

The relationship of scale reliability and validity to respondent inconsistency

Michael D. Biderman

University of Tennessee at Chattanooga

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 accepted for presentation at the 26th Annual Conference of The Society for Industrial and Organizational Psychology, Chicago, IL. 2011.

Poster

TITLE

The relationship of scale reliability and validity to respondent inconsistency

ABSTRACT

The relationship of reliability and validity estimates for personality scales to respondent inconsistency measured on a separate scale was investigated. Scale reliabilities were larger when estimated using the most consistent respondents. Validity of conscientiousness as a predictor of GPA was higher in groups composed of more consistent respondents.

PRESS PARAGRAPH

Respondent inconsistency has recently been identified as a characteristic to be considered when analyzing personality questionnaires. This study investigated the relationship of scale reliability and scale validity to respondent inconsistency. It was found that reliability estimates were higher for respondents who were most consistent in their responses to items within a dimension. Moreover, validity of conscientiousness as a predictor of GPA was higher among the more consistent respondents. Implications for selection and test construction are discussed.

When one thinks of reliability, probably the first thing that comes to mind is the test that is being considered. Reliability estimates are typically reported as characteristics of tests or scales. Rarely are the characteristics of the respondents, the test takers, mentioned when reliability is reported. However, when sources of variance in responses to test items are listed (e.g., Schmidt, Le, & Ilies, 2003), the errors that suppress reliability are attributed to the respondent or to the interaction between the respondent and the content of the test items. For example, Schmidt et al. listed momentary variations in attention, mental efficiency, distractions on a given occasion – all respondent characteristics - as causes of random response error, one of the types of error that lower test reliability. They listed respondent-specific interpretation of the wording of questionnaire items, an interaction between respondents and items, as a source of specific factor error, another type of error that lowers test reliability. Although characteristics of the test items can affect these errors, in either case, the characteristics of the respondents clearly also can affect the estimates of test or scale reliability.

Interestingly, respondent characteristics are not often mentioned in descriptions of how to maximize reliability. If respondent characteristics are mentioned at all, the suggestion that reliability estimates depend on sample heterogeneity may be given, and most texts devoted to scale development mention only the representativeness of pilot samples for the population in which the scale will be used. One characteristic that has been given almost no attention is respondent inconsistency – the tendency for respondents to give different responses to items for which identical or nearly identical responses would seem to be appropriate. Such response inconsistency may be due to item wording differences. But it may also be due to respondent characteristics that lead to inconsistent responses to equivalent items – items representing the same dimension on a personality inventory, for example. In response to the great emphasis in test construction on item characteristics, the purpose of the present study was to examine the effect of respondent characteristics on estimates of reliability and validity. Specifically, we examined the extent to which reliability and validity estimates depend on within-person inconsistency. For this study, such inconsistency was measured as the mean of standard deviations of items indicating the same dimension on a Big Five instrument.

The study of within-person inconsistency has a long history. It has been studied under a variety of guises, including the study of metatraits (Britt, 1993), or traitedness (Britt, 1993; Dwight, Wolf, & Golden, 2002), extreme response style (e.g., Greenleaf, 1992), and variability of test scores across time (e.g., Eid & Diener, 1999; Kernis, 2005). In a group of recent studies, evidence that within-person variability be viewed as a separate personality characteristic has been presented (Fleeson, 2001; Baird, Kimdy, & Lucas, 2006; Edwards & Woehr, 2007; Fleisher & Woehr, 2008). These studies found evidence that within-person variability was a stable characteristic across time and that it moderated personality-performance relationships. Recently, Reddock, Biderman, & Nguyen (2010) found that within-person variability exhibited incremental validity over both conscientiousness and cognitive ability in prediction of an academic performance criterion. A byproduct of those analyses was the finding that groups of respondents exhibiting larger within-person variability had smaller estimates of reliability than those who exhibited smaller within-person variability. A limitation of their finding, however, was that the within-person variability was measured on the same scale on which differences in reliability were found. Thus, the relationship they found was one that might have been observed simply due to random differences in respondent inconsistency. The present study extended this finding by measuring inconsistency using one scale and then examining the effect of inconsistency from that scale on reliability and validity of other, different scales. If respondent

inconsistency were a random phenomenon, then it would not be expected to propagate across scales, and inconsistent responders as measured on one scale would not exhibit lower reliability on other scales. On the other hand, if respondent inconsistency is a stable characteristic, then inconsistent responders identified on one scale would exhibit lower reliability in different scales.

