Journal of Applied Psychology
2008, Vol. 93, No. 2, 268–279

0021-9010/08/$12.00 DOI: 10.1037/0021-9010.93.2.268

A Closer Look at the Frame-of-Reference Effect in Personality Scale Scores and Validity

Filip Lievens, Wilfried De Corte, and Eveline Schollaert
Ghent University

This article contributes to the understanding of why the use of a frame-of-reference leads to increased criterion-related validity of personality inventories. Two competing explanations are described and tested. A between-subjects ($N = 337$) and a within-subject ($N = 105$) study are conducted to test the hypothesized effects of use of a frame of reference on reliability and validity. Regarding the effects on reliability, use of a frame of reference reduces within-person inconsistency (instead of between-person variability) in responding to generic items. Use of a frame of reference further leads to higher validity as a result of the reduction of between-person variability and within-person inconsistency. Yet, reducing these inconsistencies is not enough. It is also important to use a frame of reference that is conceptually relevant to the criterion. Besides implications for contextualized personality inventories, these results provide an explanation for the moderate validities of generic personality inventories.

Keywords: frame of reference, personality scales, criterion-related validity, reliability, item responding

In the personality domain, recent research has experimented with the use of more contextualization in items. In particular, it has been argued that the common use of generic (or noncontextualized) personality items (e.g., "I pay attention to details") is one reason for the relatively low criterion-related validities of personality scales (Bing, Whanger, Davison, & VanHook, 2004; Schmit, Ryan, Stierwalt, & Powell, 1995). Therefore, contextualized personality inventories impose a specific frame of reference (e.g., "I pay attention to details *at work*") on test takers when responding to personality items. Empirical research has found considerable support for the use of a frame of reference as a way of improving the criterion-related validity of personality scales (Bing et al., 2004; Holtz, Ployhart, & Dominguez, 2005; Hunthausen, Truxillo, Bauer, & Hammer, 2003; Robie, Schmit, Ryan, & Zickar, 2000; Schmit et al., 1995).

However, a fundamental question has remained unanswered: Why is the criterion-related validity of contextualized personality inventories higher than that of noncontextualized personality inventories? Traditionally, it is assumed that use of a frame of reference increases reliability by reducing between-person inconsistency in item interpretation (Bing et al., 2004; Holtz et al., 2005). This reduction in between-person inconsistency is then also assumed to increase criterion-related validity. This study contrasts this traditional explanation with an alternative explanation of how provision of a frame of reference might lead to higher validity. For each explanation, we outline the rationale, theoretical background, and hypothesized effects on reliability and validity. Next, we used a between-subjects and within-subject design to test the explanations.

Filip Lievens and Eveline Schollaert, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; Wilfried De Corte, Department of Data Analysis, Ghent University, Ghent, Belgium.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium. E-mail: filip.lievens@ugent.be

Prior Research on the Frame-of-Reference Effect

Conceptually, the use of contextualized personality scales is based on the cognitive-affective system theory of personality (Mischel & Shoda, 1995). This theory posits that cross-situationally consistent behavior can be expected only when situations elicit psychologically similar cues and demands. As individuals' behaviors are conditional on the situation, Wright and Mischel (1987) referred to the underlying tendencies as conditional dispositions. The key measurement implication of this theory is that prediction of people's behavior can be improved when people are given a context, or frame-of-reference, when asked to describe themselves.

Empirically, prior research has found considerable support for the frame-of-reference effect in personality scales and criterion-related validities. The original study of Schmit et al. (1995) examined the frame-of-reference effect in a student setting. These authors hypothesized that use of a frame of reference would lead to more positive scale scores as compared with a noncontextualized format, because behavioral expectations are more constrained in specific contexts than in general. Furthermore, they expected an at-school frame of reference to increase the criterion-related validity of a Conscientiousness factor for predicting grade point average (GPA). Both hypotheses were confirmed. In another study, Robie et al. (2000) examined the impact of use of a frame of reference on the measurement properties of the Revised NEO Personality Inventory (NEO PI-R). Results showed more error variance in ratings when no frame of reference was used. Hunthausen et al. (2003) investigated the effects of frame-of-reference use on validity in a field setting. An at-work frame of reference moderated the validity of two Five-Factor Model factors (Extraversion and Openness) for predicting the job performance of customer service managers. These frame-of-reference scales also had incremental validity over cognitive ability. Whereas all prior

studies had used a between-subjects design, Bing et al. (2004) used a within-subject design, in which students were asked to complete both generic and contextualized personality inventories. Contextualized versions had incremental validity over noncontextualized versions and cognitive ability. Finally, Holtz et al. (2005) focused on test perceptions. They expected that candidates would have more positive face validity perceptions of contextualized inventories. However, there was no effect of frame of reference on students' perceptions.

Explanations of the Frame-of-Reference Effect

Traditional Explanation

The traditional explanation posits that adding a frame of reference reduces between-person variability in responding to generic personality items. According to this traditional explanation, test takers who respond to a generic personality inventory might be divided into different subgroups. Some test takers rate all generic items with a specific frame of reference, whereas other test takers use a different frame of reference across all generic items. Yet, it is assumed that each test taker consistently uses the same frame of reference across all items. For example, Holtz et al. (2005) noted,

"It is thought that when global personality items are used, it is difficult to predict what cues individuals will focus on. *Some* test-takers may respond in accordance with how they perceive their personality across situations while *others* may respond specifically to how they view themselves at work, home, or elsewhere [italics added]" (p. 76, see also Schmit et al., 1995, pp. 607–608).

A contextualized personality inventory is thought to reduce this between-person variability in the frame of reference adopted, as all groups are asked to conceptualize the items with an imposed frame of reference.

This traditional explanation can be visualized in a model, such as the one presented in Figure 1. This diagram shows that two (in this simplified example) latent variables influence item responses. For example, the first latent variable might be Conscientiousness (or any other personality trait) at school, and the second might be Conscientiousness (or any other personality trait) at work. For some test takers, the true model contains large loadings from the two items (in this simplified example) on the first latent variable but not on the second latent variable (top of Figure 1), whereas for other test takers, the opposite is true (bottom of Figure 1).

