the validity of .20 from the Generic condition ($p < .05$, one-tailed). Hypothesis 1 was thus supported.

Hypothesis 2 stated that reliability of Big Five scales would be greater in the FOR condition than in the Generic condition. Corresponding reliabilities for each condition from Table 1 were compared using a test presented by Kim and Feldt (2008). The reliabilities of extraversion and agreeableness were significantly higher in the FOR condition. However, none of the other differences were significant, with reliabilities of conscientiousness and stability larger in the Generic conditions and that of openness larger in the FOR condition. These results provided mixed support for H2.

Hypothesis 3 stated that the measures of inconsistency would exhibit acceptable levels of reliability. The diagonal entries in Table 1 showed that all three measures of inconsistency within both the Generic and FOR conditions had reliability estimates larger than .7, with those of the latent variable-based factor scores $> .8$. These estimates reflect the positive correlations between within-person standard deviations for the five dimensions and suggest that there is a general characteristic of inconsistency underlying all of them. This argument is also supported by the positive loadings of the standard deviations on the latent inconsistency variable in the model of Figure 1. All of these results support H3.

The fourth hypothesis was that generally inconsistent respondents as indicated by the inconsistency measures would exhibit lower scale reliabilities than more consistent respondents. To test this hypothesis, median splits on each *V* measure within each condition were performed, forming two groups – a high inconsistency group and a low inconsistency group for the Generic condition and the FOR condition. The reliabilities of the Big Five scale scores were then computed for each inconsistency group within each condition. These are presented in Table 2. In each comparison, the scale reliability computed from the more inconsistent group was smaller than that computed from the more consistent group. All differences were significant at $p < .05$, indicating that the inconsistency of responding represented by the *V* variable was reflected in the familiar coefficient α. These results support Hypothesis 4.

Hypotheses 5A and 5B stated that within-person variability would be negatively related to GPA both simply and incrementally over conscientiousness. Hypothesis 5A was tested by computing the simple correlation coefficients of GPA with the inconsistency measure from each condition. Incremental validity over conscientiousness in each condition (Hypothesis 5B) was assessed in two-predictor regressions in which GPA was the dependent variable. Table 3 presents the results of these analyses. As shown in the table, the inconsistency measure from each condition was a valid predictor of GPA by itself ($p < .05$ for both conditions). Moreover, the *V* variable within each condition contributed incremental validity over conscientiousness scale scores when used in two-predictor equations ($p < .01$ for each condition). In all cases, the relationship of GPA to the *V* variable was negative, indicating that persons who were more inconsistent had lower GPAs. Both hypotheses were supported.

Hypothesis 6 stated that within-person variability would moderate the relationship of GPA to conscientiousness. To test it, moderated regression analyses were conducted in which GPA was regressed onto conscientiousness, within-person variability, and their product. The test of the product variable in this regression provided the test of the moderation hypothesis. The results are presented in Table 4. As shown in the table,

Table 1. Means, standard deviations, and correlations of Big Five scale scores, cognitive ability, inconsistency scores, and end-of-semester GPA from the FOR and generic conditions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Generic extraversion	88																	
2. Generic agreeableness	28	76																
3. Generic conscientiousness	04	19	80															
4. Generic stability	25	04	10	86														
5. Generic openness	18	17	10	20	79													
6. FOR extraversion	78	25	−01	29	19	90												
7. FOR agreeableness	29	70	15	09	06	45	82											
8. FOR conscientiousness	−04	24	70	00	13	−04	19	79										
9. FOR stability	18	06	11	75	25	25	03	15	84									
10. FOR openness	20	14	13	20	84	27	13	25	23	80								
11. Cognitive ability	−03	−03	−14	16	24	−00	−06	−01	14	22	81							
12. Gen V	−11	−02	10	−17	07	10	−03	20	−11	03	−25	77						
13. Gen adjusted V	−01	10	18	−11	12	−02	06	26	−06	08	−24	99	81					
14. Gen latent V	−03	07	15	−15	14	−04	03	21	−09	10	−27	96	96	92				
15. FOR V	−09	02	18	−07	09	−09	−07	24	−01	08	−19	73	73	70	72			
16. FOR adjusted V	−03	11	22	−04	13	−02	06	29	02	14	−18	73	75	71	99	77		
17. FOR latent V	02	09	25	−06	17	03	04	29	00	19	−18	71	73	73	95	96	89	
18. End-of-semester GPA	−05	11	20	−04	07	−04	06	27	03	10	32	−16	−14	−18	−16	−14	−16	NA
Mean	4.46	5.20	4.84	4.40	4.80	4.15	4.81	5.03	4.44	4.65	2.24	1.12	0.00	0.00	0.00	0.00	0.00	2.97
Standard deviation	1.00	0.73	0.82	0.95	0.74	1.09	0.84	0.78	0.93	0.78	6.00	0.32	0.33	0.22	0.31	0.32	0.21	0.60

