

Web-based Multisource Reference Checking: An investigation of psychometric integrity and applied benefits

Cynthia A. Hedricks*, Chet Robie** and Frederick L. Oswald***

*SkillSurvey, Inc., 565 E. Swedesford Road, Ste. 315, Wayne, PA 19087, USA. chedricks@skillsurvey.com

**School of Business and Economics, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada

***Department of Psychology, Rice University, Houston, TX 77005, USA

Reference checking is a near-universal practice within personnel selection systems, and legal pressures to gather job-relevant and structured feedback from references is mounting. Despite this state of affairs, reference checking is a woefully under-researched method for obtaining psychometrically sound and behaviorally informative data that predict task, team, and leadership behavior at work. From studies of job candidates in applied settings, this article reports on the reliability, validity, and compliance of multisource reference feedback gathered using a web-based methodology. Acceptable levels of internal consistency, inter-rater reliability, and test-retest reliability of the reference-checking instrument were realized. Results of survival analyses found support for prediction of involuntary, but not voluntary turnover. No practically significant differences were found in overall mean scores across demographic subgroups. Finally, the web-based reference-checking system evinced high degrees of efficiencies across a range of metrics (e.g., reference response time, reference response rate, candidate response time).

1. Introduction

Reference checking in personnel selection is a near-universal practice, yet has long been an under-researched area (Muchinsky, 1979). The relatively low validities reported for reference-checking data, combined with the collective belief that reference checking provides a dearth of information that is not consistent, reliable, or psychological in nature, are historical forces that probably served to direct personnel selection research away from reference checking and toward standardized measures of job applicants' cognitive ability or personality traits instead (Hunter & Hunter, 1984; Reilly & Chao, 1982; Schmidt & Hunter, 1998).

Within the past decade, however, there has been increased interest in reference checking and other means for gathering information external to job applicants' own

test scores and self-reports. This interest emerges from the confluence of at least five important factors: (a) the dual emphasis by human resource (HR) professionals on improving 'quality of hire' as reflected in past work behavior, while also achieving a diverse workforce; (b) legal changes that favor the practice of gathering feedback from job references; (c) the need for additional time and cost savings that could be reaped from initial reference screening; (d) studies validating observer ratings; and (e) validation studies providing empirical support for many different 'alternative' approaches to selection (i.e., those not involving traditional cognitive or personality assessments).

Even when high levels of unemployment have resulted in a higher volume of job applications per open position, HR professionals are still engaged in what has been termed a 'war for talent' (Michaels, Handfield-Jones, & Axelrod, 2001); there are concerns for ensuring quality-of-hire and diversity, regardless of fluctuations in the job market. Employers cannot solely rely on resumes, applications, or interviews to gather accurate information as to a candidate's past experiences and competence, as candidates often embellish positive information

This article is developed from research presented at the 23rd and 24th annual conferences of the Society for Industrial and Organizational Psychology in San Francisco (2008) and New Orleans (2009).

and/or omit critical negative information (e.g., Levashina & Campion, 2007; Wood, Schmidtke, & Decker, 2007).

One solution is to gather reference feedback from external sources, namely those who have had some experience working with the candidate, such as in the role of supervisor, coworker, or subordinate (Bliss, 2001; Fernandez-Araoz, 2007). Note, however, that it is typically rare to find explicitly 'negative' or 'developmental' information in letters of recommendation (Grote, Robiner, & Haut, 2001). Alternatively, asking references to respond to structured questions concerning the candidate's job-related competencies leads to a standardized measure that is consistent and fair across applicants; it also provides the ability to identify and compare profiles of applicants' work-relevant traits and behaviors, and thus identify performance patterns, rating discrepancies, and areas for development. Compared with unstructured letters of recommendation, structured reference surveys may also be more likely to reduce idiosyncrasies and systematic differences in reference feedback (for the latter, see Madera, Hebl, & Martin, 2009, for an example of gender differences). Moreover, as with other more systematic approaches to personnel selection, structured reference checking has been found to increase validity over nonstructured reference checking (Aamodt, Bryan, & Whitcomb, 1993; McCarthy & Goffin, 2001; Taylor, Pajo, Cheung, & Stringfield, 2004; Zimmerman, Triana, & Barrick, 2010).

Legal changes that now make reference checking more likely include an increase in negligent hiring lawsuits that arise when employers fail to check the background and past experiences of an employee (cf. *Deerings West Nursing Center v. Scott*, 1990; *McLean v. Kirby Co.*, 1992; *Tallahassee Furniture Co., Inc. v. Harrison*, 1991). This liability is of particular concern for applicants who, if hired, would pose a threat to other individuals, such as patients, children, customers, or coworkers. Fortunately, the reference-checking process is facilitated by the conditional privilege or immunity offered by most states (Gatewood, Feild, & Barrick, 2007) that protects reference providers who share both positive and negative job-related information, so long as the information shared is truthful and well-intentioned, and not malicious or defamatory.

Thus there are several strong reasons for the need to confirm a candidate's claims about his or her past work experience, and to use this information to improve selection-related decisions. Although groups such as the Society for Human Resource Management (SHRM) and other HR experts recommend reference checking as a method to do so (Bliss, 2001; SHRM, 2005), certain obstacles continue to limit its use.

In fact, when the first author conducted two separate webcasts in 2008,¹ 68 of the 150 attendees (45%) reported that the main reason that reference checking was not routinely conducted at their companies was

that the 'HR staff was too busy.' Survey results from companies across multiple industries found that HR departments spend approximately 48% of their time sourcing and screening applicants, and another 18% processing new hires (Automatic Data Processing (ADP), Inc., 2008). The amount of additional time for HR to conduct and complete a telephone reference check can be a deterrent because professional references are often busy and cannot be reached by telephone during traditional business hours. Unfortunately, any delay in reference checking has the potential to increase the time to fill an open position, resulting in lost opportunity and productivity.

