Applied HRM Research, Volume 11, Number 2, pages 107-118

Respondent Perceptions of Integrity and Personality Measures: Does Response Format Make a Difference?

Morell E. Mullins
Xavier University

Jaclyn M. Polson
Wright State University

Trisha Lanch & Kathryn Kehoe
Xavier University

Two studies are reported that examined the effect on respondent reactions of modifying the number of points on a response scale. Results indicate that for an integrity test, respondents who completed a measure using a 5-point Likert scale perceived themselves to have performed better on the test and perceived the measure to be significantly more fair and more face valid than respondents who completed an identical scale using a dichotomous response format. No such differences, however, emerged between dichotomous, 5-point, and 9-point versions of a personality inventory, indicating that although respondent reactions may be affected by the number of points in a response scale, some types of personnel selection measures may be more affected by number of response points than others.

In the past two decades, much research has emerged on applicant reactions to a variety of employee selection methods (Anderson, 2004). One area that has received relatively little attention is the impact methodological decisions may have on applicant reactions (Chan & Schmitt, 2004) and the extent to which the impact of those methodological factors may depend on the content of the test. Chan and Schmitt (1997) examined test medium as a potential factor, but to our knowledge, specific methodological factors other than those related to technology (e.g., Bauer, Truxillo, Paronto, Weekley, & Campion, 2004; Chapman, Uggerslev, & Webster, 2003) have not generally been focal to research on applicant reactions. However, other micro-level decisions made during the construction of an assessment tool may also affect respondent reactions. In the current study, we sought to understand whether reactions to different types of tests differs on the basis of whether responses are made on a dichotomous (true/false) scale or on a scale with more than two response options.

Study One: Integrity

Despite significant relationships between integrity scores and work behaviors (e.g., Ones, Viswesvaran, & Schmidt, 1993; Sackett & Wanek, 1996), test takers often perceive integrity tests as being unrelated to the job (lacking face validity) and unfair (Rosse, Ringer, & Miller, 1996; Rynes & Connerley, 1993; Steiner &

Gilliland, 1996). Although researchers hypothesized that these negative perceptions of integrity tests were the result of the invasiveness or offensiveness of the personal questions, little or no support has been found for this claim (Neuman & Baydoun, 1998). Therefore, the items themselves may not be causing negative reactions, but the response format of the test might.

Research indicates that negative reactions to testing may partially result from the response format of the test, and that the available range of response can influence assessment fairness perceptions. Specifically, participants responding to attitudinal scales reported feeling unable to give an accurate answer when being assessed with a scale with fewer than 5 points (Chicchetti, Showalter, & Tyrer, 1985), preferring scales with 5-9 points, and feeling that scales with 2-3 points provided inaccurate assessments (Cox, 1980). That the response format of integrity tests is often dichotomous (yes/no, true/false; Rudner, 1992) suggests a potential problem. If integrity items with dichotomous response formats lead to perceived inaccurate assessment, this may lead to lower perceptions of likely performance on the test, which have been shown to be related to perceptions of face validity (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998; Macan, Avedon, Paese, & Smith, 1994; Rynes & Connerly, 1993). Perceptions of low performance may in turn result in perceptions of low face validity. Research has demonstrated that perceptions of face validity influence perceptions of instrument fairness (Smither, Millsap, Stoffey, Reilly, & Pearlman, 1996). Hence, we hypothesized that individuals who received an integrity test with a 5-point response scale would perceive the test to be significantly more fair and face valid and would perceive themselves to have performed better than individuals who received a content-identical scale with a dichotomous response format.

Method

Participants

Eighty-six undergraduate students at a small Midwestern university took part in the study. Of the participants, 79% were women, 87.2% were white, 9.3% were African American/Black, and 3.5% were Hispanic/Latino. Participants ranged in age from 17-56. Although participants were not asked about their work history, 24.4% of the sample were above age 25 and as such could be expected to possess at least some work experience.

Materials

A pencil and paper integrity test was created by examining a published integrity test and writing items of similar overt content; all items were written to ensure that they tapped relevant content domains, including stealing, lying, or various illegal activities, and were consistent with the thematic composites identified by Wanek, Sackett, and Ones (2003). The final survey contained 30 items derived in this manner. A sample item is as follows: "Workers steal due to being unsatisfied with their company." The response scales and scoring key were constructed for the current study.