Note. $|r| \geq .11$ is significantly different from 0. Reliabilities are on the diagonal. Decimal points were omitted from correlations to increase clarity. Cognitive ability means, standard deviation, and reliability were estimated from WPT responses of 310 participants. $N = 329$ for all other measures. Reliability estimates for Latent V values are factor determinacies.

Table 2. Reliability coefficients for consistent respondents and inconsistent respondents as represented by the *V* measure of intra-individual variability

	Big Five dimension									
	Generic condition					FOR condition				
	E	A	C	S	O	E	A	C	S	O
Median splits on unadjusted *V*										
Consistent	.92	.84	.85	.90	.86	.93	.86	.83	.89	.87
Inconsistent	.84	.71	.76	.83	.74	.87	.80	.71	.80	.73
p <	.001	.001	.010	.001	.001	.001	.050	.010	.001	.001
Median splits on adjusted *V*										
Consistent	.92	.82	.86	.89	.83	.93	.87	.85	.87	.85
Inconsistent	.84	.73	.76	.83	.76	.87	.80	.71	.82	.76
p <	.000	.003	.000	.002	.009	.000	.004	.000	.013	.002
Median splits on latent *V*										
Consistent	.92	.82	.85	.90	.83	.93	.87	.83	.87	.85
Inconsistent	.85	.73	.76	.83	.75	.88	.80	.74	.82	.74
p <	.000	.004	.001	.001	.010	.000	.003	.003	.010	.000

Table 3. Simple and incremental validity of *V*, adjusted *V*, and factor scores of the latent inconsistency variable of Figure 1 in Generic and FOR conditions

Inconsistency measure			
Latent *V*			
Variable	Unadjusted *V*	Adjusted *V*	Factor scores
Generic condition			
Simple regression			
V	−.16[b]	−.14[a]	−.18[b]
Multiple regression			
Conscientiousness	.22[c]	.23[c]	.23[c]
V	−.18[b]	−.18[b]	−.21[c]
Multiple *R*	.27[c]	.26[c]	.29[c]
FOR Condition			
Simple regression			
V	−.16[b]	−.14[a]	−.16[b]
Multiple regression			
Conscientiousness	.32[c]	.34[c]	.34[c]
V	−.24[c]	−.24[c]	−.26[c]
Multiple *R*	.35[c]	.35[c]	.36[c]

Note. Regression coefficients are standardized coefficients. [a]$p < .05$. [b]$p < .01$. [c]$p < .001$.

the hypothesis was supported for the Generic Condition data ($p < .05$ for all tests). Moreover, the sign of the product term was as hypothesized for the FOR condition, although the size of the product term coefficient did not reach traditional levels of statistical significance. To elucidate the moderation, the validity of conscientiousness for the groups formed by median splits on Unadjusted *V* used in the analyses testing H4 was obtained. For the Generic condition, validity was .10 for the inconsistent responders and .37 for the consistent responders. For the FOR condition, validity was .23 for the inconsistent responders and .35 for the consistent re-

sponders. Thus the differences in validity ranged from .12 to .27 between the two types of responders, depending on the instructional condition.

