Applied HRM Research, Volume 11, Number 2, pages 97-106

Measurement Equivalence of a Personality Inventory Administered on the Internet Versus a Kiosk

Chet Robie
Wilfrid Laurier University

Douglas J. Brown
University of Waterloo

This study examined the measurement equivalence of a personality inventory across Internet and kiosk modes of administration. A large nationwide sample of data was collected from a range of industries. The analyses indicated no evidence of differential item functioning across modes of administration. A higher percentage of men, Whites, and Asians chose the Internet mode. Those in the Internet sample were more likely to disclose their race/ethnicity than those in the kiosk sample. Scale reliabilities and intercorrelations appeared to be unaffected by mode of administration. Two of the three Big Five scales (conscientiousness and agreeableness) evidenced small mean differences across mode of administration. A ¼ standard deviation difference between the Internet and kiosk samples emerged for the emotional stability scale. Finally, we found no evidence of main or interactive (with mode) effects for race or gender on the personality scores. Implications for faking on personality tests and comparisons to previous findings in the area of measurement equivalence are discussed.

The present research was carried out to examine whether data obtained from an off-site Internet administration of a personality inventory was subject to any bias compared to data obtained from an on-site kiosk mode of administration. Kiosks are small stations that are located on the employers' premises. Kiosk stations are designed to allow respondents to enter responses to application material much like the Internet mode but without much of the necessary fundamental computer skills (e.g., windows operating system; web browsing). Kiosk stations are also ideally suited for walk-in candidates and those without access to computers.

Ensuring measurement equivalence across modes of administration is important so that an applicant would get the same score regardless of testing mode and the validity of the test is equivalent across testing modes. Additionally, the same norms and cutoffs can be used on measures that have been found to be equivalent.

Research has examined measurement equivalence of individual differences measures across various testing modes (Church, 2001; Ployhart, Weekley, Holtz, & Kemp, 2003). Church (2001) compared opscan (i.e., paper-and-pencil) with online and automated phone (i.e., IVR) responses on organizational survey items. Overall, survey method accounted for a relatively small percentage of unique variance in the data (0% to 4%). Ployhart et al. (2003) examined the measurement equivalence of personality, biodata, and situational judgment tests across proctored Web-based and paper-and-pencil modes. Relative to the applicants completing the paper-and-pencil measures, the Web-based measures showed: (a) better distributional properties, (b) lower means, (c) more variance, (d) higher internal consistency reliabilities, and (e) stronger intercorrelations.

Other research has examined measurement equivalence from a broader view; specifically, whether certain sub samples of individuals are more or less likely to use one modality over another (Hattrup, O'Connell, & Yager, 2006). For example, Hattrup et al. (2006) found that African Americans and Hispanics were more likely to use the IVR method, whereas Whites and Asians tended to use the Web approach more frequently.

No research has yet examined the measurement equivalence of a personality instrument administered on the Internet versus kiosk. Two major factors may lead to these two administration modes resulting in nonequivalent scores. First, the demographics of the individuals who use the Internet versus those who choose to use kiosks may be different. For example, Whites and Asians typically have greater access to technology (e.g., computers in the home) than Blacks and Hispanics (Ford & Whaley, 2003). Second, contextual factors may differ across administration modes. In an off-site Internet administration mode, applicants may feel freer to take their time in answering and be free of distractions in the privacy of their home. In an on-site kiosk administration mode, applicants may feel pressured to complete their responses in a reasonable period of time because of other applicants waiting for the kiosk and may be not be free of in-store distractions (e.g., loud shoppers).

Based on past research, we attempted to answer the following research questions:

1. Do personality items function differently across Internet and kiosk modes?
2. Are there race and gender differences in the *use* of the Internet and kiosk modes?
3. Are there differences in the two modes regarding the extent to which applicants will identify their race and gender?
4. Are there personality score differences in the two modes?
5. Racial and gender differences in personality scores:
    a. Are there racial and gender differences in personality scores across the two modes?
    b. Are the racial and gender differences in scores the same across the two modes?

Method

Sample

Our data set, which was obtained from an applicant tracking software firm, consisted of a nationwide sample of 370,122 unique applicants for positions with 61 different organizations. The organizations represented full service dining, hospitality, grocery, music retail stores, video stores, consumer electronics retail stores, and call centers. All of the jobs involved hourly workers.

