SUBMISSION TYPE

Poster

TITLE

Automated Online Reference Checks: A Review and Empirical Investigation

ABSTRACT

Automated Online Reference Checks (AORCs) are a new, standardized reference check tool. To investigate the utility of AORCs, we addressed a series of research questions using data from an AORC system. Supervisors gave the most critical references, no sex biases were present, and the qualitative and quantitative ratings shared a positive relation.

PRESS PARAGRAPH

Automated online reference checks (AORCs) have revolutionized reference checking by reducing costly resources and increasing standardization. Although the practice of AORCs has skyrocketed, the research literature is in its infancy. We first discuss issues with traditional reference checks and review the literature on AORCs. We then address a series of research questions using data from an AORC system, Reference Hunter. We examine the relation between qualitative and quantitative responses to AORC items, the presence of sex biases, and differences due to type of reference (e.g., supervisor, colleague). Our findings suggest that AORCs are an effective means of standardizing the reference check process, and as a result, reduce biases that have polluted traditional reference checks.

WORDS

 2989

Automated online reference checks (AORCs) are a new phenomena that uses the internet to automate the reference checking process, a process that previously had been viewed as time consuming and prone to issues of poor validity, lack of candor among references, and perceived legal barriers to participating (see Hedricks et al., 2019). With AORCs, candidates are asked to supply contact information for references and then organizations use the AORCs to solicit reference checks over the internet, thus saving person hours involved in tracking down and arranging phone conversations. AORCs can also enforce standardization in background checks by requiring referees to respond to a fixed set of items, with no possibility of getting sidetracked by irrelevant conversations. This is a new area of practice with few empirical investigations into the utility, reliability, and validity of AORCs (Hedricks et al., 2013; Hedricks et al., 2018; Hedricks et al., 2019). In this manuscript, we examine data from one AORC system, Reference Hunter (The Hire Talent, 2019), to investigate some research questions that have yet to be examined in the research literature. Specifically, we examine the role of the sex and type of reference in influencing reference ratings, as well as the relation between qualitative and quantitative data in an AORC system.

**Issues with Traditional Reference Checks**

**Discretion with the Choice of References**

The most prevalently researched version of traditional reference checking is letters of recommendation, which are known to have low predictive validity (Reilly & Chao, 1982; Hunter & Hunter, 1984; Kuncel et al., 2014). Potentially contributing to the lack of predictive ability is the amount of discretion that candidates have in choosing referees. Nearly everyone can find someone who will say something positive about them. Klonsky and Oltmanns (2002) referred to this phenomenon as the “letter of recommendation problem” and “pal-serving bias.” In what is known as the “friendship effect,” Leising et al. (2010) demonstrated the difference between *liking* and *knowing* the candidate. Their study showed that referees who liked the candidate gave more positive ratings, whereas referees who did not like the candidate, but knew the candidate just as well, provided more accurate ratings. Supporting this finding, a meta-analysis by Sutton et al. (2013) revealed a strong positive relation between referee-liking and performance ratings.

Being that the candidate-referee relationship can influence the type of reference provided, candidates face a tough decision regarding whom they should select as referees. Supervisors, coworkers, family, and friends are commonly asked to serve as referees, and each may offer a unique perspective on the same individual. Previous research has demonstrated that supervisor, coworker, and self-ratings differ in validity. Studies repeatedly reveal that supervisors provide the most valid ratings, followed by coworkers, and then the self (Hoffman et al., 1991; Mount et al., 1994; Small & Diefendorff, 2006). In this manuscript, we examine whether rating sources differ in terms of positivity in an AORC system.

**Sex Bias**

Research investigating sex differences in performance evaluations and reference checks is prevalent. In their meta-analysis investigating the presence of sex bias in performance appraisals, Bowen et al. (2000) found that pro-male bias occurred in performance ratings when all of the raters were male. Regarding letters of recommendation, Madera et al. (2009) revealed that female candidates were described using communal terminology (e.g., kind, sympathetic) whereas male candidates were described using agentic terminology (e.g., assertive, independent). Additionally, letters of recommendation that contained communal language were negatively related to hiring decisions. Filippou et al. (2019) reported similar results in which letters of recommendation for male candidates contained significantly more agentic terms compared to females. Madera et al., (2019) examined the prevalence of doubt raisers in references and found that men and women referees used significantly more doubt raisers for female candidates, and evaluators viewed recommendations that contained doubt raisers less favorably. In this manuscript, we examine whether sex differences exist in AORC data.