Although no hypotheses were presented for incremental validity of inconsistency over that afforded by cognitive ability, exploratory analyses were conducted in which the increase in validity associated with adding the conscientiousness scale score and the measure of inconsistency to WPT scores was assessed. The results are presented in Table 5. Inspection of Table 5 shows that each variable – cognitive ability, conscientiousness, and *V* – added significantly to validity controlling for the other

Table 4. Moderated regression testing the hypothesis that within-person variability moderates the conscientiousness-GPA relationship

Inconsistency measure

Latent variable

Variable	Unadjusted V	Adjusted V	Factor scores
Generic condition			
Conscientiousness	.66[b]	.25[c]	.25[c]
V	.55	.59	.60
C × V	−.91[b]	−.79[a]	−.82[a]
Multiple R	.30[c]	.30[c]	.32[c]
FOR condition			
Conscientiousness	.58[b]	.34[c]	.35[c]
V	.20	.18	.30
C × V	−.56	−.43	−.56
Multiple R	.36[c]	.36[c]	.37[c]

Note. Regression coefficients are standardized coefficients. [a]$p < .05$. [b]$p < .01$. [c]$p < .001$.

Table 5. Simulataneous regression of end-of-semester GPA onto cognitive ability, conscientiousness, and V

Inconsistency measure

Latent variable

Variable	Unadjusted V	Adjusted V	Factor scores
Generic condition			
Cognitive ability	.33[c]	.33[c]	.32[c]
Conscientiousness	.25[c]	.26[c]	.26[c]
V	−.10[a]	−.11[a]	−.13[a]
Multiple R	.41[c]	.43[c]	.43[c]
FOR condition			
Cognitive ability	.29[c]	.29[c]	.28[c]
Conscientiousness	.31[c]	.32[c]	.32[c]
V	−.18[b]	−.18[b]	−.20[c]
Multiple R	.45[c]	.45[c]	.46[c]

Note. Regression coefficients are standardized coefficients. [a]$p < .05$. [b]$p < .01$. [c]$p < .001$.

two. Although not shown in the table, moderation results for these analyses were similar to those for the analyses of Table 4, with V moderating the relationship of GPA to C in the Generic condition but not in the FOR condition.

5. Discussion

In this study, we examined the efficacy of two methods with the potential to increase the validity of personality predictors of performance. The results provided further support to previous studies of FOR effects on responses to personality questionnaires. Our study findings showed that providing a FOR resulted in increased validity. Schmit *et al.* (1995) found an increase of .16 in validity of conscientiousness between a noncontextualized and an

'at school' FOR condition. Bing *et al.* (2004) found an increase of .09 in validity of conscientiousness from .42 to .51 when moving from a generic to a FOR condition. Likewise, Lievens *et al.* (2008), in Study 2, also a within-subjects comparison, found an increase in validity of *C* from .05 to .38, a .33 difference. The difference of .07 found in the present study from .20 to .27 is comparable to those of Bing *et al.* (2004), providing a conservative estimate of the increase in validity that could be expected from what it would have been using generic instructions.

The results of the present study add to the body of evidence suggesting that within-person inconsistency is a general characteristic of responding. The positive correlations between standard deviations across scales suggest that the characteristic is not specific to a single personality dimension and the positive correlations across instructional conditions suggest that it is not specific to a single administration of a questionnaire. All of these point to a general characteristic, one that is possessed by the respondents and that would affect responses to any questionnaire. This conclusion supports findings by Baird *et al.* (2006) and Fleisher and Woehr (2008). It is in contrast to the findings from the traitedness literature in which the focus has been on differences in inconsistency from one trait to another. Our findings demonstrate that inconsistency is a general characteristic, not confined to just a few traits.[3]

Our findings suggest that in certain instances, a simple definition of inconsistency based on the mean of within-dimension standard deviations is useful. We believe that the use of a 7-point response scale was a primary factor in uncoupling the central tendency of the scales from variability of responses to the items within the scales. The high correlations between unadjusted V and the two adjusted measures testifies to the efficacy of the un-adjusted measure. We acknowledge, however, that even with a 7-point scale, there may be instances in which average response levels within a group will be so high or low that correlations between level and variability across respondents within the group will not be negligible, as they were here. In those cases, both the unadjusted V computed as the mean of residuals and the latent V would be available as alternative measures. The results found in this study for the two adjusted V measures were nearly identical to those for the unadjusted V, suggesting that they would serve well in other more demanding situations.