Fairness perceptions, face validity, and perceived performance were assessed using a 9-item measure adapted from Chan et al. (1998). This measure used a 5-point response scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

Procedure

A between-subjects design with two groups (dichotomous vs. 5-point) was used. Each participant read an identical scenario describing the application process of a hypothetical employer. The scenario read as follows: "Suppose you are seeking employment as a manager at a manufacturing company. Duties include supervising employees, devising more efficient operations, maintaining a budget, and inventory control. The survey that you are being asked to complete is being used as part of your application process." Participants then completed the integrity test. Half of the participants received a form of the integrity test for which they were instructed to "indicate whether you believe the following statements are TRUE or FALSE by writing EITHER True OR False" in the blank provided. The other half of the participants received a content-identical integrity test but were asked to indicate their level of agreement with each integrity item utilizing a 5-point Likert scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Once participants completed the integrity test, they completed the reaction measures and the demographic questions.

Results

Table 1 includes descriptive statistics for the three reaction variables across conditions, and Table 2 presents intercorrelations and internal consistency reliability estimates for all study variables. The study hypothesized that participants completing the integrity test with the 5-point scale would report higher perceptions of fairness, face validity, and performance than participants completing the integrity test with the dichotomous format.

Table 1
Descriptive Statistics for Study 1

Dimension	Scale Type	Mean	Standard Deviation
Fairness			
	Dichotomous scale	2.59	1.02
	5-point scale	3.74	0.87
Face Validity			
	Dichotomous scale	3.12	0.76
	5-point scale	3.55	0.73
Performance			
	Dichotomous scale	3.02	0.87
	5-point scale	3.81	0.66

Table 2
Intercorrelations and internal consistency reliability for Study 1

Variable	1	2	3	4	5
1 Fairness	(.86)				
2 Face validity	.53**	(.61)			
3 Performance	.63**	.43**	(.76)		
4 5-point integrity test	-.13	.05	.07	(.78)	
5 Dichotomous integrity test	-.29	-.27	-.29	--	(.72)

*$p < .05$, two tailed; **$p < .01$, two tailed.

Note. Coefficient alpha reliabilities are listed in parentheses

To test these hypotheses, a multivariate analysis of variance (MANOVA) was run, with test form as the independent variable and the three perceptual outcomes as dependent variables. The composite dependent variable was significantly affected by test form, F (3, 82) = 11.92, $p < .01$. Descriptive statistics for the independent variables (dichotomous scale, 5-point scale) on the dependent variables (perceived fairness, face validity, and performance) are presented in Table 1. An examination of the between-subjects effects from the MANOVA demonstrates that test form had a significant effect on each of the three dependent variables. Fairness perceptions (F (1, 84) = 31.83, $p < .01$), face validity (F (1, 84) = 7.30, $p < .01$), and perceived performance (F (1, 84) = 22.66, $p < .01$) all demonstrated significant effects based on the response scale utilized. Furthermore, all of the differences were in the hypothesized direction: Individuals who received the 5-point response scale perceived it to be significantly more face valid and fair and perceived themselves to have performed better than individuals who received the dichotomous response scale.

The pattern of correlations in Table 2 is interesting on a number of levels. First, a negative (but non-significant) relationship emerged between scores on the dichotomous integrity test and the perceived fairness, perceived face validity, and perceived performance on that test, indicating that individuals scoring higher on the measure of integrity potentially felt disadvantaged by the use of this test form. While it is true that these negative relationships were not statistically significant, it is worth noting that these correlations were computed for a sample of n = 43 participants and that correlations of the same magnitude were significant based on our full sample (n = 86). That scores on the 5-point integrity test were unrelated to perceptual variables seems to indicate that actual performance on the measure was relatively unrelated to reactions.

Discussion

Analyses indicated that range of response in answer formats influences perceptions of fairness, face validity, and performance regarding integrity tests. Although other studies have examined determinants of applicant reactions in personnel selection, to the best of our knowledge, no research has examined whether the range of response contributes to perceptions of unfairness. Results also revealed that performance perceptions on the test were positively related to perceptions of

both fairness and face validity. All three hypotheses were supported and in two of the cases, the effect size approached a full standard deviation difference in reactions between individuals who received the dichotomous form and those who received the 5-point form of the same measure.

The implications of the first study are apparent. Changing the response scale from a 5-point Likert scale to a dichotomous format not only resulted in predictable psychometric results ($\alpha = .78$ for the 5-point scale vs. $\alpha = .72$ for the dichotomous) but was also associated with respondents reporting more negative reactions in terms of perceived fairness, perceived face validity, and perceived performance.