**Existing Research on Standardized Reference Checks**

The AORC is a new tool, resulting in a scarce amount of previous literature from which to draw inferences. What is established is the validity of structured selection methods. Although there is an overwhelming amount of evidence in support of structured selection methods, many organizations still rely on unstructured methods such as supervisor intuition (Highhouse, 2008).

One potential benefit of AORCs is the structured format may decrease the likelihood of referees incorporating bias (e.g., sex bias) into their references (Hedricks et al., 2013; Madera et al., 2009). Bias can still occur with standardized reference checks, but the standardization forces the referees to focus their responses on the items found in the survey. Hedricks et al. (2013) examined the reliability and validity of AORCs and found them to have sufficient test-retest reliability and high internal consistency. Additionally, AORCs significantly predicted involuntary turnover, and there was a significant positive relation between the pre-hire ratings given by the referee and post-hire supervisor ratings.

AORCs still allow for referees to provide qualitative information, and in fact, it is a common practice. Hedricks et al. (2018) revealed that 80% of the reference checks they examined contained qualitative information, and the majority of the qualitative information pertained to soft skills, such as communication and ability to work with others.

Providing raters with standardized prompts is different than simply asking raters to write a letter of recommendation. In a study that surveyed 173 members of the medical field who wrote letter of recommendations for residency positions, Keim et al. (1999) found that 75% of evaluators felt that standardized letters of recommendation (SLORs) were better at discriminating between applicants compared to narrative letters of recommendation (NLORs); 84% said SLORs are easier to complete and rank applicants, and 90% said they would like to use SLORs going forward. Friedman et al. (2017) revealed that SLORs contained significantly less sex bias compared to NLORs. Although SLORs contain qualitative portions, NLORs are entirely unstandardized, allowing for rater bias to become problematic.

 **Initial Evidence of the Effectiveness of AORCs**

 Evidence from previous literature suggests that standardized reference checks are an improvement upon traditional unstructured methods, implying that AORCs may be a valid tool for predicting future job performance. Before making any conclusions regarding the utility of AORCs, additional research needs to be conducted. In this section, three research questions regarding AORCs are presented. To address each question, we present findings using data from an AORC system to provide initial evidence that AORCs are an effective reference check tool. All of the findings presented are intended to provide researchers and practitioners with a general idea of the type of information that AORCs yield, and whether AORCs are an effective means of obtaining references.

*Research Question 1*: Consistent with previous findings, do supervisors provide the most critical performance evaluations of candidates in AORCs?

*Research Question 2*: Does the standardized nature of AORCs reduce sex differences that are common in traditional unstandardized reference checks?

*Research Question 3*: Does the qualitative data acquired using standardized prompts in AORCs relate to the quantitative performance ratings provided by referees?

**Reference Hunter**

Reference Hunter is an AORC platform developed by The Hire Talent. Organizations who use Reference Hunter are instructed to email job candidates a link that allows the candidates to enter names, phone numbers, and emails of potential referees. Once those phone numbers and emails have been entered, potential referees are sent a link via email, automated phone call, and text message to a survey that includes a fixed set of items (though individual employers can customize the questions sent to referees).

The AORC that we examined contained quantitative and qualitative items in which the referees responded to regarding the candidate whose performance they were asked to evaluate. Before responding to the performance-related items, the referees indicated the type of relationship they had with the candidate (i.e., supervisor, coworker, etc.). For the quantitative items, referees rated the candidate’s overall performance, and they also indicated whether or not they would rehire or work with the candidate in the future. The three items that required qualitative responses were standardized across all surveys, and they asked the referees to describe the overall experience working with the candidate, why they rated the candidate’s performance the way they did, and what the candidate could have done to receive a higher performance rating.

**Participants**

A sample of 354 references for 181 candidates was acquired through Reference Hunter’s AORC system. Respondents were supervisors (56%), family/friends (12%), colleagues (23%), mentor/advisors (4%), or others (5%). Among the candidates, 57% were male and 43% were female. Among the referees, 63% were male and 37% were female. Candidates were applying for jobs that ranged from administrative to management roles.