In response to the critical need to gather reference feedback on job candidates in a standardized, timely, and cost-efficient manner, a web-based reference-checking system was developed. Compared with traditional reference-checking procedures, the more efficient system described herein combines best practices in competency modeling (e.g., Spencer & Spencer, 1993; Wood & Payne, 1998), multisource feedback methodology, and survey and data-management technologies. This article describes the nature of this web-based reference-checking system, with empirical results describing its psychometric reliability, validity for predicting employee turnover over time, along with race/ethnic group differences and their implications for adverse impact.

2. Method

2.1. General measure and procedures

A web-based technology² was used whereby multiple individuals who have worked with a job candidate provided structured feedback on the candidate's competency level on several different dimensions of job-related behaviors. This prehire process was used by companies at various points in the candidate selection process to guide the interview, to confirm a candidate's claims as to his/her past work experiences, and to assist with a hiring decision.³

All reference-checking surveys included items reflecting the following four dimensions of job-related behaviors at a minimum: Professionalism (PR); Interpersonal Skills (IS); Problem Solving and Adaptability (PSA); and Personal Value Commitment (VAL). Example items are: 'Work in a disciplined and organized way to complete goals in a timely manner' (PR); 'Build strong, positive working relationships with manager and peers and maintain them over time' (IS); 'Make high-quality decisions based upon facts and business priorities' (PSA); and 'Demonstrate trustworthiness, honesty, and high personal standards in dealings with others' (VAL).

In order to tailor the reference-checking process and include additional dimensions of job performance, job

analysis information was collected from each prospective employer to understand the organization-specific competency areas for job success, which then led to choosing an appropriate survey from the existing library of competency-based surveys. The prospective employer set the parameters as to the number of reference providers within each rater group (e.g., manager [or teacher], peer, direct report, or client) who would receive requests for feedback.

At the request of the prospective employer, candidates entered some standard information into the Internet survey about their job references, including current contact information and job title, and information pertaining to when they worked with the reference (e.g., company name, their job title, and dates worked together). Candidates were also asked to complete an optional electronic form that gathered demographic data (gender, race/ethnicity, age group); they were informed that this information was for research purposes only and would not be shared with references or prospective employers.

Once the candidates entered the required information, their references received an e-mail from the candidate requesting the following: (a) salutation (e.g., Mr., Ms.); (b) contact information; (c) work relationship to the candidate (e.g., manager, peer); and (d) job title, company name, and dates worked together. References also provided ratings on a survey indicating the extent to which the candidate engaged in a number of job-related behaviors, with all items using a 7-point Likert-type scale ranging from 1 (*never*) to 7 (*always*) or the option to check N/O (*not observed*), as well as comments to two open-ended questions, one asking for at least three of the candidate's strengths and another asking for at least three areas in which the candidate could improve. Automated, survey-specific scoring keys were used to aggregate the ratings and verbatim comments across multiple raters. Feedback was treated confidentially; neither the behavior ratings nor the verbatim comments were associated with a given reference provider on the candidate report to the prospective employer. The typical number of references received (median = 4) allowed for comments to be fairly anonymous in nature.

Measures included: reference provider contact information; candidate optional demographic information; *candidate response time* (i.e., the number of business days elapsed between when the prospective employer entered the candidate into the software and when the candidate completed his/her portion of the process); *reference response time* (i.e., the number of business days that elapsed between when the candidate entered reference contact information into the software and when the reference submitted the completed survey); *report turnaround time* (i.e., the number of business days from when the recruiter sent an e-mail to the

candidate to initiate the reference-checking process to when the client-determined required amount of references had completed the survey); *number of references completed (total)*; *number of references completed (managers)*; *reference response rate* (the number of completions divided by the total requests); and a set of overall scores: an average rating on all performance dimensions, overall scores for each one of the specific performance dimensions, and an average rating for each individual behavior within a given performance dimension. All sets of overall scores reflected averages across all reference providers. Whenever two or more managers had provided feedback on the job candidate, aggregated ratings from the managers only were also calculated and made available. Open-ended comments as to the candidate's strengths and areas for improvement were presented verbatim.

Several policies are designed to protect the privacy of both candidate and reference provider (Woska, 2007). First, candidates must authorize reference providers to be contacted; in fact, candidates themselves enter the reference contact information into the software. Second, we do not provide the reference feedback report to candidates, and strongly urge our clients not to provide copies to the candidate. Third, we do not generate a report unless at least three or more references have responded. Fourth, we never associate a reference provider's name to reference feedback.

Several policies are designed to guard against possible fraudulent references. First, we collect reference names, e-mail addresses, job titles, companies, phone numbers, etc. Second, we inform the candidate that it is possible that the reference provider will be contacted (and client organizations sometimes do outside of the SkillSurvey process) to mitigate any thoughts of gaming the system. Perceived risk of identification (i.e., getting caught) has been found to lead to less faking in personality tests (Dwight & Donovan, 2003). Finally, we also collect and report the IP (Internet protocol) addresses of the candidates and of the reference providers. Each host or network has a unique IP address. If the IP address of the candidate matches that of a reference provider, or if the IP address is the same for two or more of the references, the duplicate IP addresses appear in red font on the candidate's reference-checking report. Our recommendation to the prospective employer is then to verify the reference providers with a phone call.

2.2. Reliability studies

2.2.1. Internal consistency

We computed internal consistency of the items (alphas by dimension) on eight representative competency surveys from a library of approximately 250 surveys. We randomly selected ($N = 1,000$) for each survey from

our total database of organizations; thus, the total sample size for the internal consistency analyses was ($N = 8,000$).

2.2.2. Test–retest reliability

A representative sample of organizations was selected to collect test–retest data. Reference providers who had given feedback on a candidate received a request two weeks later to provide a second set of responses on the same behavior items for quality control purposes. These reference providers were informed that like the ratings and comments they provided earlier, the second set of responses would be treated confidentially. Sample sizes ranged from 174 to 238 across surveys. The sample sizes ranged from $N = 77$ to 238 depending on the source (all references or managers only) and survey (mean $N \approx 162$). The total sample size was 1,298.