Personality Measure

Development of the personality measure began with a literature search, followed by a job analysis, to determine which behaviors were important for each construct. Test developers then determined what personality traits should be useful in predicting these behaviors and developed a large item pool to measure those traits. The item pool was then reduced to 50 items (strongly disagree to strongly agree) for each scale by selecting those items that met various key criteria, such as minimizing social desirable responding, eliminating adverse impact, and providing comprehensive coverage of intended characteristics. Finally, a criterion-related validity study was performed with the overall scale, producing a correlation of .37 ($p < .01$; $N = 834$).

We selected the subscales of Conscientiousness (14 items); Agreeableness (10 items); and Emotional Stability (11 items) for further study because these three constructs tend to be related to customer-service orientation (Frei & McDaniel, 1998; Hogan, Hogan, & Busch, 1984). The remaining 15 items measured constructs not directly related to customer service (e.g., Thrill-Seeking; School Performance). We focused our study on constructs related to customer service measures so that our results may generalize to that class of measure.

Results

Do Personality Items Function Differently Across Internet and Kiosk Modes?

Our first set of equivalence analyses involved examining possible differential item functioning (DIF). DIF occurs when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item after matching on the underlying ability that the item is intended to measure. We first split each sample into a derivation and a cross-validation sample of equal sample sizes. We then examined, for every scale, the percentage of variance accounted for by the first principal component. Reckase (1979) suggests that the first principal component must account for at least 30% of the variance for unifactor models to be applied. All of the scales met this criterion. We then applied Zumbo's (1999) ordinal DIF model to each of the scales. This method has shown to perform extremely well under conditions of both uniform and nonuniform DIF (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). Zumbo's model uses ordinal logistic regression in a hierarchical fashion to identify DIF. The outcome variable is the item response. The logistic regression is conducted in a stepwise fashion with chi-square difference tests and r^2 difference tests conducted at each step:

Step #1: One first enters the conditioning variable (i.e., the total score);

Step #2: The group variable is entered (Internet versus kiosk); and finally

Step #3: The interaction term (Step 1 × Step 2) is entered into the equation.

A practically significant increase in r^2 of .01 or greater between Step #1 and Step #2 signals the presence of uniform DIF. Similarly, A practically significant increase in r^2 of .01 or greater between Step #2 and Step #3 signals the presence of non-uniform DIF. Uniform DIF is said to apply when differences between groups in item responses are found at all trait levels, while in non-uniform DIF an interaction is found between trait level, group assignment, and item responses (Camilli & Shepard, 1994).

Our DIF analyses did not show any items that evidenced either uniform or non-uniform DIF in either the derivation or cross-validation samples. The average unique percentage of variance accounted for by the group variable (Step #2 or uniform DIF) and the average unique percentage of variance accounted for by the interaction term (Step #3 or non-uniform DIF) were equal across derivation and cross-validation samples (r^2 = .0012 and r^2 = .0003, respectively). These analyses suggested that the items functioned similarly across modes of administration.

Are There Race and Gender Differences in the Use of the Internet and Kiosk Modes?

The sample characteristics are shown in Table 1. The Internet sample evidenced a higher percentage of men (61%) than the kiosk sample (54%). This difference was practically significant (r_{Φ} = .07). Moreover, the kiosk sample evidenced a higher percentage of Blacks (23%) and Hispanics (14%) and a lower percentage of Whites (57%) and Asians (3%) than the Internet sample (19%; 11%; and 65%, and 4%, respectively). These differences were also practically significant (r_{Φ} = .10).

Are There Differences in the Two Modes Regarding the Extent to Which Applicants Will Identify Their Race and Gender?

No practically significant difference existed between modes in the extent to which applicants were willing to identify their gender (r_{Φ} = .01) with 93.8% in the Internet sample willing to identify their gender and 93.3% in the Kiosk sample willing to identify their gender. The Internet sample evidenced a higher percentage of individuals (91.2%) who were willing to identify their race/ethnicity than those individuals in the kiosk sample (87.4%). This difference was practically significant (r_{Φ} = .06).

Are There Personality Score Differences in the Two Modes?