 References were nested within candidates, which warranted the use of multilevel modeling to address Research Question 2. Determining an appropriate sample size for conducting multilevel analysis is a step that is often overlooked. Previous studies that observed multiple scores nested within individuals have recommended a minimum sample of 100 participants (Ohly et al., 2010), and Scherbaum and Ferreter (2009) stated that Level 2 sample sizes smaller than 30 (which is common in organizational research) may produce biased results. Our sample size surpasses these suggestions, implying the appropriateness of using multilevel modeling.

**Measures**

***Performance***

 Performance was measured with a single item in which the referees rated the candidates’ overall performance on a 9-point likert-type scale with 1 being the lowest and 9 being the highest possible rating. The item stated, “Nine being the highest, how would you rate (candidate name)’s overall performance in the role(s) they were in?”

***Rehire or Work With in the Future***

 The referees indicated whether they would rehire or work with the candidate in the future by responding “Yes” or “No” to an item that stated, “Assuming the role was appropriate for their experience, would you rehire or want to work with (candidate name) in the future?”

***Qualitative Items***

For the first qualitative item, the referees described their overall experience working with the candidate (OE). The prompt stated, “Tell me about your experience working with (candidate name).” The second qualitative item asked referees to explain the performance rating that they gave the candidate (WR). The prompt stated, “Tell me more about why you rated (candidate name)’s overall performance the way you did.” For the third qualitative item, the referees were asked to describe what the candidate could have done to receive a higher performance rating (DB). The prompt stated, “What could (candidate name) do to get from the overall performance rating you gave to the next score level?”

 **Positivity of Response.** For the three qualitative items, two of the authors read and rated the positivity of each response from 1 to 5. We calculated the ICC*A,1* statistic to evaluate interrater agreement and the ICC*C,1* statistic to evaluate interrater consistency for OE (ICC*A,1* = 0.67, 95% CI = [0.49, 0.77]; ICC*C,1* = 0.71, 95% CI = [0.65, 0.76]), WR (ICC*A,1* = 0.54, 95% CI = [0.42, 0.64]; ICC*C,1* = 0.57, 95% CI = [0.49, 0.65]), and DB (ICC*A,1* = 0.61, 95% CI = [0.48, 0.70]; ICC*C,1* = 0.64, 95% CI = [0.56, 0.71]). All ICC values surpassed moderate reliability standards (Koo & Li, 2016), therefore, the average of the two authors’ ratings was taken for each response in order to create an average score for OE, WR, and DB.

 **Word Count**. In addition to rating the positivity of the responses given for each of the qualitative items, the word count was also calculated for each response to determine whether there was a relation between the amount that a reference wrote and the other items.

**Results**

 Table 1 provides descriptive statistics and correlations between all observed variables excluding sex and referee type. Descriptive statistics of the dependent variables broken out by candidate sex are located in Table 2.

To address Research Question 1, we conducted a one-way ANOVA to examine whether performance ratings differed across referee types. The overall model was significant, *F*(4, 240) = 2.77, *p* < .05. A Tukey HSD post-hoc test revealed a significant difference in performance ratings between supervisor (*M* = 8.37, *SD* = 0.84) and family/friend (*M* = 8.83, *SD* = 0.47). The effect size for the mean difference (*d* = 0.68) exceeds Cohen’s (1988) cutoff for a moderate effect. Consistent with previous findings, supervisors rated candidates’ performance significantly lower compared to family/friends. All other pairwise comparisons were not significant.

A MANOVA was then conducted to examine whether word count totals for the three qualitative items differed depending on referee type. There was no significant multivariate effect, implying that word count totals did not statistically significantly differ across referee types. We also conducted a MANOVA to test if the positivity of responses given for the three qualitative items differed depending on referee type, and no significant multivariate effect was found.

 To address Research Question 2 we conducted multilevel modeling with the lme4 package (Bates et al., 2007) in R being that references were nested within candidates. Standard maximum likelihood estimation with random effects was used. For all analyses, the Level 1 variable was the dependent variable, and the Level 2 variables were candidate sex, referee sex, and the interaction of the two sex variables.