There is some support for the contention that test takers completing personality inventories can be divided into different groups. Most of this research has taken a person-centered approach wherein respondents (instead of items) served as variables and were clustered according to their personality scores (e.g., Asendorpf, Borkenau, Ostendorf, & Van Aken, 2001; De Fruyt, 2002). In addition, and more relevant to the current study, other research has tried to uncover groups of respondents on the basis of their response styles. Along these lines, different subgroups of respondents (regular responders, slight fakers, extreme fakers, etc.) have been identified (Rosse, Stecher, Miller, & Levin, 1998; Zickar, Gibby, & Robie, 2004; see also McFarland & Ryan, 2000). Finally, there is evidence that test takers can be distinguished in terms of the strategies used to respond to personality items. Some test takers relate items to previous experiences and behaviors, whereas other test takers think about how relevant others have characterized them with respect to the trait-relevant behavior mentioned (Gordon & Holden, 1998). The common thread running through these studies is that test takers can be distinguished on the basis of their response styles and strategies. Likewise, the traditional explanation assumes that test takers might differ in the specific frame of reference (at work, at home, at school, etc.) they adopt to answer items of a generic personality inventory (Bing et al., 2004).

According to the traditional explanation, the reliability of contextualized personality inventories is higher than that of noncontextualized personality inventories because between-person variability in the frame of reference used is reduced. Further, it is hypothesized that this reduction in between-person variability subsequently leads to an increase in validity. These assumptions are reflected in the following quotation from Bing et al. (2004):

> Noncontextualized personality items are open to interpretation by respondents in comparison to context-specific items. As a result, when answering test items, *one* respondent may consider the way he or she behaves at work, and *another* respondent may consider the way he or she behaves in social situations; thus these respondents, in essence, are not responding to the same item when taking into account their differences in item interpretation. Such differences in item interpretation lead to increases in measurement error and *subsequent* reductions in validity [italics added]. (p. 151)

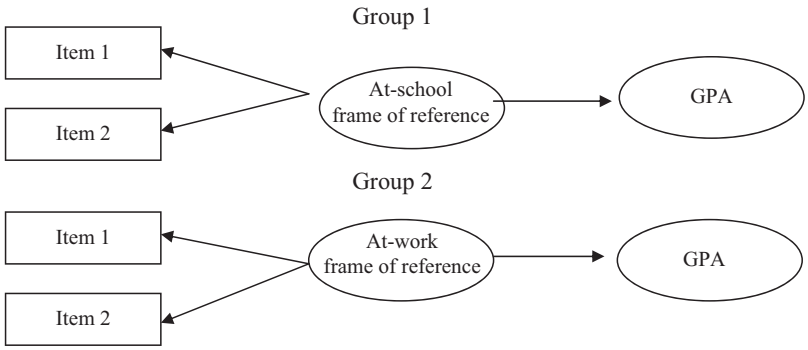

Figure 1. Example model of between-person variability in responding to generic personality items. GPA = grade point average.

Alternative Explanation

Although the aforementioned arguments have traditionally been used to explain the effects of using a contextualized frame of reference, there may be an alternative explanation. Specifically, imposing a frame of reference might reduce two different sources of variability: between-person variability and within-person inconsistency. In the remainder, we discuss the hypothesized effects of these sources of variability on reliability and validity. Hereby, we make the simplifying assumptions that test takers use only two frames of reference in responding to generic items and that these items are parallel measures of the underlying personality traits.

Similar to the traditional explanation, the alternative explanation posits that between-person variability is relevant when comparing generic with contextualized personality inventories. However, contrary to the traditional explanation, the effects of between-person variability on reliability are hypothesized to be minimal. This hypothesis is based on the formula of the specific index (i.e., Cronbach alpha) that is typically used to compute reliability (Murphy & Davidshofer, 2001). Cronbach alpha is only minimally affected by between-person variability as long as (a) test takers are consistent within themselves in their usage of a frame of reference and (b) the reliabilities of the separate frames of reference used are relatively similar. When test takers differ in the frame of reference used (while being consistent within themselves), the item (co)variances in the total sample will be a weighted sum (across different test takers) of the (co)variances of items that are all rated with a similar frame of reference. When the reliabilities, and thus the item, (co)variances of the frames of reference are similar, then any weighted sum of the within frame-of-reference item (co)variances will deviate only marginally from these within frame-of-reference item (co)variances themselves.

Similar to the traditional explanation, the alternative explanation hypothesizes that between-person variability in the frame of reference adopted might impact on validity. The reasoning is that the criterion-related validity of a generic personality inventory will increase when people interpret the items using a frame of reference that shows conceptual overlap with the criterion (i.e., correct frame of reference). Conversely, the validity will decrease when people interpret generic items using a frame of reference that does not show conceptual overlap with the criterion (i.e., incorrect frame of reference). In the context of completing a generic personality inventory, one might then expect that validity will be positively related to the number of people who use a correct frame of reference (i.e., a frame of reference that conceptually overlaps with the criterion). Essentially, this reasoning builds on the common notion that validity is related to conceptually matching the predictor and the criterion (Binning & Barrett, 1989; Goldstein, Zedeck, & Goldstein, 2002; Warr, 2000). For instance, a well-known example is that cognitive criteria are best predicted by cognitively oriented selection procedures, whereas noncognitive criteria are best predicted by noncognitively oriented selection procedures (see also Campbell, McCloy, Oppler, & Sager, 1993; Lievens, Buyse, & Sackett, 2005).

Apart from between-person variability, the alternative explanation also posits that contextualized personality inventories reduce within-person inconsistency in the frame of reference adopted. This inconsistency stems from a test taker's use of one frame of reference to answer one generic item and another frame of reference to answer another item. Thus, there might be within-person inconsistency in terms of the number of items rated with specific frames of reference. An example of such a model is represented in Figure 2. In this model, test takers interpret half of the items with an at-school frame of reference and the other half with an at-work frame of reference. This represents only one scenario, as test takers might also interpret 10%, 20%, and so forth of the items with differing frames of reference. A contextualized personality inventory is then expected to reduce this within-person inconsistency in the frame of reference adopted across items.

Conceptually, the notion of within-person inconsistency in the frame of reference adopted assumes that test takers think about each separate item when they respond to personality items, which is supported by think-aloud studies (Rogers, 1974a, 1974b) and by research on item context effects (Knowles, 1988; McFarland, Ryan, & Ellis, 2002). Specifically, within-person inconsistencies are in line with schematic theories of item responding (Holden, Fekken, & Cotton, 1990). According to schema theory, a respondent compares the content of each test item with a cognitive schema (Aronson & Reilly, 2006). This schema provides a cognitive context for processing relevant self-information. In particular, respondents will use the schema to conduct a selective memory search to find self-related information. There is ample evidence that respondents will then typically choose autobiographical memories that serve them best (Kunda & Sanitioso, 1989; Sanitioso, Kunda, & Fong, 1990; Sorrentino & Higgins, 1986). As a generic personality inventory consists of many different items, it can be assumed that not all items activate the same schemata. Hence,

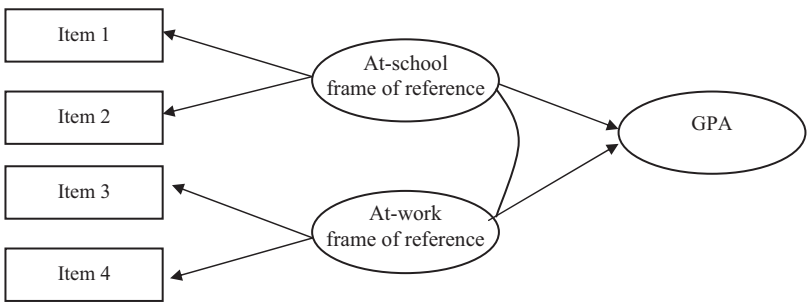

Figure 2. Example model of within-person inconsistency in responding to generic personality items. GPA = grade point average.

people might use different schemata (frames of reference) from one item to another for selecting relevant self-information.