2.2.3. Inter-rater reliability

Another representative sample of organizations was chosen to examine inter-rater reliability. The sample consisted of 20,822 reference providers and 4,236 candidates.

2.3. Turnover and validity studies⁴

The total sample consisted of 7,243 newly hired employees across six different organizations ($n_1 = 330$, $n_2 = 412$, $n_3 = 1,095$, $n_4 = 1,566$, $n_5 = 1,568$, $n_6 = 2,272$). Organizations 1, 3, 4, and 5 were US healthcare organizations. Organization 2 was a national US chain of retail stores. Organization 6 was a global high-tech company. The average length of the study observation period across organizations was approximately 234 days (standard deviation [SD] = 178 days, with some positive skew). All analyses were aggregated across companies because the performance behaviors that were assessed were critical across organizations (i.e., organization specific dimensions were excluded). The majority of respondents were female (67.5%), White (58.2%), and 39 years of age or younger (75.0%).

The first goal of the studies was to investigate associations between specific aspects of the process that were related to employee turnover. Specifically, we examined the relationship between both involuntary and voluntary turnover with: (1) overall score on the prehire reference assessment; (2) candidate response time; (3) median reference response time; and (4) reference response rate. Organization and specific vocational preparation (SVP; U.S. Department of Labor, 1991) were used as control variables in these analyses (combined as a stratification variable – see later). SVP is defined as ‘the amount of elapsed time required by a typical worker to learn the techniques, acquire the information,

and develop the facility needed for average performance in a specific job–worker situation’ (DOT, Vol. II, p. 1009). Research has shown that SVP is conceptually related to and highly correlated with other measures of job complexity and job level (Robie, Ryan, Schmieder, Parra, & Smith, 1998).

The second goal of the studies was to validate the prehire reference assessment process; specifically, to correlate overall performance dimension scores provided by the candidates’ references with supervisor performance evaluations once these individuals were on the job. Toward this end, supervisors were asked to rate their direct reports. It is important to emphasize that they used the same methodology and survey items that had been used to collect reference feedback during selection; therefore, the reference feedback can be directly matched with supervisors’ performance evaluations. We collected posthire feedback for 239 of the 330 new hires (72.4%) in one of the healthcare organizations. The number of months the new hire was on the job was used as a control variable in this analysis as a proxy for the temporal effects of increasing cohort turnover because of poor performance and increasing job performance because of job experience.

2.4. Group differences

Recent candidates (a subsample of $N = 150,000$) were presented with an online form that asked them to respond to three questions pertaining to their sex, age group, and race/ethnicity group ($N = 146,562$ for age group, however, as 3,438 candidates came from companies that declined collection of this information). Candidates were informed that responses to any or all of these questions were entirely optional and for research purposes only, and that the information that they provided would not be included in the reference-checking process. Thus, 18,140 (12.1%) did not report their sex; 18,038 (12.0%) did not report their age group; and 22,171 (14.8%) did not report their race/ethnicity group.

These 150,000 candidates were applying for jobs at one of 516 companies. A wide spectrum of industries was represented; for example: healthcare (59.6%), financial services (9.5%), staffing (8.0%), high-tech (4.6%), education (4.1%), engineering and construction (2.8%), manufacturing (2.5%), retail (2.2%), business services (2.0%), telecom (1.4%), energy and utilities (1.0%), and a number of industries with <1% (e.g., nonprofit, aerospace, entertainment, government, real estate, automotive). A total of 194 different competency surveys were represented; candidates were applying for positions related to nursing (16.7%), general professional (16.4% – those with <3% each-HR, legal, research and development, marketing, operations, production, etc.), allied health (16.2%), customer service (9.4%), midlevel

management (8.2%), administrative (6.9%), general hourly (6.6%), information technology (6.3%), sales (4.1%), finance (3.7%), engineering and construction (3.6%), executive-level management (1.2%), and education (0.6%).

2.5. Efficiency metrics

We examined several indices to measure the efficiency of the web-based multisource reference-checking process: (1) candidate response time (in business days); (2) reference response time (in business days); (3) report turnaround time (in business days); (4) number of references requested (total and managers only); (5) number of references completed (total and managers only); and (6) reference response rate. All of these indices are important in evaluating the practical benefits of the system in an applied setting, given the desire for quick turnaround on receiving reference feedback from a large and representative number of references to inform the hiring decision. These analyses were conducted using the same sample of $N=150,000$ as was used for the analyses of group differences.

3. Results

3.1. Reliability

3.1.1. Internal consistency

Eight representative competency surveys from a library of approximately 250 surveys were evaluated for internal consistency (alpha). These eight surveys included RN (24 items), Sales Professional (22 items), IT Professional (22 items), Finance Professional (22 items), Admin Professional (22 items), Customer Service Hourly (18 items), General Professional (18 items), and General Hourly (18 items). Alpha coefficients, calculated on 1,000 recent randomly selected candidates from each of these surveys, were very high, ranging between 0.96 and 0.98. High reliability (low measurement error variance) is especially important for making personnel selection decisions; the ritualistic .70 minimum level for alpha is not nearly high enough (Lance, Butts, & Michels, 2006; Nunnally, 1978).

3.1.2. Test-retest reliability

The reference response rate to the research requests for a second rating two weeks later was 48.8%. The percentage of raters in each rater group was essentially the same (manager: 48.9% vs. 48%; peer: 39.9% vs. 40.6%; client: 7.1% vs. 6.8%; and direct report: 4.1% vs. 4.6%) for the first and second set of ratings, respectively. For all but one of the five surveys, the sample size was greater than 30. Results on responses from all references, as well as those for managers only, are presented

Table 1. Two-week test-retest correlations on behavioral competency ratings of job candidates: all references versus managers only

Web-based survey name	All references		Managers Only	
	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>
General professional, v1	238	.78	128	.74
General professional, v2	196	.78	90	.72
IT professional	271	.64	124	.77
Sales professional	174	.65	77	.83

Note: All test-retest correlation coefficients of data gathered from the web-based multisource procedure were significant at $p < .001$.

in Table 1. Test-retest reliabilities ranged from .64 to .83 and on average were somewhat higher for managers (.77) than for all references combined (.71). These coefficients are lower than the alpha reliability coefficients reported, which is very typical for psychological measures.

