Increasing the validity of personality questionnaires.

Craig M. Reddock
Michael D. Biderman
University of Tennessee at Chattanooga

Nhung T. Nguyen
Towson University

Authors' Note: Correspondence regarding this article should be sent to Michael Biderman, Department of Psychology / 2803, U.T. Chattanooga, 615 McCallie Ave., Chattanooga, TN 37403. Tel.: (423) 425-4268. Email: Michael-Biderman@utc.edu

Paper submitted for presentation at the 25th Annual Conference of The Society for Industrial and Organizational Psychology, Atlanta, GA 2010.

Poster

TITLE

Increasing the validity of personality questionnaires.

ABSTRACT

The efficacy of use of frame-of-reference instructions and a measure of intra-individual variability in prediction of GPA was investigated. Validity of conscientiousness was greater when frame-of-reference instructions were given. Intra-individual variability was found to add incremental validity over that of conscientiousness alone.

PRESS PARAGRAPH

Forty years after Guion and Gottier's (1965) cautions concerning the use of personality measures as predictors of job performance, uncorrected estimates of validity of conscientiousness, the best predictor from the Big Five, hover in the .2 range. Recent research has suggested possible ways to enhance the validity of such measures. The efficacy of two such methods on the validity of conscientiousness as a predictor of undergraduate GPA was investigated. Use of frame-of-reference instructions resulted in a significant increase in validity. Measuring intra-individual variability also increased validity of predictions from a personality questionnaire. Implications of the results are discussed.

The usefulness of personality measures for predicting performance has been the subject of much debate. For example, Guion and Gottier (1965) stated that, for a variety of reasons, personality measures as predictors of job performance were of no practical value. This study is frequently cited by both critics and proponents of personality measures who view it as the seminal work on the topic. Since then there has been considerable research on how personality can be measured and used to predict performance focusing recently on the Big 5 personality factors (Dudley, Orvis, Lebiecki, and Cortina, 2006; Murphy and Dziewecynski, 2005; Ones, Dilchert, Viswesvaran, and Judge, 2007; Smith, Hanges and Dickson, 2001). Whereas these studies have shed some light on the predictive ability of personality tests, 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."

Recently, two factors that might enhance the validity of personality assessment have been investigated. The first is the frame of reference provided by the instructions given with personality questionnaires. Some studies found that when specific frame-of-reference (FOR) instructions were given (e.g., Bing, Whanger, Davison, & VanHook, 2004; Lievens, De Corte, & Schollaert, 2008), validity was enhanced in FOR conditions when compared to generic instructions, although at least one study (Schmit, Ryan, Stierwalt, & Powell, 1995) found no significant differences between FOR and generic instruction conditions. The second factor to receive attention is intra-individual or within-person variability. Recent research has suggested that within-person consistency may be a personality construct in its own right and may be a predictor of performance (Fleisher, & Woehr, 2008; Biderman, 2007).

The purpose of the present study was to investigate the extent to which the above two factors could augment the validity of a personality questionnaire as a predictor of academic performance. First, the validity of a personality measure obtained under generic instructions was compared with validity when frame-of-reference instructions were given to respondents. Then, the relationship of performance to intra-individual variability as assessed by a measure of inconsistency of responding was investigated.

The criterion used in the present study was academic performance operationalized as undergraduate grade point average (GPA). Because of the general finding that conscientiousness is the only Big Five trait to consistently predict academic performance (e.g., Dollinger & Orf, 1999; Goff & Ackerman, 1992; Dwight, Cummings, & Glenar 1998; Conard, 2006; O'Connor & Paunonen(2007) in the present study, we focused only on conscientiousness and assessed its validity as a predictor of GPA.

Frame-of-reference Instructions

Recent attempts to increase the validity of personality tests have focused on the impact of adding FOR instructions to personality measures (Bing, et al., 2004; Holtz, Ployhart, and Dominguez, 2005; Hunthausen, Truxillo, Bauer, & Hammer,2003; Lievens, et al., 2008; Schmit et al., 1995). FOR instructions are context-specific tags that are added to individual personality items. For example, changing "I am always prepared." to "I am always prepared at school" creates an "at school" frame of reference.