We first tested for sex differences in performance ratings, and found no statistically significant differences (see Table 3). We then tested for sex differences in response positivity. A separate test was conducted for OE, WR, and DB respectively, and no statistically significant differences were found (see Table 4). Lastly, we tested for sex differences in word count by conducting separate tests for OE, WR, and DB, and no statistically significant differences were found (see Table 5). Overall, no sex differences were found in the AORC data.

To address Research Question 3 correlations between performance and OE, WR, and DB response positivity and word count were observed. Performance was significantly positively correlated with OE (*r* = .25), WR (*r* = .43), and DB (*r* = .50) response positivity. For all three qualitative items, higher performance ratings were associated with more positive responses given by referees. Regarding word count, performance was significantly positively correlated with WR (*r* = .12). The lengthier the referees’ explanation for why they rated candidates a certain way was associated with higher performance ratings.

Point-biserial correlations between rehire and OE, WR, and DB response positivity and word count were also observed. Rehire was significantly positively correlated with OE (*rpb* = .26) and WR (*rpb* = .22), but not DB (*rpb* = .09) response positivity. For OE and WR, referees that provided positive responses were more likely to rehire or desire to work with the candidate in the future. There were no significant point-biserial correlations between rehire and OE, WR, and DB word count. Longer responses were not associated with whether a referee would rehire or desire to work with a candidate in the future.

**Conclusions**

 In this manuscript, we argue that traditional reference check methods are inadequate, and that AORCs are a new and potentially useful tool for assessing job candidates. We reviewed the existing literature on AORCs and came up with a set of research questions that we addressed using data from an AORC system.

Although the practice of AORCs is growing, the research literature in this area is just beginning. The few studies that have been published confirm that AORCs are a useful tool and provide significant advantages compared to traditional reference checks. Not only are AORCs more convenient than traditional reference checks, but initial evidence suggests that responses provided in AORCs are valid predictors of turnover intention and future supervisor performance ratings, while also demonstrating acceptable levels of test-retest reliability and internal consistency (Hedricks et al., 2019). In addition, referees are more likely to respond to online systems.

 Our brief empirical examination highlights the importance of understanding the source of the references, finding similar results that others have found for 360 degree feedback systems. References from friends and family tend to be inflated and so should be discounted when considering references. Similar to previous findings, our results suggest that supervisors provide the most critical evaluations of candidates. In addition, the effects of sex of referee and candidate are important and need to be considered with traditional reference checks, as they are prone to sex differences (e.g., Filippou et al., 2019; Madera et al., 2019). We demonstrated that AORCs are an effective means of reducing sex differences in both the quantitative and qualitative responses provided by referees.

 Finally, we found that the amount of writing that referees did in the qualitative responses was positively related to their overall performance ratings and their self-reported likelihood of rehiring. Many AORC systems rely on quantitative responses but provide the ability for referees to include qualitative responses. In this system, the qualitative responses were optional, though generally completed. It is unclear how to best use the qualitative responses in conjunction with the quantitative responses, but our study shows that the amount of words written correlates with the positive tone of the qualitative responses as well as the quantitative data. In general, if people have something negative to say about an individual, they are less likely to write a lot about them. AORC systems that want to extract negative information about candidates might need to seek other mechanisms.

 Our study has limitations. The performance measure was a one-item measure and so it was impossible to assess its reliability, though it was positively correlated with the likelihood of rehiring single-item measure as well, thus giving us confidence on the validity of the measure. In addition, we did not have external data to follow up to see whether the hiring decisions made based on recommendations from this system had predictive validity. Work by Hedricks and colleagues demonstrated, however, that AORC systems can provide predictive validity.

 Given the relatively small sample, our findings should be replicated with larger samples and additional AORC systems, but given the paucity of research in this area, we felt that these data provide additional and needed information that further helps explain the functioning of AORC systems and helps support its growing practice.