Table 2 presents the descriptive statistics and intercorrelations for the Internet and kiosk samples. We used mean scale scores to enable easy interpretations across scales and modes. The intercorrelations and alpha coefficients are highly similar across samples (difference range .00-.02). The relatively high intercorrelations across samples are consistent with the "ideal employee" factor that is often found in applicant samples with personality data (cf. Schmit & Ryan, 1993). Standardized difference scores were negligible across modes for Conscientiousness (d = .13) and

Agreeableness (*d* = .09) but practically significant for Emotional Stability (*d* = .24). The mean for the Internet sample was approximately ¼ standard deviation higher than that of the kiosk sample for the Emotional Stability scale.

Are There Racial and Gender Differences in Personality Scores Across the Two Modes?

Personality means by race/ethnicity and gender are shown in Table 3. Results of the Mode of Administration × Gender × Race/Ethnicity ANOVAs (see Table 4) for each of the three personality variables found negligible effects for gender (.000, .000, and .000 partial eta-squared's for conscientiousness, agreeableness, and emotional stability, respectively) and race/ethnicity (.005, .009, and .004 partial eta-squareds for conscientiousness, agreeableness, and emotional stability, respectively).

Table 1
Sample Characteristics

	Overall Sample	**Internet**	**Kiosk**
Gender	370,122	201,495	168,627
Male	200,370 (58%)	115,350 (61%)	85,020 (54%)
Female	146,012 (42%)	73,689 (39%)	72,323 (46%)
No response	23,740	12,456	11,284
Race/ethnicity	370,122	201,495	168,627
White	204,273 (62%)	119,957 (65%)	84,316 (57%)
Black	68,562 (21%)	34,640 (19%)	33,922 (23%)
Hispanic	41,013 (12%)	19,785 (11%)	21,228 (14%)
Asian	11,505 (4%)	7,476 (4%)	4,029 (3%)
South Pacific Islander	1,933 (<1%)	823 (<1%)	1,110 (<1%)
Native American	2,370 (<1%)	1,070 (<1%)	1,300 (<1%)
Other	1,568 (<1%)	10 (<1%)	1,558 (1%)
No response	38,898	17,734	21,164

Note. Percentages are all derived by removing the "No response" frequencies from the total.

Table 2
Descriptive statistics and intercorrelations for Internet and kiosk samples

Variable	M_I	SD_I	M_K	SD_K	1	2	3
1. Conscientiousness	3.54	0.29	3.50	0.31	.78 (.77)	.63	.57
2. Agreeableness	3.36	0.32	3.33	0.33	.63	.68 (.67)	.56
3. Emotional Stability	3.37	0.32	3.29	0.35	.57	.57	.70 (.70)

N_{Internet} = 201,495. N_{Kiosk} = 168,627. I = Internet. K = Kiosk. Correlations controlled for race/ethnicity and gender changed on average approximately |.005|; we therefore report the uncorrected correlations. Correlations below the diagonal are for the Internet sample. Correlations above the diagonal are for the Kiosk sample. Alpha coefficients are arranged along the diagonal (Kiosk sample alpha coefficients are in parentheses). Sample means and standard deviations for the entire sample (*N* = 370,122) are: Conscientiousness (*M* = 3.52, *SD* = 0.30); Agreeableness (*M* = 3.35, *SD* = 0.33); Emotional Stability (*M* = 3.33, *SD* = 0.34).