However, the first study is rather narrow and rightly open to critique. Perhaps most critically, there is no way to know the extent to which the results obtained in this study are content-specific. That is, having demonstrated that respondent reactions are more negative when an integrity test utilizes a dichotomous format rather than format more scale points, does this finding generalize to other content domains? In the second study, we attempted to address this issue using a personality inventory rather than an integrity test.

Study Two: Personality

The second study sought to replicate the findings from study one in a different content domain. The hypotheses as such paralleled those from the first study. We expected that individuals who received a personality inventory with a 5-point response scale would perceive the test to be significantly more fair and more face valid and would perceive themselves to have performed better than individuals who received a content-identical scale with a dichotomous response format. In addition, because the personality inventory as originally published used a 9-point scale, we included a third condition in which participants received the measure as presented by Saucier (1994). We did not hypothesize differences between the 5-point and 9-point scales, though we did expect differences consistent with those from the first study to emerge between the personality inventory with the 9-point scale and the same inventory with the dichotomous response scale.

Method

Participants

Two hundred four undergraduate students at a small Midwestern university took part in the study, with 63 participants receiving the dichotomous form of the personality measure, 70 receiving the 5-point form, and 71 receiving the 9-point form. Of the participants, 53% were women, 83.4% were white, 8.8% were African American/Black, and 2% were Latino/Hispanic. Participants ranged in age from 18-40 years, with a mean of 20.5 years. In addition, because of potential generalizability concerns, participants in the second study were asked about their work histories. Of 204 participants, only nine reported never having held a job; 187 had held three or more jobs over the past five years, and 140 reported an average length of service of at least one year for their previous jobs. As such, although a

student sample was utilized, the data suggest that our sample possessed at least a minimal amount of work experience upon which to base their judgments of and reactions to the measures presented.

Materials

The Big Five Personality Factors were assessed using a 40-item adjective-based measure derived from a personality inventory reported by Saucier (1994). Three forms of the test were developed. The first form used a dichotomous response scale and asked respondents to "indicate whether you believe the following trait describes you by writing 'T' for True or 'F' for False in the blank provided". The second form utilized a 5-point response scale, with response points adapted from the original Saucier form of the measure and ranging from 1 (*Very Inaccurate*) to 5 (*Very Accurate*). The third form utilized the 9-point form of the response scale reported by Saucier, with the same terminal anchors as the 5-point. Internal consistency reliabilities for the five factors are presented in Table 3. The reaction measures used in the first study were again used, and demonstrated internal consistencies of $\alpha = .57$ (perceived fairness), $\alpha = .77$ (perceived face validity), and $\alpha = .70$ (perceived performance). All reaction measures were assessed on a 5-point Likert scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

Table 3
Internal consistency reliabilities for personality dimensions

Personality Dimension	Dichotomous	5-point	9-point
Extraversion	.82	.82	.93
Conscientiousness	.78	.79	.88
Agreeableness	.78	.82	.83
Openness to Experience	.64	.71	.81
Neuroticism	.72	.84	.84

Procedure

A between-subjects design with three groups (dichotomous vs. 5-point vs. 9-point) was used in which all groups were again presented with a scenario describing the application process of a hypothetical employer. The only difference between the instructions received by participants in the second study and those in the first was that the context of Study Two was a retail store rather than a manufacturer. All duties and responsibilities in the position description remained identical. Participants took the personality inventory, followed by the reaction measures.

Results

A multivariate analysis of variance (MANOVA) was run, with test form as the independent variable and the three perceptual outcomes as dependent variables. The composite dependent variable was not significantly affected by test form, F (6, 398) = 0.22, *ns*. Descriptive statistics for the three conditions on perceived fairness,

face validity, and performance are presented in Table 4. An examination of the between-subjects effects from the MANOVA demonstrates that test form had no significant effect on any of the three dependent variables. Follow-up pairwise comparisons did not alter the pattern of results: No differences emerged in reactions to this personality inventory as a function of the response scale utilized, nor did any consistent trend present itself in the mean reactions across the three conditions.

Discussion

In the case of the personality inventory, the response scale chosen appears to have had no effect on perceived fairness, face validity, or performance, associated with the adjective checklist form of the Big Five Personality Factors utilized in this study. The question that must be addressed is why the differences observed in Study One did not emerge in Study Two. The manipulation of the response scales was identical to that utilized in Study One, and in fact added another condition that had the potential (based on a comparison of a 9-point scale to the dichotomous form) to show stronger effects than those observed previously. It seems unlikely that the modification of the study's framing instructions (moving from a manufacturing context to a retail store) had an effect on reactions; if anything, this should have made issues of fairness *more* salient to our sample based on their age and experience level and, thereby, strengthened our manipulation.