Because of the relationship of reliability to validity, it would be expected that any factor that affects reliability would also affect validity. For that reason, it was expected that criterion related validity would be highest among respondents who exhibited the least inconsistency. The study included three Big Five measures of conscientiousness each of which was used to predict end-of-semester undergraduate grade point averages (GPA) of respondents. The GPA data were used to investigate the relationship of the validity of conscientiousness to respondent inconsistency. We focused on the validity of conscientiousness because it is the Big Five variable most consistently found to predict academic criteria (Poropat, 2009; Trapmann, Hell, Hirn, & Schuler, 2007). The data were originally gathered as part of a separate investigation and details of that investigation will be reported elsewhere.

Method

Participants.

Participants were 206 undergraduates at a medium sized southeastern university, participating for course credit. Sixty-four were male. Mean age was 19.32 ($SD=4.86$). Percentage of Whites was 67.96, Black/African-American was 25.24, and Other was 6.80.

Measures

Big Five Scale 1. The scale used to measure respondent inconsistency was the IPIP 50-item Big Five Sample Scale (www.ipip.ori.org) administered via paper and pencil to all participants under instructions to respond honestly. Participants responded to each item on a scale of 1="Completely Inaccurate" to 7="Completely Accurate".

Big Five Scale 2. A second IPIP Big Five scale was used to assess reliability. It was the second block of 50 items available on the IPIP web site, taken from the 100-item Sample Scale.

Thompson Mini-Markers. The Thompson Mini-Markers (Thompson, 2008) represent an alternative measure of the Big Five. Each item is a single word rather than a descriptive phrase. Examples are "Shy", "Talkative", "Energetic", "Quiet", "Extraverted". This instrument was administered as part of a separate investigation. Respondents indicated how accurately each word described them on the same 7-point scale used for the Big Five scales.

Depression. The Costello and Comrey (1967) Depression scale was administered as part of a separate investigation. Sample items include: "When I wake up in the morning I expect to have a miserable day." and "I wish I were never born." Responses were on the same seven-point scale and higher scores reflected higher levels of chronic depression.

Self Esteem. The Rosenberg Self Esteem scale (Rosenberg, 1965 taken from Carmines and Zeller, 1979) was administered to all respondents as part of a separate investigation. Sample items include: "I feel I have a number of good qualities." and "I wish I could have more respect for myself." Responses were on the same seven-point scale and higher scores reflected higher levels of self-esteem.

Proactive Personality. A shortened version of the Bateman and Crant (1993) proactive personality scale was administered as part of a separate investigation. Sample items include "I am constantly on the lookout for new ways to improve my life." and "Wherever I have been, I have been a powerful force for constructive change." Responses were on the same seven-point scale, with larger values representing greater proactive personality.

GPA. Undergraduate grade point averages were obtained from school records immediately after completion of the semester in which respondents filled out the questionnaire.

Within-person Inconsistency. Within-person inconsistency was measured by computing the standard deviation of responses to the 10 items from each Big Five dimension in Big Five Scale 1 and then computing the mean of those standard deviations.

Procedure.

Participants completed the paper and pencil questionnaires in small groups of 2 to 15. After participants filled out an informed consent sheet, the Wonderlic Personnel Test (WPT) was administered, followed by the 100-item IPIP Big Five scale (from which the Big Five Scale 1 and Big Five Scale 2 were taken), the depression scale, the self esteem scale and the proactive personality scale in that order.

Results

Table 1 presents means and standard deviations of the scales administered to respondents. Table 2 presents correlations of the five dimensions from Big Five Scale 1 and the individual scale standard deviations, and the within-person inconsistency measure. Inspection of Table 2 reveals that the Big Five scale scores are somewhat positively correlated, that the standard deviations are more highly correlated than the scale scores, and that the measure of overall inconsistency exhibits very low correlations with the scale scores.

The positive correlations between the individual dimension standard deviations, leading to a reliability coefficient of .602 for the overall measure of inconsistency, provides some indication of stability of the inconsistency variable across the Big Five dimensions. This suggests that inconsistency may be a characteristic of the respondents that would affect their responses to other questionnaires.