3.1.3. Inter-rater reliability

Inter-rater reliability was assessed by the use of intraclass correlations that were calculated by using generalizability theory (Shavelson & Webb, 1991). The sample consisted of 4,236 ratees and 20,822 raters from a database that represented 33 companies and 1,344 job titles. Variance components were computed using minimum norm quadratic unbiased estimation (or MINQUE) given its relative ease to compute with large datasets. Five variance components were derived based on the generalizability design of references-within-applicants interacting with items (see Greguras & Robie, 1998, for an example): (a) all references across the 18-item survey; (b) references who were managers across the 18-item survey; (c) all references for each of the four competency areas; and (d) references who were managers for each of the competency areas. The five variance components were: (a) person/ratee (or 'true') variance; (b) item main effect; (c) person $\times$ item interaction; (d) an undifferentiated variance term that contained both the reference main effect and the person $\times$ reference interaction; and (e) an undifferentiated residual term that contained the reference $\times$ item interaction, the three-way interaction of ratee, reference, and item, and any remaining unmeasured sources of error.

In order to estimate variance components accurately, we sampled as large a subsample of the dataset as was computationally possible ($N=200$)⁵; then we repeated this subsampling 10 times to derive mean variance component estimates. Mean variance components were then used as input into what generalizability theory calls *decision studies* that project reliability estimates in which one or more sources of error variance have been manipulated simultaneously. More specifically, we held the number of items constant in the decision study as those

Table 2. Generalizability coefficients for web-based feedback gathered from all references and from managers only

# Raters	All References					Managers Only				
	Overall	IS	PR	VAL	PSA	Overall	IS	PR	VAL	PSA
1	.13	.12	.13	.11	.11	.19	.18	.19	.17	.16
2	.22	.21	.23	.20	.20	.31	.30	.32	.29	.28
3	.30	.29	.31	.26	.28	.40	.39	.41	.38	.36
4	.37	.35	.38	.32	.34	.47	.45	.48	.44	.43
5	.42	.40	.43	.37	.39	.53	.50	.54	.49	.48
6	.46	.44	.47	.41	.43	.57	.55	.58	.54	.53
7	.50	.48	.51	.44	.47	.61	.58	.61	.57	.56
8	.53	.51	.54	.47	.50	.64	.61	.64	.60	.59
9	.56	.54	.57	.50	.53	.67	.63	.67	.62	.62
10	.59	.56	.59	.52	.56	.69	.65	.69	.64	.64
11	.61	.58	.62	.54	.58	.71	.67	.71	.66	.66
12	.63	.60	.64	.56	.60	.72	.69	.72	.68	.68
13	.65	.62	.65	.58	.62	.74	.70	.74	.69	.70
14	.66	.63	.67	.59	.63	.75	.72	.75	.70	.71
15	.68	.64	.68	.61	.65	.76	.73	.76	.72	.72

Note: Data from the 18-item General Professional, v2 were used for all computations. IS = Interpersonal Skills; PR = Professionalism; VAL = Personal Value Commitment; PSA = Problem Solving and Adaptability.

that were present in the survey *in situ* and examined the reliability when increasing number of references in steps, from 1 to 15.

Table 2 lists the generalizability coefficients for all references combined and for those of just managers. The range for generalizability coefficients (from 1 to 15 references) for all reference types was .11/.13 to .61/.68 across scales and for managers was .16/.19 to .72/.76. The average number of references responding was approximately five; thus, one would expect typical inter-rater reliability of the ratings for all references to be .37 to .42 depending on the scale. The average number of managers responding was approximately three; thus, one would expect typical inter-rater reliability of the ratings to be .36 to .41 for managers depending on the scale.

3.2. Turnover

The two criterion variables that we examined for retention were involuntary and voluntary turnover. Involuntary turnover occurs when employees are fired or laid off. Voluntary turnover occurs when employees quit of their own volition. We examined these criterion variables separately because they have different meanings to organizations, and research has found these two variables to have different antecedents and relationships with job performance (cf. Bycio, Hackett, & Alvares, 1990).

We used Cox regression analysis to examine the relationship between the candidate response time, reference response rate, and overall score (prehire) and the two turnover criterion variables mentioned earlier. Cox regression is a form of survival analysis that is particularly useful with a binary outcome variable and multiple

independent variables, any or all of which can be measured over time (Morita, Lee, & Mowday, 1989, 1993; Somers, 1996; Somers & Birnbaum, 1999). We did not include in the regressions number of references requested or number of references completed because these two variables contribute to reference response rate. We included a stratification variable that was a fully crossed combination of job level (5 levels) $\times$ organization (6 levels) to ensure that the proportionality of hazards assumption of Cox regression was met. Standardized variables were used in these analyses to aid in the interpretation of the odds ratios (ORs).

Tables 3 and 4 show the results of the Cox regression analyses for predicting voluntary and involuntary turnover, respectively. The model was significant for involuntary turnover. Reference response rate significantly predicted involuntary turnover ($p = .001$) such that higher reference response rates were associated with lower involuntary turnover ($\beta = -.204$, $OR = 0.82$). Overall score (prehire) significantly predicted involuntary turnover ($p = .003$) such that higher scores were also associated with lower involuntary turnover ($\beta = -.156$, $OR = 0.86$). The OR is the exponentiated regression weight from logistic regression (e^{β}) and represents the increase in odds of surviving for a one-unit increase in the covariate (Tabachnick & Fidell, 2001). Turning to our results, the OR of 0.82 for reference response rate estimates that an increase in the reference response rate by one SD is associated with an 18% reduction in the odds of being involuntarily terminated (controlling for other variables in the model). The OR of 0.86 for overall score (prehire) suggests that an increase in one SD in overall score (prehire) is associated with a 14% reduction in the odds of being involuntarily terminated (controlling for other variables in the model). The

Table 3. Cox regression analysis predicting voluntary turnover

Variable	β	SE	Wald	OR	df	Probability
Candidate response time	.018	.035	0.27	1.01	1	.604
Reference response time	-.007	.035	0.04	0.99	1	.840
Reference response rate	-.071	.037	3.71	0.93	1	.054
Overall score (prehire)	-.001	.037	0.00	1.00	1	.988

Note: $N = 7,230$ ($N = 6,462$ not terminated and $N = 768$ voluntarily terminated). Model $\chi^2(4) = 4.21$, $p = .378$. df = degrees of freedom; OR = odds ratio; SE = standard error.