**References**

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

Bowen, C. C., Swim, J. K., & Jacobs, R. R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology*, *30*(10), 2194–2215.<https://doi.org/10.1111/j.1559-1816.2000.tb02432.x>

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Filippou, P., Mahajan, S., Deal, A., Wallen, E. M., Tan, H. J., Pruthi, R. S., & Smith, A. B. (2019). The presence of gender bias in letters of recommendations written for urology residency applicants. *Urology*, *134*, 56-61. <https://doi.org/10.1016/j.urology.2019.05.065>

Friedman, R., Fang, C. H., Hasbun, J., Han, H., Mady, L. J., Eloy, J. A., & Kalyoussef, E. (2017). Use of standardized letters of recommendation for otolaryngology head and neck surgery residency and the impact of gender. *The Laryngoscope*, *127*(12), 2738-2745. <https://doi.org/10.1002/lary.26619>

Hedricks, C. A., Robie, C., & Oswald, F. L. (2013). Web‐based multisource reference checking: An investigation of psychometric integrity and applied benefits. *International Journal of Selection and Assessment*, *21*(1), 99-110. <https://doi.org/10.1111/ijsa.12020>

Hedricks, C. A., Rupayana, D. D., Fisher, P. A., & Robie, C. (2019). Factors affecting compliance with reference check requests. *International Journal of Selection and Assessment*, *27*(2), 139-151. <https://doi.org/10.1111/ijsa.12243>

Hedricks, C. A., Rupayana, D., Puchalski, L., & Robie, C. (2018). Content of qualitative feedback provided during structured, confidential reference checks. *Personnel Assessment and Decisions, 1*(4). https://doi.org/10.25035/pad.2018.004

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333-342.  <https://doi.org/10.1111/j.1754-9434.2008.00058.x>

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self versus supervisor ratings. *Personnel Psychology, 44*(3), 601–619.  <https://doi.org/10.1111/j.1744-6570.1991.tb02405.x>

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72.<https://doi.org/10.1037/0033-2909.96.1.72>

Keim, S. M., Rein, J. A., Chisholm, C., Dyne, P. L., Hendey, G. W., Jouriles, N. J., ... & Wightman, J. M. (1999). A standardized letter of recommendation for residency application. *Academic Emergency Medicine*, *6*(11), 1141-1146.  <https://doi.org/10.1111/j.1553-2712.1999.tb00117.x>

Klonsky, E. D., & Oltmanns, T. F. (2002). Informant‐reports of personality disorder: Relation to self‐reports and future research directions. *Clinical Psychology: Science and Practice*, *9*(3), 300-311.  <https://doi.org/10.1093/clipsy.9.3.300>

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.<https://doi.org/10.1016/j.jcm.2016.02.012>

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment*, *22*(1), 101–107.<https://doi.org/10.1111/ijsa.12060>

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of personality and social psychology*, *98*(4), 668. [https://doi.org/10.1037/a0018771](https://psycnet.apa.org/doi/10.1037/a0018771%22%20%5Ct%20%22_blank)

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591. http://dx.doi.org/10.1037/a0016539

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2019). Raising doubt in letters of recommendation for academia: gender differences and their impact. *Journal of Business and Psychology*, *34*(3), 287-303. https://doi.org/10.1007/s10869-018-9541-1

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79*(2), 272. <https://doi.org/10.1037/0021-9010.79.2.272>

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research: An introduction and some practical recommendations. *Journal of Personnel Psychology*, *9*(2), 79. <https://doi.org/10.1027/1866-5888/a000009>

Reilly, R.R. & Chao, G.T. (1982), Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*: 1-62.  <https://doi.org/10.1111/j.1744-6570.1982.tb02184.x>

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347–367. <https://doi.org/10.1177/1094428107308906>

Small, E. E., & Diefendorff, J. M. (2006). The impact of contextual self-ratings and observer ratings of personality on the personality–performance relationship1. *Journal of Applied Social Psychology, 36*(2), 297–320. https://doi.org/10.1111/j.0021-9029.2006.00009.x

Sutton, A. W., Baldwin, S. P., Wood, L., & Hoffman, B. J. (2013). A meta-analysis of the relationship between rater liking and performance ratings. *Human Performance*, *26*(5), 409-429. <https://doi.org/10.1080/08959285.2013.836523>

The Hire Talent (2019). Reference Hunter.