To examine the effect of inconsistency on reliability estimates, the distribution of the inconsistency measure was examined and three groups of respondents defined – those least inconsistent, those with a midrange of inconsistency, and those most inconsistent. The three-way split of the distribution was done so that there were 68 or 69 persons at each level of inconsistency. Mean values of inconsistency for each group were .90, 1.16, and 1.43 for the low, medium and high inconsistency groups respectively.

The sample was split into three groups and reliability estimates were obtained for all the study variables for each group. Those reliability estimates are presented in Table 3. The reliability estimates are presented in two panels. In the top panel are reliabilities from the scales used to compute the inconsistency measure. Due to the inverse relationship of inconsistency to reliability based on the mathematics of the definition of reliability it would be expected that there would be differences in reliability between the three groups for Big Five Scale 1 – the scale used to define inconsistency - even if the differences in inconsistency were random across respondents. As expected, an inverse relationship was found with the mean of reliability estimates from the least inconsistent group equal to .889, that of the middle group equal to .841, and that of the most inconsistent group equal to .762.

A more salient comparison is of reliabilities of the scales not used to measure inconsistency. Differences in reliability estimates across inconsistency classifications would have to be due to the carryover of inconsistency as measured using one scale to affect reliability estimated in another scale. The bottom panel of Table 3 presents the reliability estimates from the scales whose responses were not used to define inconsistency. Mean of the reliability estimates varied from .864 for the group least inconsistent when responding to Big Five Scale 1 down to .780 for the most inconsistent group. A one way analysis of variance was performed on

reliability estimates in the bottom part of Table 3. The null hypothesis of equal mean reliabilities across inconsistency groups was rejected ($F(2,36) = 5.00, p < .05$) with R^2 equal .218. Post hoc tests using Tukey's b statistic indicated that the mean of the most inconsistent group was significantly different from the mean of the least inconsistent group.

To assess the extent to which validity was related to inconsistency group, correlations of GPA with measures of conscientiousness were computed for each group. Three measures of conscientiousness were available – those of Big Five Scale 1, Big Five Scale 2, and the Mini-Marker scale. Table 4 presents validity coefficients of the measures of conscientiousness for each group. As can be seen from inspection of the table, the validity was roughly the same for both the consistent group and the middle group, but fell off dramatically for the most inconsistent group. Three moderated regression analyses – one for each conscientiousness scale - were conducted to provide a more formal test of the extent to which inconsistency moderated the validity of conscientiousness. In each regression GPA was regressed onto the inconsistency measure, the conscientiousness measure, and the product of inconsistency and the measure of conscientiousness. If validity were lower for larger values of inconsistency, the regression coefficient for the product term would be negative. The standardized regression coefficient for the product term was -1.288 ($p < .05$ one-tailed), $-.848$ ($p < .10$ one-tailed) and $-.705$ ($p < .10$ one-tailed) providing some support for the suggestion in Table 4 that validity fell off as inconsistency increased.

Discussion

This study was conducted to examine the role of respondent inconsistency in the evaluation of tests for use in selection situations. Recent studies have suggested that within-person inconsistency may be a stable respondent characteristic. There is an increasing body of research suggesting that such within-person variability may play an important role in employee selection systems. This research replicates and extends key previous findings concerning within-person variability.

First, we found that within a Big Five instrument, inconsistency of responding to each of the dimensions was fairly stable, stable enough to allow an overall measure of inconsistency to have marginally acceptable reliability. The moderately positive correlations between within-scale standard deviations from Big Five Scale 1 suggest that respondents to a questionnaire who are inconsistent in their responses to one scale will tend to be inconsistent in their responses to other scales within the questionnaire. These results agree with those of Fleeson (2001), who measured inconsistency across periods of many days and found persons more inconsistent at one time to be more inconsistent at other times. They also agree with the results of Edwards and Woehr (2007) who found reliable self reports of inconsistency.

Second, and more importantly for the study of within-person inconsistency, an overall measure of inconsistency computed based on the responses to one instrument was related to the mean of reliability estimates from five different instruments. The relationship of reliability to inconsistency was not markedly different in the new instruments than it was in the instrument on which the measure was based. This is evidence for the stability of the inconsistency measure across at least the duration it took to administer the several questionnaires. If the differences in inconsistency observed in Big Five Scale 1 had been random, there would have been no relationship of mean reliability in the other scales to inconsistency group defined by responses to Big Five Scale 1. Since that relationship was significant with at least a moderate effect size, the results provide evidence consistent with the suggestion that within-person variability is a general personality characteristic that can affect estimates of reliability across personality scales.