According to the alternative explanation, within-person inconsistency is hypothesized to affect internal consistency reliability. If items are rated from a different perspective (frame of reference) by a test taker, the covariance among these items will be essentially a covariance between "different" items. Next, it is hypothesized that reliability will not be linearly affected by between-person variability. As reliability refers to consistency of measurement (regardless of what one measures; Schmidt, Viswesvaran, & Ones, 2000), reliability will be negligibly affected when the majority of items are rated with a specific frame of reference. Conversely, reliability will be affected when about half of the items are rated with one frame of reference and half of the items with another frame of reference, because respondents are then the most inconsistent within themselves. Thus, there will be a curvilinear relationship between the number of items rated with a specific frame of reference and reliability.

With regard to validity, the number of items rated with a specific frame of reference is posited to have large effects on validity. Again, the notion of conceptual overlap (Binning & Barrett, 1989; Goldstein et al., 2002) serves as a basis of this hypothesis. Using the correct frame of reference for a large number of items will increase validity, whereas the opposite will be true for using an incorrect frame of reference for a large number of items. Thus, validity will be positively related to the number of items that is rated with a correct frame of reference.

Summary

Taken together, the differences between the alternative explanation and the traditional explanation are threefold. First, the traditional explanation deals only with between-person variability in the frame of reference adopted. Conversely, the alternative explanation addresses both between-person variability and within-person inconsistency in the frame of reference adopted. Second, the traditional explanation suggests that reducing between-person variability will lead to an increase in reliability. The alternative explanation posits that reduction of between-person variability will not lead to an increase in reliability. Instead, the alternative explanation posits that within-person inconsistency will affect reliability. Third, according to the traditional explanation, reducing between-person variability is the sole explanation for the increase in criterion-related validity obtained with contextualized personality inventories. In opposition, the alternative explanation posits that reducing both between-person variability and within-person inconsistency leads to the increased validity of contextualized personality inventories.

Overview of Studies

We conducted two studies to test the rival explanations related to the frame-of-reference effect in personality inventories. The first study focused on between-person variability in the frame of reference adopted in all items and examined the two explanations' competing hypotheses with regard to the effects of between-person variability on reliability and validity. To this end, we used a between-subjects design in the first study, wherein participants were randomly assigned to three conditions: (a) a noncontextualized/generic condition, (b) a contextualized/frame-of-reference condition (at school), and (c) a contextualized/frame-of-reference condition (at work). As the criterion measure was academic performance (GPA), the at-school context conceptually overlapped with the criterion and therefore constituted the correct frame of reference, whereas the at-work context represented the incorrect frame of reference.

The second study focused on within-person inconsistency in the frame of reference adopted across items and examined its effect on reliability and validity, as hypothesized by the alternative explanation. To this end, we used a within-subject design in the second study. Again, GPA served as criterion measure. In this study, participants completed a personality inventory with both an at-school (correct) frame of reference and an at-work (incorrect) frame of reference.

Study 1

Method

Sample and Procedure

The sample consisted of 337 students who were in their final year in college. The participants were predominantly students majoring in law, economics, and social sciences. Of the sample, 33% were male and 67% were female. Mean age was 22 years (SD = 2 years). Virtually all participants had considerable work experience (student jobs, summer jobs); the mean number of jobs held was 4.

Participants were recruited by an invitation e-mail for a preparation session on psychological testing and assessment. They could subscribe to either an Internet-based session or a paper-and-pencil session. At the start of the session, participants were given the explanation that the advantage of taking part in this session was that they could increase their experience with taking a variety of tests. Next, participants completed a series of psychological tests. Although these tests differed somewhat across administration modes (e.g., a cognitive ability test and a situational judgment test were included in the paper-and-pencil session, whereas an Internet-based in-basket was included in the Internet-based session), a short resume (assessing demographic variables and GPA) and a personality inventory were always included as the first two parts of the session. Of the various tests, only participants' responses to the personality inventory were used in this study. We examined whether the personality scale scores and validities differed across administration mode and found no significant differences. A couple of weeks later, participants received feedback about their test results via e-mail.

Design and Manipulations

Students were randomly assigned to three conditions. In the generic condition, participants did not receive instructions to complete the personality scales with a specific frame of reference in mind. They simply received the standard instructions. The two contextualized conditions (at school and at work) were operationalized by specific instructions provided to participants. Similar to Hunthausen et al. (2003), we did not add context tags after each item but added the context (either at work or at school) before and during the administration. For example, in the contextualized

at-school condition, participants were instructed to think about how they were at school when responding to each statement. When turning a page, a reminder was also inserted to emphasize that participants should think about how they behaved at school when responding to the items. The contextualized at-work condition was operationalized in a similar way.

As a manipulation check, we asked which frame of reference (general, at work, or at school) participants had used. Thirty-three (10%) participants were removed across the contextualized conditions because they had left the manipulation check item blank, indicated multiple frames of reference, or indicated a different frame of reference than the one of their designated condition. This reduced the sample to 304 participants.

Personality Scales

The Big Five personality traits were assessed with Goldberg's (1999) 50-item International Personality Item Pool. The scale is composed of 10 (positively worded or negatively worded) statements related to the respondent's standing on each of the Big Five traits: Conscientiousness, Openness to Experience, Emotional Stability, Extraversion, and Agreeableness. Respondents are asked to indicate how accurately each statement describes them, using a Likert-type scale ranging from 1 (*very inaccurate*) to 5 (*very accurate*). A composite score was computed by summing the items related to a given trait. Goldberg (1999) reported the mean coefficient alpha for each of the five scales (10 items each) as .84, indicating an acceptable degree of internal consistency. Our data were consistent with this finding, with alphas of .81, .76, .89, .88, and .84, respectively, for the five scales (see diagonal of Table 1).