Table 3
Personality means and standard deviations by race/ethnicity and gender

	Conscientiousness		Agreeableness		Emotional Stability	
	M	*SD*	*M*	*SD*	*M*	*SD*
Internet						
Race/Ethnicity						
White	3.53	0.29	3.35	0.31	3.39	0.31
Black	3.58	0.28	3.43	0.32	3.37	0.33
Hispanic	3.59	0.28	3.40	0.31	3.36	0.33
Asian	3.59	0.28	3.40	0.32	3.41	0.32
South Pacific Islander	3.49	0.30	3.30	0.31	3.33	0.31
Native American	3.55	0.28	3.35	0.31	3.38	0.31
Other	3.32	0.23	3.26	0.32	3.19	0.27
Did not indicate	3.46	0.32	3.29	0.33	3.27	0.34
Gender						
Male	3.55	0.30	3.37	0.32	3.39	0.32
Female	3.53	0.29	3.37	0.31	3.37	0.31
Did not indicate	3.45	0.32	3.28	0.34	3.26	0.34
Kiosk						
Race						
White	3.49	0.31	3.32	0.33	3.32	0.33
Black	3.53	0.30	3.39	0.34	3.27	0.36
Hispanic	3.52	0.30	3.36	0.33	3.26	0.35
Asian	3.51	0.31	3.36	0.34	3.29	0.36
South Pacific Islander	3.49	0.30	3.30	0.32	3.34	0.32
Native American	3.48	0.31	3.31	0.34	3.29	0.35
Other	3.47	0.32	3.38	0.32	3.13	0.36
Did not indicate	3.43	0.32	3.27	0.35	3.20	0.36
Gender						
Male	3.50	0.32	3.33	0.34	3.29	0.35
Female	3.50	0.30	3.35	0.33	3.29	0.34
Did not indicate	3.45	0.32	3.30	0.34	3.22	0.35

Are the Racial and Gender Differences in Scores the Same Across the Two Modes?

Results of the Mode of Administration × Gender × Race/Ethnicity ANOVAs (see Table 4) for each of the three personality variables found negligible effects for the mode × gender interaction (.000, .000, and .000 partial eta-squared's for conscientiousness, agreeableness, and emotional stability, respectively) and mode × race/ethnicity interaction (.000, .000, and .001 partial eta-squared's for conscientiousness, agreeableness, and emotional stability, respectively).

Conclusions

This research attempted to answer five questions related to the measurement equivalence of a personality inventory administered on the Internet versus kiosk. We attempted to answer the first question, "Do personality items function differently across Internet and kiosk modes?" through the use of DIF methodology. The results suggest that, at the item level, the personality inventory appears to function equivalently across administration modes. These results are consistent with a past DIF study that found item-level equivalence of an employee attitude survey across Internet and paper-and-pencil modes (Sireci, Harter, Yang, & Bhola, 2003).

Table 4
ANOVA Results for mode of administration, gender, and race/ethnicity

Source	SS	df	MS	F	*p* <	Partial Eta-Squared
Conscientiousness						
Mode (M)	0.00	1	0.00	0.05	N.S.	.000
Gender (G)	0.30	1	0.30	3.35	N.S.	.000
Race (R)	148.28	6	24.71	279.80	.001	.005
M × G	0.10	1	0.10	1.17	N.S.	.000
M × R	12.11	6	2.02	22.84	.001	.000
G × R	1.34	6	0.22	2.52	<.05	.000
M × G × R	0.49	6	0.08	0.92	N.S.	.000
Error	29252.11	331,184	0.09			
Agreeableness						
Mode (M)	0.01	1	0.01	0.14	N.S.	.000
Gender (G)	0.13	1	0.13	1.28	N.S.	.000
Race (R)	314.45	6	52.41	508.24	.001	.009
M × G	0.11	1	0.11	1.07	N.S.	.000
M × R	1.84	6	0.31	2.97	.01	.000
G × R	1.58	6	0.26	2.56	.05	.000
M × G × R	0.64	6	0.11	1.03	N.S.	.000
Error	34150.64	331,184	0.10			
Emotional Stability						
Mode (M)	0.57	1	0.57	5.36	.05	.000
Gender (G)	0.37	1	0.37	3.46	N.S.	.000
Race (R)	129.16	6	21.53	201.24	.001	.004
M × G	0.12	1	0.12	1.14	N.S.	.000
M × R	30.93	6	5.15	48.18	.001	.001
G × R	5.07	6	0.85	7.90	.001	.000
M × G × R	0.45	6	0.08	0.70	N.S.	.000
Error	35428.05	331,184	0.11			

The second question we attempted to answer was, "Are there race and gender differences in the *use* of the Internet and kiosk modes?" We found a higher percentage of men, Whites, and Asians chose the Internet mode, whereas a higher percentage of females, Blacks, and Hispanics chose the kiosk mode. These results are consistent with those of Hattrup et al. (2006) who found that men, Whites, and Asians chose the Web version more frequently than the IVR version as opposed to females, African Americans, and Hispanics. As noted previously, these results can be explained by the fact that Whites and Asians typically have greater access to technology (e.g., computers in the home) than Blacks and Hispanics (Ford & Whaley, 2003). Moreover, research published in the last 20 years suggests that females are at a disadvantage relative to men when learning about computers or learning other material with the aid of computer-assisted software (cf. Cooper, 2006). It makes sense then that females in general would choose the least computer-intensive mode in these contexts.