It might also be suggested that the content of the Saucier (1994) personality measure lacked the ability to evoke negative reactions, independent of methodological manipulations. We believe that this is not the case, because the measure's 40 adjectives included words such as moody, inefficient, temperamental, and unintellectual, all of which have negative meanings and/or connotations and thus have the *a priori* potential for aversive reactions.

Table 4
Descriptive statistics for personality inventory perceptions

Dimension	Number of Scale Points	Mean	Standard Deviation
Fairness			
	2	3.11	.68
	5	3.17	.76
	9	3.14	.77
Face Validity			
	2	3.23	.97
	5	3.18	.94
	9	3.09	.94
Performance			
	2	3.85	.72
	5	3.82	.74
	9	3.86	.78

General Discussion

Although little attention has been paid to methodological factors as determinants of respondent reactions, the current study suggests that this could prove a fertile area for future research. In Study One, our results indicated that the response scale chosen for our integrity measure affected perceptions of fairness, face validity, and performance. In Study Two, the number of scale points did not affect reactions to a personality inventory.

Taken together, the results of our two studies indicate that the number of scale points *might* influence perceptions of fairness. From a practice perspective, we need to be wary when we adapt measures to suit our needs. Depending on the content of the test, how people respond – and thus the likelihood that we may face lawsuits – could be influenced by the response scale chosen. It is safest not to modify response scales for existing measures, or if this must be done in order to fit a measure into an existing test battery, practitioners are encouraged to conduct a validation study on the modified scale. It is possible that similar findings may emerge if we consider minor wording differences for scale anchor points (e.g., "Strongly Agree" to "Strongly Endorse"), which future research should consider.

From a theoretical perspective, the question of why the reactions change for some measures but not others needs to be further addressed. It may be, for example, that the "confrontational" nature of overt integrity items makes them more prone to evoke strong reactions. The personality items, which still utilize emotionally-charged words, may be less likely in general to evoke reactions. It may also be the case that the type of judgment being made could have an effect on reactions. The integrity items used in this study measured attitudes, and the attitudes that are relevant to integrity judgments have strong moral and ethical foundations, often dating back to childhood. The personality scale, on the other hand, was a self-description inventory, without any moral or ethical "baggage" attached. As such, any reactions might be less extreme than those related to strongly-held beliefs about lying, cheating, or stealing; it is possible, in other words, that the evaluation of who we perceive ourselves to be is less upsetting to us than the evaluation of what we believe to be "right".

Although the results of the studies are interesting and potentially relevant to both researchers and HR practitioners, they are not without flaws. The first deals with the focal measure for each study. In Study One, we utilized an integrity measure for which no construct validity evidence exists. The measure was created because, to the best of the authors' knowledge, no current, commercially-available integrity test is available with both dichotomous and 5-point response formats. As such, the integrity test was constructed specifically for this study to ensure that two forms of a test, identical at the item level, were used. Therefore, there is minimal support for the construct validity of the integrity test utilized. However, given that every item in the test dealt with attitudes or beliefs related to stealing, lying, generally dishonest behavior, or law-breaking, we believe it does sample the construct domain of interest and allows conclusions to be drawn about the effect response scale options had on reactions to the measure. Future research should

address whether our findings are replicable with tests that have empirically demonstrated construct validity, though again, such comparisons will be primarily of one construct-validated measure, and another (parallel) measure utilizing a different response scale.

In Study Two, we utilized Saucier's (1994) Mini-Markers, a 40-adjective tool to assess the Big Five personality factors. We recognize that the measure is meaningfully different than the questionnaire format found in the most popular tool used to assess the Big Five (the NEO-PI-R; McCrae & Costa, 1992) and therefore may not represent the "standard" way in which personality assessment is considered. However, research has demonstrated that Saucier's adjective-based measure does demonstrate some level of convergent validity with other questionnaire-based assessments of the Big Five (e.g., Palmer & Loveland, 2004). Further, because a body of research exists that has used the Saucier measure, the construct validity of the personality measure chosen seems solid. Whereas the first study might rightly be critiqued on those grounds were it to stand on its own, we believe the combination of the two studies clearly indicates that further exploration of the effect of scale format on respondent reactions should be undertaken.