Apart from the Big Five factors, we also measured two specific facets of Conscientiousness (i.e., Achievement Striving and Self-Discipline) because these facets have been found to be especially predictive of academic performance (e.g., De Fruyt & Mervielde, 1996; Schmit et al., 1995). Both of these Conscientiousness facets were measured with 10 statements taken from Goldberg's (1999) International Personality Item Pool. Response instructions were the same as for the Big Five factors.

Similar to Robie et al. (2000), we conducted a pilot study to check whether all 70 items used (50 Big Five items, 10 items associated with Achievement Striving, and 10 items related to Self-discipline) could be situated in general, work, and school contexts. As a result, 2 of the 70 items (both were Extraversion statements, namely, "I am the life of the party" and "I talk to a lot of different people at parties") were replaced by other Extraversion statements ("I make friends easily" and "I know how to captivate people") from the International Personality Item Pool that were relevant in different contexts.

Criterion

As our study was situated in an educational context, cumulative GPA served as the criterion measure. GPA was measured on a scale ranging from 0 to 4, with higher scores indicating better grades. Cumulative GPA was obtained from 282 (93%) participants by self-report (as part of the process of completing their resume). Although it would have been preferable to gather students' GPAs through university records, it has been demonstrated that self-reported GPA and GPA obtained from university records are highly correlated (.91; see Schmitt et al., 2003). With respect to the reliability of this criterion, GPA correlated strongly across years, with correlations between GPAs across years around .70. These values are similar to the values found in a meta-analysis on the temporal stability of GPA (Vey et al., 2003).

Results and Discussion

Table 1 presents the descriptive statistics of Study 1 variables collapsed across all conditions. The broad factor of Conscientiousness (.19) and the two Conscientiousness facet scales (.22 and .18 for Achievement Striving and Self-Discipline, respectively) were significant predictors. These results are consistent with previous research on the validity of personality scales in educational settings (e.g., De Fruyt & Mervielde, 1996; Schmit et al., 1995). In addition, the correlation between Achievement Striving and Conscientiousness was .58. Self-Discipline correlated .70 with Conscientiousness. The intercorrelation between Achievement Striving and Self-Discipline was .73. These intercorrelations are in line with the meta-analytic values reported in Dudley, Orvis, Lebiecki,

Table 1
Descriptive Statistics of Study 1 Variables

Variable	*M*	*SD*	α	1	2	3	4	5	6	7
Predictor										
1. Achievement	39.17	5.13	.82							
2. Self-discipline	35.19	7.33	.91	.73**						
3. ES	35.13	7.03	.89	.06	.20**					
4. O	35.01	6.48	.76	.25**	.16**	.26**				
5. E	36.67	4.75	.88	.25**	.07	.12**	.28**			
6. A	41.50	4.63	.84	.30**	.15**	−.02	.22**	.16**		
7. C	38.70	5.44	.81	.58**	.70**	.05	.02	.02	.26**	
Criterion										
8. GPA	1.55	0.61		.22**	.18**	.01	−.03	.04	−.02	.19**

Note. $N = 304$ for the correlations among the predictors; $N = 282$ for the correlations with the criterion. ES = Emotional Stability; O = Openness to Experience; E = Extraversion; A = Agreeableness; C = Conscientiousness; GPA = grade point average. Internal consistencies are shown on the diagonal. For the criterion, the temporal consistency is given on the diagonal.
** $p < .01$.

Table 2
Means, Standard Deviations, and Effect Sizes of Study 1 Variables Broken Down by Condition

Variable	Generic (n = 115)		At school (n = 91)		At work (n = 98)		p	Partial eta squared
	M	*SD*	*M*	*SD*	*M*	*SD*		
Achievement	38.64	4.98	37.99	5.24	40.90	4.79	.00	.06
Self-discipline	34.61	7.07	32.69	7.01	38.19	6.94	.00	.09
ES	34.29	7.61	34.78	7.23	36.45	5.94	.07	.02
E	34.02	6.32	34.57	6.74	36.58	6.18	.01	.03
O	36.87	5.23	35.82	4.71	37.23	4.08	.11	.01
A	42.15	4.74	40.57	4.68	41.59	4.36	.05	.02
C	38.37	5.20	37.76	5.42	39.97	5.56	.01	.03
GPA	1.56	0.58	1.53	0.64	1.54	0.64	.91	.00

Note. Values of significance test and partial eta squared are obtained from analyses of variance. ES = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; GPA = grade point average.

and Cortina (2006). Thus, the current results show that the personality inventory performed in a manner similar to that expected.

Table 2 presents the means and standard deviations of the personality scales broken down by condition. For all personality scales, a small effect size was observed (partial eta squared varying from .01 to .09). With the exception of Agreeableness, the work-related frame of reference produced the highest scale scores. This confirms the results of Schmit et al. (1995) that the use of specific contexts in personality scales restricts the appropriate behavior to be elicited as compared with a generic context.

Table 3 presents the internal consistency coefficients (Cronbach alphas) across conditions. To determine whether the Cronbach alphas were significantly different across conditions, we used Feldt and Ankenmann's (1998) test. There were no significant differences between the internal consistencies in the generic condition and the consistencies in the contextualized conditions for any of the personality scale scores.

Criterion-related validities of the personality scales broken down per condition are presented in Table 3. Criterion-related validities were highest for the group of participants who were randomly assigned to the at-school frame of reference (correct frame of reference). The differences between the at-school frame of reference and the other conditions were largest for conceptually relevant traits (i.e., traits that have been found to be related to the criterion of academic performance), namely, Conscientiousness (.37 vs. .09 and .16), Achievement Striving (.41 vs. .16 and .12), and Self-Discipline (.34 vs. .16 and .06). For the other traits, there were no significant differences across the conditions.