Finally, the results provide some evidence for the relationship of within-person variability to criterion-related validity of personality tests. Validity coefficients of all three measures of conscientiousness for the most inconsistent group were less than .1, less than half as large as the smallest coefficient from the two groups of more consistent responders. Although the interaction term in regressions of GPA onto inconsistency and conscientiousness met typical criteria for significance only for the Big Five Scale 1, the signs of the coefficients were negative in the regressions of both Big Five Scale 2 and Mini-Marker conscientiousness scales. The small sample sizes surely limited power to detect the interaction in these analyses.

The results of this study clearly call for further investigation of respondent inconsistency. The simple measure employed here – the mean of within-dimension standard deviations – appears to be a useful measure, easily computed. Although some have adjusted measures like those computed here to remove any possibility of correlation of variability with mean scale value, e.g., Baird et al. (2006), the correlations in Table 2 show that the simple mean of standard deviations was not significantly correlated with four the Big Five Scale 1 scores and only modestly correlated with the fifth. Investigators wishing to insure that there was no possibility of contamination of the variability measure with mean value could regress the within-dimension standard deviations onto scale scores and then base the overall measure of inconsistency on the residuals in those regressions. Reddock et al. (2010) performed such regressions and found no differences in results vs. those obtained using the standard deviation of within-dimension responses employed here.

The results presented here suggest that it makes a difference who is chosen for pilot studies and normative samples for scales. Using a sample composed of inconsistent responders could yield reliability estimates as much as .08 lower, on the average, than those of a sample of consistent responders. At the present time, we have no data on the distribution of inconsistency in the population including respondents who are not college students. It is possible that there is a greater range of inconsistency in the population than in the sample of college students used for this study. A wider range would make the inconsistency of respondents even more salient for estimation of reliability. A macabre implication of the results found here is the possibility of specially selected samples chosen for their consistency to be used in pilot studies of personality scales so that reported reliability values would be as large as possible.

The relationship of validity to inconsistency suggested here also deserves further study. Clearly, if inconsistency affects reliability, then it must also have some effect on validity. Thus, it was expected that the group that was most inconsistent would have the lowest validity coefficients. However, there is the possibility that the relationship is that of an inverted U. Validity of all three conscientiousness measures for the group composed of least inconsistent respondents were slightly lower than those in the middle inconsistency group. Because of the small sample sizes employed here, it is too soon to draw conclusions concerning some limit on gains associated with low inconsistency. But the results here clearly indicate that some of the variation in validity from sample to sample may be due to variations in respondent inconsistency. Inspection of Table 4 shows that overall validity of conscientiousness was about .2, a value found in meta-analyses of conscientiousness as a predictor of GPA (e.g., Poropat, 2010). However, if only medium or low inconsistent respondents are used, that validity could increase to .3 or possibly even .4. On the other hand, if the sample included a large number of inconsistent respondents, the validity could be below .1. Again, the specter of respondents specially selected for use in validity studies is raised. A suggestion that follows from these results would be to report the mean inconsistency of respondents along with reliability and

validity information on a test to allow potential users of tests to take this inconsistency into account when deciding whether or not to use a test.

This study adds to a growing body of research indicating that within-person variability is a personal characteristic that may play an important role in the evaluation of personality scales, particularly for employee selection. A much-needed future research effort is an assessment of the temporal stability of the measure of inconsistency. The correlations obtained during the single administration of questionnaires reported here were consistently positive. But correlations of the measure of inconsistency across longer time periods are clearly needed to assess the stability of the characteristic.

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.
- Bateman, T. S., & Crant, J. M. (1993). The proactive component of organizational behavior: A measure and correlates. *Journal of Organizational Behavior, 14*, 103-118.
- Britt, T. (1993). Metatraits: Evidence relevant to the validity of the construct and its implications. *Journal of Personality and Social Psychology, 65*, 554-562.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: SAGE Publications, Inc.
- Costello, C. G., & Comrey, A. L. (1967). Scales for measuring depression and anxiety. *The Journal of Psychology, 66*, 303-313.
- 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 of frequency-based personality measurement. *Personality and Individual Differences, 43*(4), 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.
- 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., & 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.
- Greenleaf, E. A. (1992). Measuring extreme response style. *The Public Opinion quarterly, 56*, 328-351.
- 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.
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.
- Reddock, C. M., Biderman, M. D., & Nguyen, N. T. Increasing the validity of personality questionnaires. Paper presented at the 25th Annual Conference of The Society for Industrial and Organizational psychology, Atlanta, GA 2010.
- Rosenberg, M. (1965). *Society and the adolescent self image*. Princeton, NJ: Princeton University Press.
- 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.
- Thompson, E. R. (2008). Development and validation of an international English Big-Five Mini-Markers. *Personality and Individual Differences, 45*, 542-548.
- Trapman, S., Hell, B., Hirn, J.W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Journal of Psychology, 215*, 132-151.