The similarity of the latent V measure to both the unadjusted V and the adjusted V suggests that latent variable modeling techniques may be profitably used to represent variability as well as level. This means that it may be possible to utilize the power of latent variable techniques in a new and relatively unresearched area of investigation, as illustrated in Eid and Diener (1999).

The data of this study also provided evidence that individual differences in response variability or

inconsistency are related to the reliability of scales. The results of tests of Hypothesis 4 showed that subgroups of participants with the most inconsistency had the smallest reliability for each of the Big Five dimensions. The results suggest that the inconsistency measured at the level of the individual participant is propagated to traditional measures of reliability measured at the group level. Because an inconsistency score can be computed for each participant, our method will help future researchers and practitioners in identifying respondents whose questionnaire responses would be least reliable. One implication of this result is that it presents another factor that test constructors must take into account when assessing test reliability. For example, it is well known that reliability estimates are affected by the heterogeneity of the normative sample, and knowledgeable test makers likely ensure that when estimating reliability of a new test they use as heterogeneous a sample as possible. If the findings in this study are replicated in subsequent analyses, test constructors will now have to consider the inconsistency characteristics of the sample as well as its heterogeneity.

A nefarious implication of our study is the possibility that persons wishing to market tests with high reliability estimates could select respondents on the basis of inconsistency, for example, by administering a Big Five questionnaire as a pretest and then administering the test to be marketed to those respondents who were most consistent based on the Big Five *V* estimates. Such a strategy could add as much as 10% to estimates of reliability. A more theoretical implication is that the conceptualization of reliability may have to be altered to take into account the inconsistency of the respondents. Now that there is an accumulation of evidence suggesting that there are individual differences in inconsistency, it may be appropriate at some time to ask what the reliability of a test would be if respondents were perfectly consistent or to define reliability in terms of some arbitrary value of a measure of inconsistency.

The results of this study clearly suggest that test reliability depends on the characteristics of respondents as well as the characteristics of the items of the test to which they are responding. This means that a perfectly constructed set of items for a scale could have a very low reliability estimate due to the inconsistency of the respondents. How to best characterize the inconsistency observed – as random response error or specific factor error (Schmidt *et al.*, 2003) or as something else – is not clear from the results of this initial study. We note, however, that if a substantial portion of the variability were due to differences in specific factor error, we would have expected inconsistency to have been smaller in the FOR condition, in which specific contexts for interpretation of the items was given. In the absence of large differences, it appears that the source of inconsistency is more likely random response error.

The absence of large differences in reliability between the Generic and FOR conditions replicates results of Lievens *et al.* (2008) study 1 and of Bowling and Burns (2010). Given the striking demonstration by Lievens and colleagues of the effects of manipulated alterations in FOR on reliability, this absence of a difference in reliabilities suggests that the effect of FOR instructions on within-person variability that was expected by Lievens and colleagues is very small, if it exists at all. This result is corroborated by the finding that mean inconsistency scores in the present study were slightly larger in the FOR conditions than in the Generic conditions, a result opposite of that predicted by hypothesis 3. Thus, the personal characteristics that affect inconsistency seem to be much more powerful than the context induced by FOR instructions.

Our fifth hypothesis that inconsistency would be related to validity was supported in both the Generic and FOR conditions. In each condition, *V* was a valid predictor by itself. Moreover, adding *V* scores to a regression of academic performance onto conscientiousness scale scores resulted in a significant increase in validity, by about .08 in each condition. In each case, the relationship of *V* to the criterion was negative, suggesting that when controlling for conscientiousness, those who were more inconsistent had lower academic performance (i.e., GPAs).

A question that arises out of this result is whether the characteristic whose individual differences are reflected in *V* acts directly on performance or whether it is related to performance only through its relationship to cognitive ability. The significant simple correlations of *V* with GPA argue that *V* is not merely a suppressor variable. This, along with the significant incremental validity of *V* over cognitive ability in the multiple regressions of Table 5 suggests that *V*'s validity in prediction of GPA was through some process other than its correlation with cognitive ability.