Study 1 had three key findings. First, reduction of between-person variability (by imposing a frame of reference) did not affect internal consistency reliability. The reliability of the contextualized personality inventories was not higher than that of the generic personality inventory. This result is not in line with the traditional explanation. Instead, it supports the alternative explanation, which posited that reliability would not be affected as long as individuals are consistent within themselves and as long as the frames of reference used have comparable reliabilities. Second, reliabilities did not differ significantly across conditions, whereas validities did differ significantly. Therefore, reduction of between-person variability cannot be the only explanation for the validity increase of contextualized personality inventories. This conclusion also contradicts predictions of the traditional explanation. Third, reducing between-person variability by imposing a frame of reference on participants had beneficial effects only on the contextualization that matched the criterion (in this case, the at-school contextualization) for conceptually relevant broad factors (Conscientiousness) and conceptually relevant facets (Self-Discipline and Achievement Striving). This highlights the importance of reducing

Table 3
Internal Consistencies and Criterion-related Validities of Personality Scale Scores Broken Down by Condition

Variable	Reliability			Validity		
	Generic (n = 115)	At school (n = 91)	At work (n = 98)	Generic (n = 110)	At school (n = 83)	At work (n = 89)
Achievement	$.79_a$	$.82_a$	$.85_a$	$.16_b$	$.41_a$	$.12_b$
Self-discipline	$.89_a$	$.90_a$	$.91_a$	$.16_a$	$.34_a$	$.06_b$
ES	$.91_a$	$.90_a$	$.86_a$	$.06_a$	$.09_a$	$-.13_a$
E	$.88_a$	$.89_a$	$.87_a$	$.04_a$	$.06_a$	$-.19_a$
O	$.82_a$	$.76_a$	$.70_a$	$.12_a$	$.03_a$	$-.06_a$
A	$.84_a$	$.84_a$	$.84_a$	$-.02_a$	$.05_a$	$-.09_a$
C	$.78_a$	$.79_a$	$.84_a$	$.09_b$	$.37_a$	$.16_a$

Note. Alphas with different subscripts in the same row indicate significant differences across conditions at $p < .05$. These were computed on the basis of the test in Feldt and Ankenmann (1998, p. 171). Correlations with different subscripts in the same row indicate significant differences across conditions at $p < .05$. ES = Emotional Stability; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness.

between-person variability (a measurement issue) in such a way that test takers interpret items with a frame of reference that conceptually overlaps with the criterion (a substantive issue).

As the results of Study 1 were consistent with the alternative explanation, Study 2 focused entirely on the alternative explanation, examining whether the two proposed sources of variance (between-person variability and within-person inconsistency) had the hypothesized effects on reliability and validity. Given that within-person inconsistencies are now considered as well, a within-subject design was used.

Study 2

Method

Sample and Procedure

The sample consisted of 105 industrial and psychology students who participated in Study 2 for course credit. These students were in either their third or their fourth year of college. Among the participants, 30% were male and 70% were female. Mean age was 21 years (SD = 1 year). All students had considerable work experience (student jobs, summer jobs). The mean number of jobs was 4. A couple of weeks following participation, participants received feedback about their test results via e-mail.

Design and Manipulations

Participants completed the personality scales with both an at-school and an at-work frame of reference. These frames of reference were operationalized by adding context tags after each item (see Schmit et al., 1995). In light of possible item order effects (see Knowles, 1988) four different versions (with item order randomly determined) were created.

Personality Scales

As the personality inventory would become very long if all scales were rated with both an at-school and an at-work frame of reference, Study 2 focused on the three scales (Conscientiousness, Achievement Striving, and Self-Discipline) that were valid predictors in academic settings in general and in Study 1 in particular. The items of these three scales and the Likert-type rating scale were exactly the same as in Study 1. Therefore, all participants completed 60 items: 30 items with an at-school context and the same 30 items with an at-work context. Per trait, a composite score was computed by summing the items related to a given frame of reference. Thus, two composite scores (at work and at school) were computed per individual and per trait.

Criterion

Similar to Study 1, cumulative GPA served as the criterion measure. GPA was measured on a scale ranging from 0 to 4, with higher scores indicating better grades. Cumulative GPA was obtained from all participants by self-report. In line with Study 1, GPA correlated strongly across years (around .70).

Results and Discussion

Descriptive Statistics

Table 4 presents the descriptive statistics of Study 2 variables. Similar to Study 1, use of an at-work frame of reference produced significantly higher ratings than use of an at-school frame of reference. The intercorrelation among the two frames of reference was .46 ($p < .01$) for Achievement Striving, .22 ($p < .05$) for Self-Discipline, and .49 ($p < .01$) for Conscientiousness. Additionally, a multivariate analysis of variance showed no significant multivariate effect of item order, $F(18, 272) = 1.13$, *ns,* Wilks's lambda = .82.

Tests of Effects of Between-Person Variability

As shown in Table 4, there were significant differences in the internal consistencies across the frames of reference used for all three scales. Specifically, the scales rated with an at-school frame of reference had significantly higher Cronbach alphas than did the same scales rated with an at-work frame of reference. Significant differences were determined on the basis of Feldt's (1980) test. These significant differences support neither the traditional nor the alternative explanation, because both explanations posit that the effects of frame of reference on reliability do not depend on the

Table 4
Descriptive Statistics of Study 2 Variables

Variable	*M*	*SD*	α	1	2	3	4	5	6
Predictors									
1. Achievement, at work	39.62	5.20	$.85_a$						
2. Achievement, at school	33.00	7.29	$.90_b$	.46					
3. Self-discipline, at work	38.68	5.36	$.85_a$	.74	.34				
4. Self-discipline, at school	28.00	7.53	$.91_b$	.26	.72	.22			
5. Conscientiousness, at work	38.52	5.25	$.79_a$	.68	.35	.79	.26		
6. Conscientiousness, at school	33.29	7.27	$.87_b$	.44	.75	.41	.76	.49	
Criterion									
7. GPA	0.84	0.60		$.29_a$	$.53_b$	$.13_a$	$.41_b$	$.05_a$	$.38_b$

Note. $N = 105$. Alphas with different subscripts across the at-school and at-work conditions indicate significant differences at $p < .05$. These were computed on the basis of the test in Feldt (1980). Correlations between the predictor and the criterion with different subscripts across the at-school and at-work conditions indicate significant differences across conditions at $p < .05$. These were computed on the basis of the Z test in Meng, Rosenthal, and Rubin (1992).

frame of reference used. The significantly higher reliability of the at-school frame of reference might stem from students being more familiar with the at-school context and therefore possessing more autobiographical memories of this context. A reason for the finding of reliability differences in Study 2, and not in Study 1, might be the stronger manipulation used in Study 2, as the within-subject design of that study implied that context tags were added directly after each item (as in Schmit et al., 1995), and students were asked to complete all items twice with differing frames of reference. Conversely, in Study 1, the context tag was mentioned only before and during the administration of the items, and participants rated each item once (as in Hunthausen et al., 2003).

According to the propositions of the alternative explanation, between-person variability in the frame of reference adopted is expected to impact on validity. In support of this explanation, the validities of the conceptually relevant frame of reference (at school) were significantly higher than the validities of the incorrect frame of reference (at work). According to Meng, Rosenthal, and Rubin's (1992) *Z* test for the difference between dependent correlations, this difference was statistically significant for Achievement Striving ($Z = -2.64$, $p < .01$), Self-Discipline ($Z = -2.48$, $p < .05$), and Conscientiousness ($Z = -3.62$, $p < .01$). Generally, these results are consistent with the alternative explanation and with the results of Study 1.