It has been proposed that under generic instructions respondents may use the wrong frame of reference when responding to all questionnaire items. If the frame of reference held by respondents, even though consistently applied, is not related to the context in which the criterion is measured, then 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. Thus, using an incorrect frame-of-reference by respondents will result in reduced

validity. This suggests that providing a frame-of-reference that corresponds to the criterion context will result in generally higher validity. Because the criterion in the present study is undergraduate GPA, the context of the FOR tags will be "at school". The results of the studies investigating frame-of-reference effects lead to the following hypothesis.

Hypothesis 1: The validity of conscientiousness as a predictor of academic performance will be larger for individuals responding to personality items under at school frame-of-reference instructions than for individuals under generic instructions.

Intra-individual variability

The second factor that may affect the validity of predictions from personality questionnaires is that there may be individual differences in the within-person consistency of responding to test items. According to Eid and Diener (1999), "intra-individual variability is defined as either a) intra-individual variability in item scores of a questionnaire or b) intra-individual fluctuations across situations and time." (p. 662). There has been sporadic interest in intra-individual variability when responding to personality items. In such research, intra-individual 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 personality across time (e.g., Eid & Diener, 1999; Kernis, 2005).

Since coefficient alpha, the most frequently used estimate of reliability, is based on the covariances between items, it would be expected that for a given scale, lower intra-item variability would be associated with larger covariances between items and thus larger estimates of reliability. Thus low consistency of item responses within the same dimension will adversely affect the reliability of scale scores based on those responses (e.g., Lievens et al., 2008). From this it would also be expected that stable individual differences in variability of responses to personality items across scales (e.g., Baird, Kimdy, & Lucas, 2006) would also be related to the reliability of scale scores. Thus it is expected that reliability estimates obtained from respondents for whom an overall measure of inconsistency is large would be smaller than reliability estimates from respondents judged less inconsistent by the same measure. This leads to Hypothesis 2:

Hypothesis 2: Reliability coefficients obtained from inconsistent respondents as indicated by large intra-individual standard deviations will be smaller than reliability coefficients from consistent respondents.

Very recently, researchers have considered intra-individual variability as a characteristic of responses, possibly on a par with the characteristics most often measured in personality questionnaires (e.g., Baird et al., 2006; Biderman, 2007; Fleisher, & Woehr, 2008). For example, Biderman (2007) presented a structural equation model in which standard deviations of within-person responses to items of a specific dimension served as indicators of a latent variable. This "Variability" variable was found to be related to cognitive ability in three studies. Other investigators, e.g., Fleisher and Woehr (2008) found self-reported consistency moderated personality-performance relationships. Taken together, these results suggest that a measure of intra-individual response inconsistency estimated from responses to personality questionnaires might enhance the validity of the questionnaires. This leads to the final hypothesis:

Hypothesis 3: Inclusion of intra-individual within-dimension variability will predict academic performance and significantly add to the validity of conscientiousness as a predictor of academic performance.

METHOD

Participants.

Participants were 329 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.

Measures.

Big Five. The sample 50-item scale from the International Personality Item Pool web site was used for both the Generic and Frame of Reference (FOR) conditions. For the Generic condition, the items were used without modification, with the exception that each 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. For this study, with the exception of comparisons of reliability to test hypothesis 2, only the conscientiousness dimension of the Big Five was included in the analyses.

Intra-individual variability. 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 intra-individual variability variable (labeled V) was measured as the mean of the standard deviations of responses to Big Five items. Specifically, five standard deviations were computed for each participant – the standard deviation of responses to extraversion items, of responses to agreeableness items, conscientiousness, stability, and openness items. The measure of intra-individual variability used here was simply the mean of the five standard deviations Two values of V were obtained for each respondent – one from the responses in the Generic condition and one from the responses in the FOR condition.

Other measures. The Wonderlic Personnel Test (Wonderlic, 1999) was administered to all participants prior to the personality questionnaires. 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). Neither the WPT nor these measures were analyzed for this paper.