The third question we attempted to answer was, "Are there differences in the two modes regarding the extent to which applicants will identify their race and gender?" The results of the study found no evidence for differential disclosure across gender; however, those in the Internet sample were more likely to disclose their race/ethnicity than those in the kiosk sample. This result is consistent with research that has found that the computerized Web interface may facilitate self-disclosure for sensitive information (Davis, 1999). Efforts to increase the self-disclosure of sensitive information for participants who choose the kiosk mode may include such options as locating the kiosk in a quiet and inconspicuous location. Future research should explore possible options.

The fourth research question we attempted to answer was, "Are there personality score differences in the two modes?" At the scale level, intercorrelations and alpha coefficients were comparable. However, the practically significant difference of the mean scale score between the Internet sample and the kiosk sample (with Internet scoring higher) for emotional stability and the slight elevation of the other personality scales scores in the Internet sample may be a sign of differential amounts of faking occurring in the two samples. Several studies have found that the more testing time participants used, the more they tended to fake (Chen, Lee, & Yen, 2004; Robie, Brown, & Bly, in press). It is plausible that applicants are freer from distractions and have more opportunity to fully commit their mental energy to the task of filling out a personality inventory under conditions of an Internet modality than a kiosk modality. Emotional stability likely emerged as the only scale to evidence a marked difference because it arguably is the most socially undesirable of the three scales; this may inspire applicants with additional time to pay particular attention to "shading the truth" on the emotional stability items in particular. Although research has found that the Big Five factors are equally fakeable (Viswesvaran & Ones, 1999), no research that we are aware of has examined whether the Big Five factors are *actually* differentially faked.

Our final research question revolved around the question of racial and gender differences in personality scores across the two modes. We found no evidence: (a) that differences existed across race and gender on the personality scores across the two modes; or (b) of any interaction between mode and race (or gender) on the personality scores. Little to no organizational research has yet studied this particular issue to our knowledge.

In conclusion, we have found many areas of equivalence for a typical customer service measure across Internet and kiosk samples but some areas of nonequivalence. It is likely that personality scores will largely be equivalent across modes. One may, however, expect a slight but practically significant elevation for scales saturated with emotional stability in the Internet samples. However, it is also likely that certain groups of individuals protected under affirmative action legislation will both choose one mode over another and also be less likely to divulge demographic information that organizations find useful for equal employment opportunity purposes. It would behoove practitioners to take these possibilities into account when developing norms, setting cutoff scores, and devising means to increase the rate of disclosure of sensitive information among certain subgroups of individuals.

References

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chen, C., Lee, M., & Yen, C. (2004). Faking intention on the Internet: Effects of test types and situational factors. *Chinese Journal of Psychology, 46,* 349-359.

Church, A. H. (2001). Is there a method to our madness? The impact of data collection methodology on organizational survey results. *Personnel Psychology, 54,* 937-969.

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22,* 320-334.

Davis, N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments and Computers, 31,* 572-577.

Ford, D. L., & Whaley, G. L. (2003). The digital divide and managing workforce diversity: A commentary. *Applied Psychology: An International Review, 52,* 476-485.

Frei, R. L., & McDaniel, M. A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11,* 1-27.

Hattrup, K., O'Connell, M. S., & Yager, J. R. (2006). Pre-screening job applicants with interactive voice response and web-based technologies: Are the methods equivalent? *Applied H.R.M. Research, 11*, 15-26.

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69,* 167-173.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65,* 935-953.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Applied Psychological Measurement, 4,* 207-230.

Robie, C., Brown, D. J., & Bly, J. C. (in press). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology*.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78,* 966-974.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3,* 129-150.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197-210.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Author Information

Chet Robie
Wilfrid Laurier University
School of Business & Economics
Waterloo, Ontario N2L 3C5
(519) 884-0710 ext. 2965
crobie@wlu.ca (email)

Douglas J. Brown
University of Waterloo
Department of Psychology
Waterloo, Ontario N2L 3G1