We also simulated the effects of the degree of between-person variability in responding to generic items by randomly drawing samples from the total sample ($N = 105$) without replacement. These random samples always consisted of the same 105 test takers. However, they differed in terms of the percentages of test takers who used a specific frame of reference. For example, one might draw a random sample consisting of the ratings of 10% of the sample on the at-school items (correct frame of reference) and the ratings of 90% of the sample on the at-work items (incorrect frame of reference). Table 5 summarizes the reliability and validity effects of different scenarios of between-person variability in the frame of reference used. For each scenario, 1,000 random samples of $N = 105$ were drawn. Generally, the effects of between-person variability on validity were much larger than those on reliability. For instance, the validity of Self-Discipline varied between −.05 and .53, whereas its reliability varied between .85 and .95. Thus, consistent with Study 1, these results support the alternative, rather than the traditional, explanation.

Test of Effects of Within-Person Inconsistency

The alternative explanation posited that there would be a curvilinear relationship between the number of items rated with a specific frame of reference and reliability. The within-subject design of Study 2 enabled examination of this hypothesis because all individuals completed all items with both the at-school and at-work frames of reference. Hence, it is possible to simulate the reliability and validity effects of the degree of within-person inconsistency by randomly sampling from the responses provided by the participants of the total sample ($N = 105$) without replacement. Thus, these randomly drawn samples differed in terms of the number of items rated with a specific frame of reference. Table 6 summarizes the reliability and validity effects of different scenarios of within-person inconsistency. For instance, the second row shows the results averaged across 1,000 random samples of $N = 105$, wherein all individuals rated two (randomly chosen) items with an at-school (correct) frame of reference and the eight remaining items with an at-work (incorrect) frame of reference.

As shown in Table 6, a curvilinear pattern was apparent in the relationship between the number of times that individuals used the same frame of reference and reliability. That is, reliability was highest when all individuals rated either a small number of items with one specific frame of reference or a large number of items with one specific frame of reference. Reliability was lowest when all individuals rated about half of the items with one frame of reference and the other half with another frame of reference. Although this dip in reliability was observed across all three scales, it was most noteworthy for Self-Discipline.

A different pattern was found for the effects of within-person inconsistency in the frame of reference adopted on validity. According to the alternative explanation, the effects on validity would be positively related to the number of items rated with the correct

Table 5
Summary of Reliability and Validity Results of Different Levels of Between-Persons Variability

	Reliability			Validity		
Scenarios[a]	Achievement	Self-discipline	C	Achievement	Self-discipline	C
10% participants at school, 90% at work	.87	.89	.82	.31	.15	.11
20% participants at school, 80% at work	.89	.91	.83	.31	.15	.11
30% participants at school, 70% at work	.90	.92	.84	.35	.22	.18
40% participants at school, 60% at work	.90	.92	.85	.38	.22	.22
50% participants at school, 50% at work	.90	.93	.86	.36	.20	.21
60% participants at school, 40% at work	.91	.93	.86	.40	.24	.26
70% participants at school, 30% at work	.91	.92	.86	.42	.26	.28
80% participants at school, 20% at work	.91	.92	.87	.45	.31	.30
90% participants at school, 10% at work	.91	.91	.87	.49	.35	.34
M	.90	.92	.85	.39	.23	.22
SD	.02	.01	.02	.07	.08	.08
Minimum	.86	.88	.80	.25	.09	.07
Maximum	.93	.94	.89	.53	.45	.39

Note. C = Conscientiousness.
[a] For each of the nine scenarios, 1,000 random samples were drawn. Summary statistics at the bottom are computed across all 9,000 samples.

Table 6
Summary of Reliability and Validity Results of Different Levels of Within-Person Inconsistency

Scenarios[a]	Reliability			Validity		
	Achievement	Self-discipline	C	Achievement	Self-discipline	C
1 at-school item, 9 at-work items	.81	.76	.76	.34	.18	.10
2 at-school items, 8 at-work items	.78	.69	.74	.38	.24	.15
3 at-school items, 7 at-work items	.77	.64	.73	.42	.28	.19
4 at-school items, 6 at-work items	.76	.63	.73	.46	.32	.23
5 at-school items, 5 at-work items	.77	.65	.74	.48	.36	.27
6 at-school items, 4 at-work items	.79	.69	.77	.50	.38	.30
7 at-school items, 3 at-work items	.82	.74	.79	.51	.40	.33
8 at-school items, 2 at-work items	.85	.80	.82	.52	.41	.35
9 at-school items, 1 at-work item	.88	.85	.84	.52	.41	.36
M	.80	.72	.77	.46	.33	.25
SD	.04	.08	.04	.06	.08	.09
Minimum	.70	.53	.68	.28	.12	.04
Maximum	.89	.87	.86	.57	.47	.41

Note. C = Conscientiousness.
[a] For each of the nine scenarios, 1,000 random samples were drawn. Summary statistics at the bottom are computed across all 9,000 samples. Per sample, the specific items rated with a frame of reference were not fixed across participants.

frame of reference. Table 6 shows that validity was lowest when a correct frame of reference was used for only a limited number of items, whereas it was highest when a correct frame of reference was used for a large number of items. Consistent with the alternative explanation, these results highlight that the amount of conceptual overlap between the predictor and the criterion (as indicated by the number of items rated with the correct frame of reference) is a key determinant of validity.

Structural Equation Modeling Analyses

We also used structural equation modeling to test the models associated with our two explanations. In these analyses, we were especially interested in examining which paths of these models (see Figure 1 and 2) were affected by the hypothesized sources of inconsistency. First, we conducted multigroup confirmatory factor analyses (CFAs) to test the measurement invariance of the model related to between-person variability (Figure 1) across two groups: a group of test takers who rated items with an at-work frame of reference and a group who rated items with an at-school frame of reference. To this end, data from both Study 1 and Study 2 were used. Across all analyses (available from Filip Lievens), only minor departures of measurement invariance were observed. For some traits, the path coefficient to GPA was found to be noninvariant across groups. These structural equation modeling analyses do not support the traditional explanation, as reliability was not affected. Between-person variability had effects only on validity.