Academic performance. Cumulative grade point average (GPA) was the criterion in the analyses that follow. GPAs were obtained from university records at the end of the semester in which participation occurred.

Design and Procedure

The research 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. Participants were run in groups of 1 to 20 persons. Fifty-one percent received the Generic version of the Big Five scale before the FOR version.

RESULTS

Table 1 presents means, standard deviations, reliabilities, and correlations of the Big Five scales from the Generic and FOR conditions along with end-of-semester GPA. Correlations between the Big Five scale scores in both conditions were generally positive. As expected, with the exception of generic condition agreeableness, only the correlations of conscientiousness with GPA were significant.

Hypothesis 1 stated that the validity of conscientiousness would be larger when a frame-of-reference 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 (Cohen & Cohen, 1983, p. 56) the validity from the Generic Condition of .20 was smaller than the validity of .27 from the FOR condition ($p < .05$, one-tailed). Hypothesis 1 was thus supported.

The second hypothesis was that inconsistent respondents as indicated by standard deviations of responses within dimensions would exhibit lower scale reliabilities than more consistent respondents. To test this hypothesis, the mean of the five within-dimension standard deviations was computed for each respondent forming a variable labeled V, for Variability. A value of V was obtained from each condition for each respondent. Median splits on the distribution of V within each condition were performed, forming two groups – a low variability group and a high variability group for that condition. Reliabilities of the Big Five scale scores were then computed for each variability group within the Generic and within the FOR condition. Table 2 presents the reliability coefficients for each scale by group for the two conditions. In each comparison, the scale reliability computed from the low variability, or more consistent group was larger than that computed from the high variability group. All differences were significant at $p < .05$ using a test for independent reliability coefficients presented by Kim and Feldt (2008), indicating that the inconsistency of responding represented by the V variable was reflected in the familiar coefficient alpha. These results supported Hypothesis 2.

Hypothesis 3 stated that within dimension intra-individual variability would be a significant predictor of academic performance and would significantly add to the validity of conscientiousness as a predictor. This hypothesis was tested by computing the simple correlation coefficients of GPA with the V variable from each condition. Incremental validity over conscientiousness in each condition 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 V variable from each condition was a valid predictor of GPA by itself ($p < .01$ for the Generic condition, $p < .05$ for the FOR condition). Moreover, the V variable within each condition contributed incremental validity over Conscientiousness scale scores when used in two-predictor equations ($p < .001$ for each). In all cases, the relationship of GPA to the V variable was negative, indicating that persons who exhibited less variability of responding within dimensions, i.e., were more consistent, had higher GPAs.

The overarching result of these analyses is that the validity of a personality questionnaire including Conscientiousness can vary from .20 to .34 depending on how the questionnaire is administered and scored. When Conscientiousness is used alone under generic conditions, the smallest validity could be 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 70 percent.

DISCUSSION

This study examined the efficacy of two methods with the potential to increase validity of personality predictors of performance. The results of the present study provided

support to the expectation from recent studies on frame of reference effects for respondents to personality questionnaires. Previous studies have suggested that providing an external frame of reference should result in an increase in validity. In a within-subjects comparison comparable to that of the present study, Bing et al. (2004) found an increase 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. The differences found in the present study from .20 to .27 are certainly in line with those of Bing et al. (2004) suggesting that using the FOR instruction will increase validity by .07 or more.

The data of this study provided evidence that individual differences in response variability or inconsistency are related to both the reliability of scales and to validity of those scales. These results are perhaps the most important contribution of the present study.

The tests of Hypothesis 2 showed that subgroups of participants with the smallest variability had the largest reliability for each of the Big Five dimensions. The results suggest that the inconsistency represented by variability (V) is mirrored in traditional measures of reliability. Since a variability score can be computed for each participant, this suggests a way of identifying those respondents who are most inconsistent.

Our third hypothesis that V 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 GPA onto conscientiousness scale scores resulted in a significant increase in validity, by about .08 in each condition. In each case, the partial correlation of V to the criterion was negative, suggesting that when controlling for conscientiousness those who were more inconsistent had lower GPAs.