The finding that *V* moderated the relationship of conscientiousness to academic performance in the generic condition replicates findings from the traitedness literature in which intraitem variability on the constructs being correlated, called traitedness in that literature, moderated the relationships between pairs of variables (Britt, 1993; Dwight *et al.*, 2002). The present results generalize those earlier findings, however, by suggesting that the moderator is not a characteristic that is specific to the construct being investigated, that is, traitedness, but a general characteristic of responding to all personality scales, that is, inconsistency. One implication of our findings is that future research needs to distinguish traitedness on a particular construct from the general characteristic of inconsistent responding as reported in this study as well as in Baird *et al.* (2006) and Fleisher and Woehr (2008).

5.1. Limitations and future research

An important limitation of the present study is the use of student samples in a relatively low stakes situation. We cannot be sure whether or not there would be individual differences in intraindividual variability that would be as reliable and valid as those found in this study. It might be, for example, in very high stakes situations, that respondent faking or socially desirable responding would push the average scale scores of respondents to levels that would limit intraindividual variability, thereby reducing reliability and through range restriction, reducing the validity of *V*. Of course, a high stakes situation would not eliminate the consistent and inconsistent responders from the sample, it would just make it harder to identify them and use the differences in inconsistency to increase validity. Clearly future research using high stakes testing situations is warranted.

At the present time, a better understanding of the reason for increased validity in the FOR condition is needed. Lievens *et al.* (2008) suggested that using a FOR condition would lead to decreased within-person variability. Our data, results from Lievens and colleagues' study 1 and from Bowling and Burns (2010) do not offer strong support for this hypothesis. Future research should be directed at elaborating the reasons for differences in validity between Generic and FOR instructions.

Future research directed toward discovering the nomological net surrounding inconsistency of responding is needed. Only the Dwight *et al.* (2002) and this study have investigated inconsistency as a direct predictor of criteria as opposed to a moderator, and just a few studies have considered inconsistency as measured here as a general and stable characteristic of respondents. Only the Biderman (2007) study and this one have examined the relationship of inconsistency to respondent characteristics, finding that it is negatively related to cognitive ability. However, because we found that *V* afforded incremental validity in prediction of GPA over and above cognitive ability it appears that *V* represents other characteristics. Further research on other characteristics related to inconsistency are clearly called for.

Although there have been several threads of research on inconsistency, the vast majority of research on personality has studied the level of behavior rather than its variability. This is particularly true in the area of selection in staffing research. Other than the few studies mentioned above, we are aware of no other studies of the relationship of performance to variability. The results of this study show that the study of variability may be a fruitful one that may increase the ability of selection specialists to predict performance criteria. Our study focused on the relationship of within-person variability in the predictor to level in the criterion. Other studies might examine variability in the criterion and its relationship to both level and variability in the predictor.

5.2. Conclusion

This study has pointed out the importance of two factors in the prediction of performances using personality questionnaires – the use of FOR instructions and the measurement of within-person inconsistency. Our findings reinforced previous research concerning the efficacy of FOR instructions for increasing validity of personality questionnaires. Our findings also pointed out the importance of a behavioral measure – the inconsistency of that behavior – that has received relatively little attention in organizational literature.

The overarching positive practical result of our study is the evidence that the validity of a common personality questionnaire as a predictor of academic performance can be manipulated over a fairly wide range, causing uncorrected validity to vary from .20 to .36 depending on how the questionnaire is administered and scored. When conscientiousness is used alone under generic conditions, the smallest validity is expected. However, when conscientiousness is measured under FOR instructions and a measure of inconsistency extracted from the personality questionnaire is included as a predictor, the validity of the personality questionnaire could increase by as much as 80 percent from that base value. As very little effort or change in questionnaire wording or procedures is required to achieve this gain, the practical benefit of the manipulations is available at a very small price.