A similar strategy was applied to test which paths were affected by within-person inconsistency (see model in Figure 2). Multigroup CFAs were conducted across two samples: (a) a randomly drawn sample wherein test takers rated 5 of the 10 at-school items with an at-work frame of reference and 5 of the 10 at-work items with an at-school frame of reference (moderate within-person inconsistency; see Table 6) and (b) a sample wherein test takers rated all 10 at-school items with an at-school frame of reference and all 10 at-work items with an at-work frame of reference (no within-person inconsistency). These analyses could be conducted only with Study 2 data. Substantial departures of measurement invariance were observed. For all traits, the factor covariance was found to be noninvariant. In particular, in the sample with moderate within-person inconsistency, the covariance between the two factors was much higher. In addition, 10 error variances related to both factors were found to be noninvariant, as they were much higher in the sample with moderate within-person inconsistency. These results show that the two factors became much more error-laden, as they were both determined by 5 at-work and 5 at-school items. Thus, consistent with the alternative explanation, within-person inconsistency had a major impact on reliability. The measurement error caused by within-person inconsistency translated to the substantive relationships among the factors and GPA. Results obtained with other levels of within-person inconsistency (available from Filip Lievens) confirmed this conclusion.

Taken together, results of Study 2 confirmed and refined many of the findings of Study 1. We found that reducing between-person variability through imposing a frame of reference does not lead to higher reliability. In addition, reduction of between-person variability through a frame of reference had beneficial effects on validity only when a correct frame of reference was imposed on a large number of test takers. In addition, Study 2 extended the findings of Study 1 by investigating the impact of within-person inconsistency on reliability and validity. Within-person inconsistency affected the reliability of personality scales. In turn, this had an effect on the validity of these scales. Reliability was highest when test takers interpreted a large number of items with a specific (either correct or incorrect) frame of reference. Validity was related to test takers' interpreting a large number of items with a correct frame of reference.

General Discussion

This article adds several key findings to the literature on the use of a frame of reference in personality scales. Most importantly, it contributes to our understanding of why imposing a specific frame of reference might lead to increased validity. In addition, factors

that might impact on the reliability and validity of noncontextualized and contextualized personality scales were revealed.

First, this article shows that imposing a frame of reference in personality inventories enables the reduction of within-person inconsistency in responding to generic items. Reliability was highest when respondents interpreted a large number of items with the same frame of reference. Conversely, it was lowest when respondents switched their frame of reference for many of the items. Study 2 also provided evidence that reducing within-person inconsistency had larger effects on reliability than reducing between-person variability in item responding. This was most clearly shown by our additional structural equation modeling analyses. In addition, when the degree of within-person inconsistency was varied, reliabilities differed from .53 to .87 for Self-Discipline, whereas when the degree of between person variability was varied, reliabilities ranged from .85 to .95. In Study 1, between-person variability had negligible effects on reliability. Taken together, these reliability results shed a new and different light on the source of measurement error that is reduced by using a frame of reference. In particular, our findings highlight the importance of using a frame of reference as a vehicle for reducing within-person inconsistency, as within-person inconsistency makes the traits much more error-laden. Our results do not support the traditional explanation that use of a frame of reference reduces between-person variability.

As a second contribution, this study provides insight into factors that might explain why the validity of contextualized personality inventories is higher than that of their noncontextualized counterparts. The reduction of both between-person variability and within-person inconsistency seems to play a role. For instance, in both studies, reducing between-person variability led to a significant increase in validity. Similarly, in Study 2, reduction of within-person inconsistency increased validity. However, simply imposing a frame of reference is not enough. It is equally important to ensure that test takers adopt a frame of reference that conceptually overlaps with the criterion. This is evidenced by the fact that validity was increased in both studies only when between-person variability was reduced in such a way that the contextualization matched the criterion (in this case, the at-school contextualization) for conceptually relevant traits. Likewise, in Study 2, reducing within-person inconsistency increased validity when respondents interpreted a large number of items with a correct frame of reference.

More generally, these results confirm Schmidt et al.'s (2000) treatment of reliability and validity. According to Schmidt et al., reliability refers to a measurement model, specifically, to consistency in measurement (regardless of what is measured). Conversely, validity refers to a substantive process model. This study provides evidence that reducing between-person variability in the frame of reference used does not suffice to obtain high validity, as attested by the consistently low validity for the at-work frame-of-reference condition. In order to increase validity, it is also important to reduce between-person variability in such a way that the contextualization imposed matches the criterion. The same reasoning applies to the within-person inconsistencies. Reducing those inconsistencies in the frame of reference adopted is not enough for obtaining high validity. For example, Table 6 (first rows) shows that high reliability might be coupled with low validity even when frames of reference are imposed. When one reduces within-person inconsistency, it is also key to ensure that test takers interpret items with a frame of reference that is conceptually relevant to the criterion. Thus, it is important to reduce measurement error in such a way that the conceptual meaning of the latent factor better matches the criterion. Only when this substantive issue is taken into account is validity increased.

With regard to substantive issues, the large difference in the validity coefficients for the two frames of reference (at school vs. at work) deserves some attention. One explanation might be that for students, these two frames of reference represent radically different situations and perspectives. This is confirmed by the moderate correlation among the two frames of reference in Study 2. As another explanation, the primary reason for the correlation between Conscientiousness-related factors at school and GPA might be direct causality, whereas the primary reason for the correlation between Conscientiousness-related factors at work and GPA might be that Conscientiousness-related factors at work correlate strongly with Conscientiousness-related factors at school, which, in turn, causes GPA. If this were true, then the correlation between Conscientiousness-related factors at work and GPA would equal the correlation between the two contextualizations of Conscientiousness-related factors multiplied by the correlation between Conscientiousness-related factors at school and GPA. Inspection of Table 3 shows that this is the case for both Achievement Striving ($.46 \times .53 = .24$, whereas the value is .29 in Table 4) and Self-Discipline ($.22 \times .41 = .09$, whereas the value is .13 in Table 4). Future research should delve deeper into these explanations.

Third, our studies are the first to actually show that the use of an incorrect frame of reference decreases validity. Moreover, both studies demonstrate that validity might vary to a large extent depending on the frame of reference used. For instance, in Study 2, the validity of Conscientiousness was as low as .05 and as high as .38. These are important findings because they exemplify what might happen if candidates self-contextualize generic items and (sometimes) use an incorrect frame of reference for answering them. Accordingly, our findings might constitute one explanation for the low to moderate validities of generic personality inventories. These findings also reinforce the importance of using contextualized inventories.

This study is not without limitations. First, its generalizability to an employment context might be questioned. A laboratory setting was used because it is difficult to conduct the manipulations (i.e., an incorrect frame of reference) in the field. If we would have asked actual applicants to use an incorrect frame of reference, perhaps this might have led to legal challenges. In a similar vein, it should be acknowledged that GPA served as the criterion. Therefore, future research should examine whether our results generalize to employment settings, with job performance as the criterion. It is worth noting, however, that much validation work in employment settings is against training criteria. As another limitation, our sample consisted of a restricted group of students. It can be expected that these students were high on Conscientiousness, as they had already successfully passed several academic years. In Study 1, they had even self-selected to attend the test preparation session. Despite this possible range restriction, differential validity effects were found.