This study has pointed out the importance of a behavioral measure – the inconsistency of that behavior – that has received little attention in organizational literature. Although there have been several threads of research on inconsistency, the vast majority of research has studied the level of behavior rather than its variability. This is particularly true in the area of selection in staffing research. 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 intra-individual variability in the predictor to level in the criterion. But other studies might examine variability in the criterion and its relationship to both level and variability in the predictor.

The V measures based on standard deviations open up possibilities for identification of respondents who are responding inconsistently. The relationship of V to scale reliabilities shows that estimated reliability of a scale is a function of both scale construction and of the inconsistencies of the respondents. 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. Whether low reliability is due to poor items or inconsistent respondents is not measurable with a single administration of a questionnaire. However, use of appropriate experimental designs could separate the two.

The reason for the increase in validity of the FOR condition has not been completely resolved. One possibility, suggested by Lievens et al. (2008) is that when given generic instructions, respondents pick a context and respond as they would in that context. If the context is similar to that in which the criterion behavior is obtained, validity will be maximized. But if the context is different from that of the criterion, the validity of the personality questionnaire will suffer.

We hope that future research will explore the utility of estimation of within-subject inconsistency in selection situations. We believe that such study will profit not only from the incremental validity that measures of inconsistency will bring to the selection situation, but from the better understanding of respondent behavior that a study of inconsistency may provide.

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.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* 2nd Edition. Hillsdale, NJ: Lawrence Erlbaum.

Conard, M. A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality, 40*, 339-346.

Dollinger, S. J., & Orf, L. A. (1991). Personality and performance in "personality": Conscientiousness and openness. *Journal of Research in Personality, 25*, 276-284.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., and 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., Cummings, K. M., & Glenar, J. L. (1998). Comparison of criterion-related validity coefficients for the mini-markers and Goldberg's markers of the big five personality factors. *Journal of Personality Assessment, 70,* 541-550.

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.

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

Feldt, L. S. (1980). A test of the hypothesis that Ctronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99-105.

Fleisher, M. S., & 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.

Finney, S. J., & DeStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In Hancock, G. R., & Mueller, R. O. (Eds) *Structural Equation Modeling: A Second Course*. Information Age Publishing.

Goff, M., & P. L. Ackerman. (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.

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.

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.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences, 43,* 971-990.

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.

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

Schwab, D.P. (2005). *Research Methods for Organizational Studies* (2nd ed.). 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: IL: Author.

Table 1. Means, standard deviations, and correlations of Big Five scale scores and end-of-semester GPA from the FOR and Generic conditions. Reliabilities are on the diagonal.

		1	2	3	4	5	6	7	8	9	10	11
1	Generic Extraversion	.88										
2	Generic Agreeableness	.28	.76									
3	Generic Conscientiousness	.04	.19	.86								
4	Generic Stability	.25	.04	.10	.84							
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	.00	.80	
11	End-of-Semester GPA	-.05	.11	.20	-.04	.07	-.04	.06	.27	.03	.10	NA
	Mean	4.46	5.20	4.84	4.40	4.80	4.15	4.81	5.03	4.44	4.65	2.97
	Standard deviation	1.00	0.73	0.82	0.95	0.74	1.09	0.84	0.78	0.93	0.78	0.60

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

	Big Five Dimension				
	E	A	C	S	O
Generic Condition					
Consistent	.92	.84	.85	.90	.86
Inconsistent	.84	.71	.76	.83	.74
p <	.001	.001	.010	.001	.001
FOR Condition					
Consistent	.93	.86	.83	.89	.87
Inconsistent	.87	.80	.71	.80	.73
p <	.001	.050	.010	.001	.001

Table 3 – Simple and incremental validity of V in Generic and FOR conditions.

Variable	Standardized Coefficient	Multiple R
Generic Condition		
V	-.16[b]	
Conscientiousness	.23[c]	
V	-.20[c]	.28[c]
FOR Condition		
V	-.12[a]	
Conscientiousness	.34[c]	
V	-.23[c]	.34[c]

[a] $p < .05$ [b] $p < .01$ [c] $p < .001$