Table 1. Means and standard deviations of scales used in the study.

<u>Scale</u>	<u>Mean</u>	<u>Standard Deviation</u>
Big Five Scale 1 E	4.75	1.04
Big Five Scale 1 A	5.30	0.74
Big Five Scale 1 C	4.57	0.87
Big Five Scale 1 S	4.24	0.99
Big Five Scale 1 O	4.85	0.82
Big Five Scale 2 E	4.84	0.86
Big Five Scale 2 A	5.05	0.72
Big Five Scale 2 C	4.79	0.95
Big Five Scale 2 S	4.22	0.93
Big Five Scale 2 O	4.76	0.85
Minimarker E	4.98	1.03
Minimarker A	5.67	0.78
Minimarker C	4.90	1.01
Minimarker S	3.88	0.99
Minmarker O	5.09	0.95
Depression	1.84	0.83
Self esteem	5.65	0.87
Proactive Personality	4.85	0.98

Table 2. Correlations between Big Five Scale 1 domain scores, Big Five Scale 1 standard deviations, and the overall Inconsistency measure. Values on the diagonal are reliability coefficients.

	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>
1 E Domain	.885										
2 A Domain	.317	.789									
3 C Domain	.007	.164	.823								
4 S Domain	.237	.176	-.021	.842							
5 O Domain	.244	.335	.270	.156	.812						
6 E SD	-.301	.021	.157	-.121	.040	NA					
7 A SD	.056	-.279	.049	-.179	-.024	.227	NA				
8 C SD	.073	.026	-.131	-.058	.048	.158	.140	NA			
9 S SD	.052	.048	.120	-.178	.107	.285	.336	.217	NA		
10 O SD	.140	.159	-.102	.000	-.102	.142	.231	.313	.265	NA	
11 Inconsistency	.002	-.005	.036	-.174	.022	.595	.621	.567	.688	.630	.602

Table 3. Reliability estimates for each group defined by inconsistency on V computed from Big Five Scale 1.

<u>Scale</u>	<u>Inconsistency</u>		
	<u>Low</u>	<u>Medium</u>	<u>High</u>
Big Five Scale 1 E	.920	.904	.843
Big Five Scale 1 A	.898	.783	.676
Big Five Scale 1 C	.824	.849	.799
Big Five Scale 1 S	.900	.841	.789
Big Five Scale 1 O	.897	.832	.705
	-----	-----	-----
Mean of scales defining inconsistency	.887	.841	.762
SD of scales defining inconsistency	.037	.043	.069
	-----	-----	-----
Big Five Scale 2 E	.842	.824	.743
Big Five Scale 2 A	.826	.687	.562
Big Five Scale 2 C	.784	.841	.840
Big Five Scale 2 S	.813	.803	.785
Big Five Scale 2 O	.867	.781	.663
Minimarker E	.907	.882	.825
Minimarker A	.902	.728	.763
Minimarker C	.868	.872	.849
Minimarker S	.809	.812	.772
Minmarker O	.876	.835	.774
Depression	.951	.884	.908
Self esteem	.869	.839	.839
Proactive Personality	.924	.853	.832
	-----	-----	-----
Mean of scales not defining inconsistency	.864	.819	.780
SD of scales not defining inconsistency	.048	.058	.090

Table 4. Validities of three measures of conscientiousness in prediction of GPA for the three inconsistency groups.

<u>Scale</u>	<u>Overall</u>	<u>Inconsistency</u>		
		<u>Low</u>	<u>Medium</u>	<u>High</u>
Big Five Scale 1 C	.249	.328	.413	.060
Big Five Scale 2 C	.191	.185	.375	.078
MiniMarker C	.194	.263	.346	.050