The results of the current study prompt various intriguing directions for future research. First, it is important to better under-

stand how respondents complete generic personality inventories. In this article, we assumed that generic items led to between-person variability and within-person inconsistency in the frame of reference adopted. However, we have no information about the prevalence of these inconsistencies. We also showed what might happen if test takers switched between two frames of reference. Apart from the at-school frame of reference, we selected the at-work frame of reference because this context is also relevant for students. However, existing research provides no insight into how many frames of reference test takers typically use. To shed light on these issues in responding to generic items, think-aloud or policy-capturing studies might be fruitfully conducted. In a similar vein, mixture item response theory models might be used to group test takers in latent classes (Eid & Langeheine, 1999; Zickar et al., 2004). Alternatively, Monte Carlo simulations might be conducted to examine the impact of multiple frames of reference (which vary in their intercorrelation, measurement error, etc.) on reliability and validity.

Second, one might wonder whether further increasing the contextualization of personality scales might lead to even higher criterion-related validities. According to this perspective, adding an at-work or an at-school tag is just the beginning. On the basis of the behavioral consistency model that is typically used in sample-based predictors, one might expect that adding more context would lead to even higher criterion-related validities. Alternatively, one might posit that the level of specificity in the predictor should be mapped with the level of specificity in the criterion. This would mean that using a narrower context in the predictor limits the generality of the predictions made. Although these questions are related to the bandwidth-fidelity trade-off (Cronbach, 1960), they could invoke some intriguing future research.

Third, one might also contrast this recent move to contextualized personality scales to a reverse development in sample-based predictors, such as situational judgment tests. In fact, there is growing interest to develop construct-oriented situational judgment tests (Motowidlo, Hooper, & Jackson, 2006). These tests present a job-related situation to candidates. As compared with contextualized personality inventories, this situation is much more detailed. Yet, similar to personality inventories and contrary to traditional situational judgment tests, the response alternatives are carefully developed to reflect different degrees of a given construct (e.g., Agreeableness). Situation-response inventories (e.g., Born, 1994; Furnham & Jaspars, 1983) constitute another alternative format that has some correspondence with contextualized personality inventories. Future research should compare the validity of these formats while holding the construct measured constant.

In conclusion, prior research ascribed the increase in reliability and validity of contextualized personality inventories to a reduction of inconsistencies among test takers. However, it was unclear what type of inconsistency was reduced. The current study advances prior research on the use of a frame of reference in personality inventories by testing two competing explanations. Results confirmed the alternative explanation, showing that reliability was affected by within-person inconsistency and that criterion-related validity was affected by between-person variability (the number of test-takers using a correct frame of reference) and within-person inconsistency (the number of items rated with a correct frame of reference) in item responding.

References

Aronson, Z. H., & Reilly, R. R. (2006). Personality validity: The role of schemas and motivated reasoning. *International Journal of Selection and Assessment, 14,* 372–380.

Asendorpf, J. B., Borkenau, P., Ostendorf, F., & van Aken, M. A. G. (2001). Carving personality description at its joints: Confirmation of three replicable personality prototypes for both children and adults. *European Journal of Personality, 15,* 169–198.

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.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Born, M. P. (1994). Development of a situation–response inventory for managerial selection. *International Journal of Selection and Assessment, 2,* 45–52.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey Bass.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Row.

De Fruyt, F. (2002). A person-centered approach to P-E fit questions using a multiple-trait model. *Journal of Vocational Behavior, 60,* 73–90.

De Fruyt, F., & Mervielde, I. (1996). Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality, 10,* 405–425.

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.

Eid, M., & Langeheine, R. (1999). The measurement of consistency and occasion specificity with latent class models: A new model and its application to the measurement of affect. *Psychological Methods, 4,* 100–116.

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

Feldt, L. S., & Ankenmann, R. D. (1998). Appropriate sample size for comparing alpha reliabilities. *Applied Psychological Measurement, 22,* 170–178.

Furnham, A., & Jaspars, J. (1983). The evidence for interaction in psychology: A critical analysis of the situation–response inventories. *Personality and Individual Differences, 4,* 627–644.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *g*: Is this your final answer? *Human Performance, 15,* 123–142.

Gordon, E. D., & Holden, R. R. (1998). Personality test item validity: Insights from "self" and "other" research and theory. *Personality and Individual Differences, 25,* 103–117.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1990). Clinical reliabilities and validities of the microcomputerized basic personality inventory. *Journal of Clinical Psychology, 46,* 845–849.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice: The effects of frame-of-reference and pretest 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. B., & Hammer, B. L. (2003).

A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology, 88,* 545–551.

Knowles, E. S. (1988). Item context effects on personality scales. Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.

Kunda, Z., & Sanitioso, R. (1989). Motivated changes in the self-concept. *Journal of Experimental Social Psychology, 25,* 272–285.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90,* 442–452.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812–821.

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. *Journal of Personality Assessment, 78,* 348–369.

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

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing the invariances in personality and the role of situations. *Psychological Review, 102,* 246–268.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests.* San Francisco: Jossey-Bass.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Robie, C., Schmit, M. J., Ryan, A. M., & Zickar, M. J. (2000). Effects of item context specificity on the measurement equivalence of a personality inventory. *Organizational Research Methods, 3,* 348–365.

Rogers, T. B. (1974a). Analysis of stages underlying process of responding to personality items. *Acta Psychologica, 38,* 205–213.

Rogers, T. B. (1974b). Analysis of two central stages underlying responding to personality items—Self-referent decision and response selection. *Journal of Research in Personality, 8,* 128–138.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634–644.

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology, 59,* 229–241.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53,* 901–912.

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.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88,* 979–988.

Sorrentino, R. M., & Higgins, E. T. (1986). Motivation and cognition: Warming to synergism. In R. M. Sorrentino & E. T. Higgins (Eds.), *The handbook of motivation and cognition: Foundations of social behavior* (pp. 3–10). New York: Guilford Press.

Vey, M. A., Ones, D. S., Hezlett, S. A., Kuncel, N. R., Vannelli, J. R., Briggs, K. H., & Campbell, J. P. (2003, April). *Relationships among college grade indices: A meta-analysis examining temporal influences.* Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Warr, P. (2000). Indirect processes in criterion-related validity. *Journal of Organizational Behavior, 2,* 731–745.

Wright, J. C., & Mischel, W. (1987). A conditional analysis of dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53,* 1159–1177.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168–190.

Received August 2, 2006
Revision received July 6, 2007
Accepted October 4, 2007 ■