Table 4. Cox regression analysis predicting involuntary turnover

Variable	β	SE	Wald	OR	df	Probability
Candidate response time	.016	.056	0.09	1.02	1	.769
Reference response time	-.013	.052	0.06	0.99	1	.803
Reference response rate	-.204	.059	11.88	0.82	1	.001
Overall score (prehire)	-.156	.052	9.13	0.86	1	.003

Note: $N = 7,230$ ($N = 6,951$ not terminated and $N = 279$ involuntarily terminated). Model $\chi^2(4) = 22.33$, $p < .001$. df = degrees of freedom; OR = odds ratio; SE = standard error.

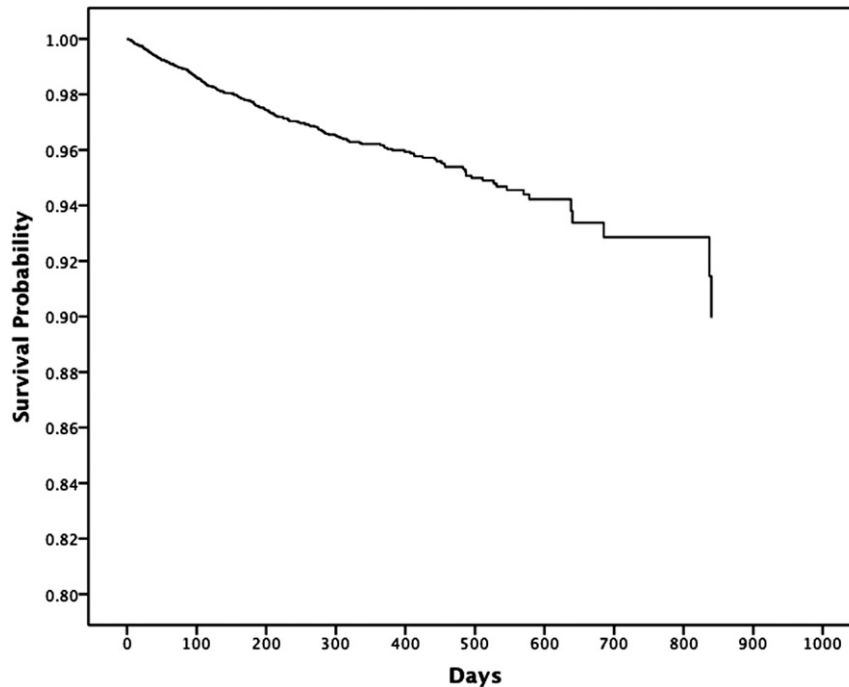


Figure 1. Survival function at one standard deviation above the mean for both reference response rate and overall score (prehire).

probability of survival at various time points and at various levels of covariates or treatments is called a survival function. Figures 1 and 2 show the survival functions at one SD above and below the means, respectively, for both reference response rate and overall score (prehire). Comparison of the two figures clearly shows that selecting individuals based on reference response rate and overall score would markedly decrease involuntary turnover across an approximately 3-year period.

3.3. Validity

Supervisor ratings were obtained on 239 of the new hires from Organization 1 who were reference checked with the web-based multisource process and still employed when the validity study was conducted. The correlation between the overall score on work-related behaviors from all references at prehire, and the overall score on actual work behaviors from supervisor ratings

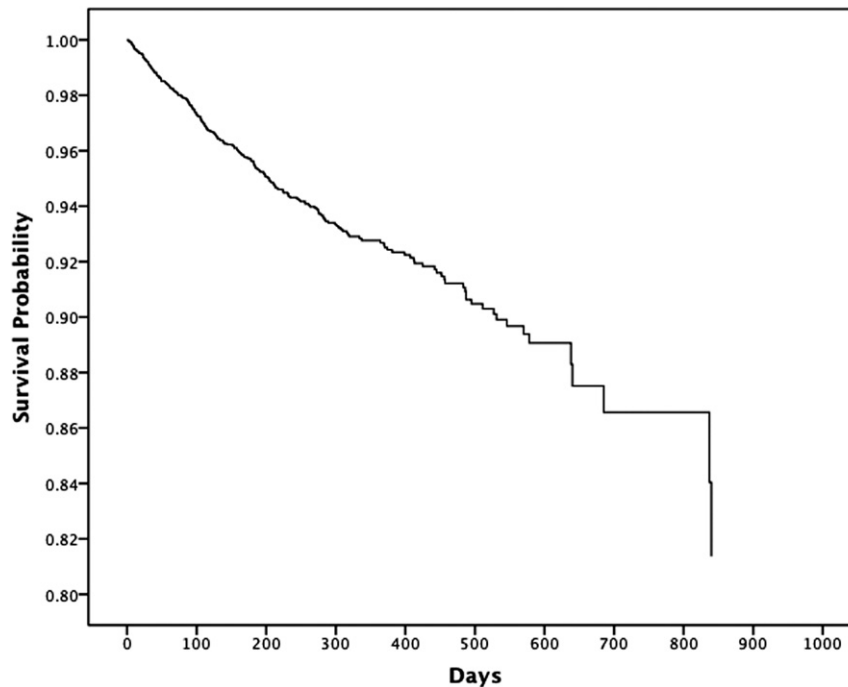


Figure 2. Survival function at one standard deviation below the mean for both reference response rate and overall score (prehire).

on the same items at posthire, was both practically significant ($r = .35$) and statistically significant ($p < .001$).