Notes

1. A fourth estimate of inconsistency – the square root of the mean of variances across dimensions within each condition – was also computed. For the Generic condition, its correlation with Gen *V* was .994. For the FOR condition, its correlation with FOR *V* was .992.
2. For readers interested in the relationships of the BIDR scales to *V* measures, the correlations of Gen *V* and FOR *V* with *SD* were .063 and .119 ($p < .05$), respectively. Correlations of IM with Gen *V* and FOR *V* were −.016 and .049. Although *SD* was significantly negatively correlated with GPA in the analyses presented in Tables 3, 4, and 5, inclusion of the BIDR scales in the regressions presented there did not change any of the conclusions drawn from the analyses.
3. To further illustrate the generality of inconsistency, correlations of interitem standard deviations from the BIDR *SD* and IM scales with Gen *V* were .585 and .571, respectively. Correlations with FOR *V* were .396 and .468.

References

Baird, B. M., Kimdy, L., & Lucas, R. E. (2006). On the nature of intra-individual personality variability: Reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology, 90*, 512–527.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13*, 225–232.

Biderman, M. D. (2007, April). *Variability indicators in structural equation models. Part of symposium: Examining old problems with new tools: Statistically modeling applicant faking*. Paper presented at the 22nd Annual Conference of The Society for Industrial and Organizational Psychology, New York, NY.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89,* 150–157.

Bowling, N., & Burns, G. (2010). A comparison of work-specific and general personality measures as predictors of work and non-work criteria. *Personality and Individual differences, 49,* 95–101.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade point average from the revised and recentered SAT I: Reasoning Test (College Board Rep. No. 2000–1)*. New York: College Entrance Examination Board.

Britt, T. (1993). Metatraits: Evidence relevant to the validity of the construct and its implications. *Journal of Personality and Social Psychology, 65,* 554–562.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance; examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91,* 40–57.

Dwight, S. A., Wolf, P. P., & Golden, J. H. (2002). Metatraits: Enhancing criterion-related validity through the assessment of traitedness. *Journal of Applied Social Psychology, 32,* 2202–2212.

Edwards, B. D., & Woehr, D. J. (2007). An examination and evaluation or frequency-based personality measurement. *Personality and Individual Differences, 43,* 803–814.

Eid, M., & Diener, E. (1999). Intra-individual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76,* 662–676.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52,* 217–250.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social Psychology, 80,* 1011–1027.

Fleisher, M. S. and Woehr, D. J. (2008, November). *The big six? The importance of within-person personality consistency in predicting performance*. Paper presented at the Meeting of the Southern Management Association, St. Petersburg, FL.

Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84,* 537–552.

Greenleaf, E. A. (1992). Measuring extreme response style. *The Public Opinion quarterly, 56,* 328–351.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18,* 135–164.

Halpin, G., Halpin, G., & Schaer, B. B. (1981). Relative effectiveness of the California Achievement Tests in comparison with the ACT Assessment, College Board Scholastic Aptitude Test, and high school grade point average in predicting college grade point average. *Educational and Psychological Measurement, 41,* 821–827.

Heller, D., Ferris, D. L., Brown, D., & Watson, D. (2009). The influence of work personality on job satisfaction: Incremental validity and mediation effects. *Journal of Personality, 77,* 1051–1084.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice: The frame-of-reference and pre-test validity information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13,* 75–86.

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology, 88,* 545–551.

Kernis, M. H. (2005). Measuring self-esteem in context: The importance of stability of self-esteem in psychological functioning. *Journal of Personality, 73,* 1569–1605.

Kim, S., & Feldt, L. S. (2008). A comparison of tests of equality of two or more independent alpha coefficients. *Journal of Educational Measurement, 45,* 179–193.

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93,* 268–279.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

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.

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance. *Human Performance, 18,* 343–357.

Muthén, B. O. (1998–2004). *Mplus Technical appendices*. Los Angeles: Muthén & Muthén.

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

Paulhus, D. L. (1984). Two-component models of social desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8,* 206–224.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607–620.

Schwab, D. P. (2005). *Research methods for organizational studies* (2nd ed.). London: Lawrence Erlbaum.

Smith, D. B., Hanges, P. J., & Dickson, M. W. (2001). Personnel selection and the five-factor model: Reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology, 86,* 304–315.

Wonderlic Inc. (1999). *Wonderlic's personnel test manual and scoring guide*. Chicago: Wonderlic Inc.