Another limitation that must be noted is our use of a student sample. We do not believe this to substantially limit the value of our results for a number of reasons. First, the sample had work experience (assessed in Study Two, but only inferred in Study One), making their reactions to hiring measures relevant. Second, the manipulation in both studies attempted to create psychological fidelity by asking respondents to consider issues from a work context. Finally, in some respects, an attitudinal measure of integrity should measure the construct equally well for 60-year-olds and 20-year-olds, and a personality measure should be able to do the same. As such, we believe that the sample, although not ideal, in no way limits the inferences that should be drawn based on a study of methodological factors.

The current studies obviously represent a preliminary approach to the research topic. The implications of our findings, both for HR practitioners and for researchers in the personnel selection domain, should be given consideration whenever new measures are being constructed or existing measures are considered for modification. Perhaps the answer to the age-old question of how many points are needed for a response scale is "it depends on the construct being measured." For example, with a question such as. "Are you a psychology major?", a dichotomous scale of yes/no is clearly the most appropriate. With an integrity test, however, it may be that two points are not enough to tap a person's agreement about a statement, whereas two points may be enough for a personality inventory.

Though we would not suggest that dichotomous scales are uniformly "bad," we would definitely argue that the truncation in response range can negatively affect reactions to our measures for some constructs. The precise boundaries, however, remain unclear. We have speculated that the "confrontational" nature of integrity items might set them apart from other hiring devices; it might also be the case that while people are sometimes comfortable thinking of themselves in "black or white" terms when it comes to personality (e.g., we either perceive ourselves as "moody" or we do not), they may be uncomfortable assigning absolutes to issues relating to honesty – shades of grey may be something we need to respond to such questions in

a way that makes us comfortable. The precise explanation is one that we suggest future research consider in detail. For now, we offer our studies as a gentle warning: Based on our results, even the smallest change in the selection process can result in substantial, often unexpected, changes in respondent reactions.

References

Anderson, N. (2004). Editorial – The dark side of the moon: Applicant perspectives, negative psychological effects (NPEs), and candidate decision making in selection. *International Journal of Selection and Assessment, 12*, 1-8.

Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment, 12*, 135-148.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159.

Chan, D., & Schmitt, N. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment, 12*, 9-23.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, (4), 232-239.

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology, 88*, 944-953.

Chicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9*, 31-36.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715-738.

McCrae, R. R., & Costa, P. T. (1992). Discriminant validity of NEO-PIR facet scales. *Educational and Psychological Measurement, 52*, 229-237.

Neuman, G. A., & Baydoun, R. (1998). An empirical examination of overt and covert integrity tests. *Journal of Business and Psychology, 13* (1), 65-79.

Ones, D. S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications of personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Palmer, J. K., & Loveland, J. M. (2004). Further investigation of the psychometric properties of Saucier's Big Five "Mini-Markers:" Evidence for criterion and construct validity. *Individual Differences Research, 2*, 231-238.

Rosse, J. G., Ringer, R. C., & Miller, J. L., (1996). Personality and drug testing: An exploration of the perceived fairness of alternatives to urinalysis. *Journal of Business and Psychology, 10* (4), 459-474.

Rudner, L. M. (1992). Pre-employment testing and employee productivity. *Public Personnel Management, 21*, (2), 133-151.

Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology, 4*, 261-277.

Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty, conscientiousness, dependability, trustworthiness, and reliability for personnel selection. *Personnel Psychology, 49*, 787-829.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*, 506-516.

Smither, J. W., Millsap, R. E., Stoffey, R. W., Reilly, R. R., & Pearlman, K. (1996). An experimental test of the influence of selection procedures on fairness perceptions, attitudes about the organization, and job pursuit intentions. *Journal of Business and Psychology, 10*, 297-318.

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81*, 134-141.

Wanek, J. E., Sackett, P. R., & Ones, D. S. (2003). Towards an understanding of integrity test similarities and differences: An item-level analysis of seven tests. *Personnel Psychology, 56*, 873-894.

Author Notes

1. Portions of this paper were previously presented at the 2005 Society for Industrial and Organizational Psychology Annual Conference in Los Angeles, California.
2. The authors wish to thank the Xavier University Faculty Development Funds for the generous financial support of this project.
3. Correspondence and requests for reprints should be sent to the first author at Xavier University's Department of Psychology, 3800 Victory Parkway, Cincinnati, OH 45207-6511, or electronically at mullins@xavier.edu