3.4. Group differences

Even though previous research suggests that a number of the aspects of this process, such as standardization and using supplied references, should reduce the potential for discrimination (Avery, McKay, & Wilson, 2007; Williamson, Campion, Malos, Roehling, & Campion, 1997), we nevertheless gathered data from our sample to explore this issue directly. Specifically, we investigated effects of age, sex, and race of the job candidate on the feedback provided by the references. We did not focus on statistical significance, because almost every result will be significant given the high sample size ($N = 150,000$).

Overall score was used as the criterion variable (i.e., average across all items in the survey). The effect sizes were all in the negligible or small range for each of the comparisons (age $d = 0.02$; sex $d = 0.12$; and race $d = 0.00$; see Table 5). These results provide additional support for this web-based multisource reference-checking technology in terms of showing little potential for adverse impact in key applicant subgroups.

3.5. Efficiency metrics

Descriptive statistics for the efficiency metrics are displayed in Table 6. On average, both candidates and references took a little over one day to complete their

Table 5. Analysis of group differences on overall score

Age						
39 or Younger			40 or Older			
<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>d</i>
87,951	6.50	0.39	40,528	6.49	0.39	0.02
Sex						
Female			Male			
<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>d</i>
86,273	6.51	0.39	45,587	6.46	0.38	0.12
Race						
White			Non-White			
<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>d</i>
91,234	6.50	0.39	36,595	6.50	0.39	0.00

Note. $d = (M_1 - M_2)/SD_{\text{pooled}}$, where Group 1 is in the left column and Group 2 is in the right column.
SD = standard deviation.

portions of the process ($M = 1.32$, $SD = 2.40$ for candidates and $M = 1.10$, $SD = 1.52$ for references). However, the distributions were positively skewed; thus, the median of less than one day to complete the required information was more representative of the distribution (median = 0.87 for candidates and median = 0.75 for ref-

Table 6. Descriptive statistics for efficiency metrics

Variable	M	SD	Median	Min	Max	Skew	Kurtosis
Candidate response time (days)	1.32	2.40	0.87	0.00	124.96	9.20	232.91
Reference response time (days)	1.10	1.52	0.75	0.00	118.83	11.59	456.35
Report turnaround time (days)	2.78	3.72	1.80	0.00	126.40	6.42	94.33
No. of references completed (total)	4.16	1.24	4.00	1.00	21.00	1.01	5.34
No. of references completed (managers)	2.29	1.07	2.00	0.00	21.00	0.70	1.75
Reference response rate	0.85	0.17	0.86	0.17	1.00	-0.80	-0.36

Note: $N = 150,000$. Days = business days. Reference response time refers to the median number of days across references for a given candidate (descriptive statistics are applied to this distribution of median days).
SD = standard deviation.

erences). The average report turnaround time was a little over 2.50 days ($M = 2.78$, $SD = 3.72$) with a median of 1.80 days. On average, candidates were rated by approximately four reference providers (overall) and two managers. The average reference response rate was 85%.

4. Discussion

The web-based multisource methodology used to gather feedback on job candidates from those who had previously worked with the individual demonstrated high internal consistency across five representative reference-checking surveys. A test-retest study in the applied setting yielded a 48.8% response rate. Test-retest reliability was slightly higher for feedback that came from managers (mean $r = .77$) compared with the test-retest reliability across all rater groups (mean $r = .71$). These reliabilities were greater than those reported for feedback gathered from structured letters of reference provided for MBA applicants (Zimmerman et al., 2010), possibly because of differences in methodologies and samples. For example, the current methodology used web-based technology (vs. letters of reference) to gather a second set of multisource ratings on actual job candidates (vs. MBA applicants), a 7-point rating scale (vs. a 5-point rating scale), and between 18 and 24 survey items (vs. 43 survey items).

It is difficult to make direct comparisons of these inter-rater agreement (ICC) results to previous research in this area, as contextual factors unique to reference checking may make such comparisons unfair. For one, the high-stakes nature of reference checking in the current sample may depress the reliability of reference-checking scores. Greguras and Robie (1998) reported ICCs that typical inter-rater agreement would be .38 for supervisors, .57 for peers, and .48 for direct reports for developmental multisource feedback ratings. Given the research showing that contextual factors such as the purpose of appraisal (e.g., administrative vs. developmental) affects the reliability of multisource ratings at least for some sources (i.e., direct reports), we believe that the ICCs for this web-

based multisource procedure are quite acceptable (Greguras, Robie, Schleicher, & Goff, 2003). We should note however that these levels of inter-rater reliability are typical of ratings gathered for reference checking (see Taylor et al., 2004).

There was a statistically significant relationship between overall reference score (prehire) and involuntary turnover. This result would likely have been even stronger but was likely reduced because of both incidental and direct range restriction on the overall score (prehire) variable. Incidental range restriction has likely occurred because the reference-checking process (in our experience) is typically conducted for those job applicants who have already been identified as finalists for a given position. These finalists have typically been screened on other measures correlated with reference-checking results, such as resume screening and face-to-face interviews. Range restriction effects may also be more direct in nature, such as when references are less likely to agree to divulge reference information to individuals having known or documented performance issues. A study by Grote et al. (2001) found that reference providers generally refused to write a recommendation for individuals with performance issues in almost one-quarter of the cases. Some authors suggest that letter-writers have an ethical duty to refuse to write or warn applicants of poor or neutral letters of reference (Range et al., 1991).

The lack of statistically significant relationships between reference (and candidate) response times and either voluntary or involuntary turnover may be the result of irrelevant variance in the predictors. Our assumption in using candidate response time as a predictive variable was that it might be indicative of a candidate with lack of initiative that in turn could affect organizational outcomes. Similarly, our assumption in using reference response time as a predictive variable was that reference providers would be more likely to respond quickly to provide reference information for candidates they held in high regard. In both instances, external factors such as candidate or reference provider workload would act to mitigate any predictive value.

It is likely that this latter range restriction effect on overall score (prehire) is actually responsible for the negative relationship between reference response rate and involuntary turnover. If reference providers are less likely to divulge reference information to candidates with performance issues, then lower response rates should exist for employees who are less likely to be retained. This result suggests that willingness of the references to comply with the reference request could be a positive predictor of employee behavior at work, such as turnover.

There was a significant, positive relationship between the overall ratings by references at prehire, and the overall ratings by supervisors on the same items at posthire. The correlation between pre and postscores was $r = .35$. The correlation between reference feedback and performance was higher than values previously reported (e.g., Hunter & Hunter, 1984; Reilly & Chao, 1982; Taylor et al., 2004). The correlation was in line with similar correlations reported by Zimmerman et al. (2010), and any differences may reflect both the reference-checking methodology as well as the criterion measure used. In the present study, supervisor evaluations were collected within the context of a study to investigate the quality of new hires, as compared with evaluations of performance from managers, coworkers, or class team members.

The reliability and validity data, as well as the results of the subgroup analyses (for race/ethnicity, gender, and age group), lend support to the use of this web-based multisource methodology for gathering reference feedback on job candidates. Specifically, this study found support for the methodology to predict two critical post-hire outcomes: involuntary turnover and work performance. The cost and time efficiency of this methodology, as well as its psychometric qualities, have the potential to make this tool a valuable component in the employee selection process.

The efficiency of web-based reference checking cannot be overstated. Participant organizations in a 2004 study (Society for Human Resource Management, 2005) reported that the predominant modality for conducting reference checks was the least efficient – the telephone (60% reported ‘always’ to use of telephone). The lesser used modality was the more efficient – Internet (8% reported ‘always’ and 28% reported ‘sometimes’). We have found in our practice that phone reference checks generally take 30–60 min to complete per candidate, and this component of the hiring process is completed after about 3–7 days with about 50% of references responding. Assuming a good reference check takes an average of 30 min, a company that hires 250 external job candidates a year will save 125 hours of recruiting time by eliminating phone-based reference checks.

These results suggest that this web-based prehire reference assessment can assist employers with four

major concerns in the area of reference checking, and these are the need for a process that (a) reduces bias, (b) is legally defensible, (c) gathers work-relevant behavioral information on which to differentiate job candidates, and (d) is time and cost efficient. Initial results did not reveal any biases in the overall score with respect to gender, age group or race/ethnicity of the job candidates.

5. Future research

Several areas appear promising for future research in this area. Applicant reactions to reference checks tend to be more favorable than personal contacts, graphology, honesty tests, or personality tests, but less favorable than biodata, cognitive tests, interviews, resumes, or work samples (Anderson, Salgado, & Hülsheger, 2010). We could find no studies that examined applicant reactions to Internet-based reference checking. However, in a review of the literature, Anderson (2003, p. 125) stated: ‘So far, almost universally positive applicant reactions have been reported for Internet-based testing batteries.’ In our own practice, in an approximately 1-year period between March 2011 and May 2012, over one million reference providers gave feedback on approximately a quarter million candidates. We received 50 complaints via phone or e-mail concerning our reference process during this time period (a very small fraction of 1% or 0.02%). We have not, however, systemically collected data on applicant reactions. Efforts are underway to do so.

The majority of the data reported in this article were US-based. Considerable divergence in personnel selection practices and legal contexts occurs across nations (cf. Huo, Huang, & Napier, 2002; Myers et al., 2008). Cultural influences might affect the relationships we found in our study. Moreover, constraints on how Internet-based data may be used in other countries should be considered.

Future research endeavors should examine the theoretical underpinnings for the relationships we found in the current study. Cross-validation in other contexts will strengthen arguments for generalizability given that a large proportion of our samples were derived from the healthcare industry. Other personnel selection instruments similar to reference checks (e.g., biographical data) have not been found to experience reductions in validity (i.e., validity decay) over time (see Breagh, 2009, p. 223 for a review of relevant studies). However, validity studies must also be conducted across time in the area of reference checking to examine the extent of possible validity decay.

A particularly interesting area for future research is the base rate of fraudulent references and methods to guard against them. We are not aware of any systematic

study that has examined the incidence of fraudulent references. In the area of personality testing, warnings of potential identification and the consequences of faking have shown to reduce faking (cf. Dwight & Donovan, 2003). It remains to be seen whether warnings affect the incidence of fraudulent references.

A final area for future research would be to examine rater–ratee race and sex effects on the feedback provided by the references. Research has been conducted in this area in performance evaluation (Mount, Sytsma, Hazucha, & Holt, 1997; Pulakos, White, Oppler, & Borman, 1989; Sackett & DuBois, 1991). We know of no such research that has examined these issues in the area of reference checking.

Notes

1. In April and May 2008, the first author conducted webcasts on reference checking for HR generalists and managers, hosted by companies providing recruitment process outsourcing.
2. Software is not purchased by the user or installed on the user's device; instead, it is accessed 24/7 via the Internet.
3. Across a wide variety of industry sectors, company sizes, and positions, the SkillSurvey Pre-Hire® multisource patent-pending tool has been used to gather reference feedback with Web 2.0 technology on 1.2 million job candidates to date. An average of 4.2 references per candidate have completed and submitted a survey in less than one business day, with an average reference response rate of 85%. The majority of all reference feedback (71%) is received back from the reference's corporate e-mail address. Approximately 55% of references gathered come from the candidate's former or current manager(s); teachers can serve as manager-type references for entry-level job candidates with no prior work experience.
4. We conducted turnover, validity, and group differences analyses by overall score instead of by dimension. A recent survey of SkillSurvey clients ($N = 215$) revealed that 28.6% of respondents found overall score and manager vs. non-manager breakouts in overall score to be the most important in evaluating a candidate (vs. 11.6% who responded that dimension scores were the most important).
5. Data used were from the SkillSurvey General Professional survey (18 items).

References

- Aamodt, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting performance with letters of recommendation. *Public Personnel Management*, 22, 81–90.
- Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment*, 11, 121–136.
- Anderson, N., Salgado, J. F., & Hülshager, U. T. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, 18, 291–304.
- Automatic Data Processing (ADP), Inc. (2008). The current state of recruitment technology: Technology's emerging role in the 'war for talent.' Available at http://pre-employmentservices.adp.com/forms/PES_Recruitment Research (accessed January 17, 2013).
- Avery, D. R., McKay, P. F., & Wilson, D. C. (2007). Relational demography and employment discrimination: Does similarity help? Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, New York.
- Bliss, W. (2001). *Legal, effective references: How to give and get them*. Alexandria, VA: Society for Human Resource Management.
- Breaugh, J. A. (2009). The use of biodata for employee selection: Past research and future directions. *Human Resource Management Review*, 19, 219–231.
- Bycio, P., Hackett, R. D., & Alvares, K. M. (1990). Job performance and turnover: A review and meta-analysis. *Applied Psychology: An International Review*, 39, 47–76.
- Deering, West Nursing Center, Div. of Hillhaven Corp. v. Scott, 787 S.W.2d 494, 1990 Tex. App. LEXIS 528 (Tex. App. El Paso 1990).
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16, 1–23.
- Fernandez-Araoz, C. (2007). *Great people decisions: Why they matter so much, why they are so hard, and how you can master them*. Hoboken, NJ: John Wiley & Sons.
- Gatewood, R. D., Feild, H. S., & Barrick, M. (2007). *Human resource selection* (6th ed.) Mason, OH: Thomson South-Western.
- Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, 83, 960–968.
- Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, M. III. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology*, 56, 1–21.
- Grote, C. L., Robiner, W. N., & Haut, A. (2001). Disclosure of negative information in letters of recommendation: Writers' intentions and readers' experiences. *Professional Psychology: Research and Practice*, 32, 655–661.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72–98.
- Huo, Y. P., Huang, H. J., & Napier, N. K. (2002). Divergence or convergence: A cross-national comparison of personnel selection practices. *Human Resource Management*, 41, 31–44.
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9, 202–220.
- Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology*, 92, 1638–1656.
- Madera, J. M., Hebl, R. M., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentive and communal differences. *Journal of Applied Psychology*, 94, 1591–1599.
- McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three

- standardized reference forms. *Military Psychology*, 13, 199–222.
- McLean, v. Kirby Co., 490 N.W.2d 229, 1992 N.D. LEXIS 173 (N.D. 1992).
- Michaels, E., Handfield-Jones, H., & Axelrod, B. (2001). *The war for talent*. Boston, MA: Harvard Business School Publishing.
- Morita, J. G., Lee, T. W., & Mowday, R. T. (1989). Introducing survival analysis to organizational researchers: A selected application to turnover research. *Journal of Applied Psychology*, 74, 280–292.
- Morita, J. W., Lee, T. W., & Mowday, R. T. (1993). The regression-analog to survival analysis: A selected application to turnover research. *Academy of Management Journal*, 36, 1430–1464.
- Mount, M. K., Sytsma, M. R., Hazucha, J. R., & Holt, K. E. (1997). Rater–ratee effects in developmental performance ratings in managers. *Personnel Psychology*, 50, 51–69.
- Muchinsky, P. M. (1979). The use of reference reports in personnel selection: A review and evaluation. *Journal of Occupational Psychology*, 52, 287–297.
- Myors, B., Lievens, F., Schollaert, E., Van Hove, G., Cronshaw, S., Mladinic, A. et al. (2008). International perspectives on the legal environment for selection. *Industrial and Organizational Psychology*, 1, 206–246.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.) New York: McGraw-Hill.
- Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, 74, 770–780.
- Range, L. M., Menyhert, A., Walsh, M. L., Hardin, K. N., Ellis, J. B., & Craddick, R. (1991). Letters of recommendation: Perspectives, recommendations, and ethics. *Professional Psychology: Research and Practice*, 22, 389–392.
- Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1–62.
- Robie, C., Ryan, A. M., Schmieder, R. A., Parra, L. F., & Smith, P. C. (1998). The relation between job level and job satisfaction. *Group & Organization Management*, 23, 470–495.
- Sackett, P. R., & DuBois, C. L. Z. (1991). Rater–ratee effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, 76, 873–877.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage.
- Society for Human Resource Management. (2005). *2004 Reference and background checking survey report*. Alexandria, VA: Society for Human Resource Management.
- Somers, M. J. (1996). Modelling employee withdrawal behaviour over time: A study of turnover using survival analysis. *Journal of Occupational and Organizational Psychology*, 69, 315–326.
- Somers, M. J., & Birnbaum, D. (1999). Survival versus traditional methodologies for studying employee turnover: Differences, divergences and directions for future research. *Journal of Organizational Behavior*, 20, 273–284.
- Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance*. New York: John Wiley & Sons.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.) Boston, MA: Allyn & Bacon.
- Tallahassee Furniture Co. v. Harrison, 583 So. 2d 744 (Fla. App. 1991).
- Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology*, 57, 745–772.
- U.S. Department of Labor. (1991). *Dictionary of occupational titles* (Rev. 4th ed.) Washington, DC: U.S. Government Printing Office.
- Williamson, L. G., Campion, J. E., Malos, S. B., Roehling, M. V., & Campion, M. A. (1997). Employment interview on trial linking interview structure with litigation outcomes. *Journal of Applied Psychology*, 82, 900–912.
- Wood, J. L., Schmidtke, J. M., & Decker, D. L. (2007). Lying on job applications: The effects of job relevance, commission, and human resource management experience. *Journal of Business Psychology*, 22, 1–9.
- Wood, R., & Payne, T. (1998). *Competency based recruitment and selection: A practical guide*. New York: John Wiley & Sons.
- Woska, W. J. (2007). Legal issues for HR professionals: Reference checking/background investigations. *Public Personnel Management*, 36, 79–89.
- Zimmerman, R. D., Triana, M. C., & Barrick, M. R. (2010). The criterion-related validity of observer ratings of personality and job-related competencies using multiple raters and multiple performance criteria. *Human Performance*, 23, 361–378.