<https://www.preemploymentassessments.com/testing/reference-hunter/> (accessed 9/10/2020

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Means, standard deviations, and correlations with confidence intervals* |
|  |  |  |  |  |  |  |  |  |  |
| Variable | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|  |  |  |  |  |  |  |  |  |  |
| 1. Performance | 8.52 | 0.73 |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 2. Rehire | - | - | **.33\*** |   |   |   |   |   |   |
|   |   |   | [.23, .43] |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 3. OEWC | 30.63 | 37.07 | .10 | .05 |   |   |   |   |   |
|   |   |   | [-.01, .20] | [-.06, .16] |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 4. WRWC | 22.10 | 22.56 | **.12\*** | .08 | **.48\*** |   |   |   |   |
|   |   |   | [.01, .23] | [-.03, .19] | [.39, .55] |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 5. DBWC | 9.80 | 11.37 | .00 | .07 | **.30\*** | **.42\*** |   |   |   |
|   |   |   | [-.11, .11] | [-.04, .18] | [.20, .39] | [.33, .50] |   |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 6. OEPR | 4.06 | 0.74 | **.25\*** | **.26\*** | **.52\*** | **.29\*** | **.22\*** |   |   |
|   |   |   | [.14, .36] | [.15, .36] | [.44, .60] | [.19, .39] | [.11, .33] |   |   |
|   |   |   |   |   |   |   |   |   |   |
| 7. WRPR | 4.04 | 0.68 | **.43\*** | **.22\*** | **.25\*** | **.58\*** | **.29\*** | **.31\*** |   |
|   |   |   | [.34, .52] | [.11, .33] | [.14, .36] | [.50, .65] | [.18, .39] | [.20, .41] |   |
|   |   |   |   |   |   |   |   |   |   |
| 8. DBPR | 3.98 | 0.72 | **.50\*** | .09 | .09 | **.13\*** | **.16\*** | **.16\*** | **.27\*** |
|   |   |   | [.40, .58] | [-.04, .21] | [-.03, .21] | [.00, .24] | [.04, .28] | [.04, .28] | [.16, .39] |
|   |   |   |   |   |   |   |   |   |   |
| *Note.* *M* = mean, *SD* = standard deviation. Values in square brackets indicate the 95% confidence interval for each correlation. OE = overall experience, WR = why rated, DB = do better, WC = word count, PR = positivity of response. \* indicates *p* < .05. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2 |  |  |  |  |
|  |  |  |  |  |
| *Means and standard deviations split by sex of candidate* |
|  |  |  |  |  |
|  | Male | Female |
| Variable | *M* | *SD* | *M* | *SD* |
|  |  |  |  |  |
| Performance | 8.44 | 0.81 | 8.60 | 0.60 |
| OEWC | 31.22 | 36.09 | 31.92 | 40.29 |
| WRWC | 22.81 | 22.17 | 21.83 | 23.90 |
| DBWC | 11.21 | 12.27 | 8.24 | 10.14 |
| OEPR | 3.98 | 0.77 | 4.17 | 0.69 |
| WRPR | 4.01 | 0.69 | 4.08 | 0.63 |
| DBPR | 3.95 | 0.77 | 4.06 | 0.65 |

*Note.* *M* = mean, *SD* = standard deviation. OE = overall experience,

WR = why rated, DB = do better, WC = word count, PR = positivity of response.

|  |  |  |
| --- | --- | --- |
| Table 3 |  |  |
|  |  |  |
| *MLM results for candidate sex, referee sex, and the interaction predicting performance* |
|  |  |  |
|  | Performance |
| Predictors | Estimate | *t* |
|  |  |  |
| Candidate sex | 0.12[-0.13, 0.38] | 0.94 |
| Referee sex | 0.16[-0.11, 0.42] | 1.16 |
| Interaction | -0.01[-0.38, 0.37] | -0.04 |
| *Note.* Values in square brackets indicate the 95% confidence interval for each estimate. Candidate and referee sex were coded 0 = male, 1 = female. OE = overall experience, WR = why rated, DB = do better, WC = word count, PR = positivity of response. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 4 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *MLM results for candidate sex, referee sex, and the interaction predicting positivity of response* |
|  |  |  |  |  |  |  |
|  | OEPR | WRPR | DBPR |
| Predictor | Estimate | *t* | Estimate | *t* | Estimate | *t* |
|  |  |  |  |  |  |  |
| Candidate sex | 0.18[-0.08, 0.44] | 1.40 | -0.01[-0.25, 0.24] | -0.04 | 0.19[-0.11, 0.48] | 1.26 |
| Referee sex | 0.02[-0.25, 0.30] | 0.16 | 0.11[-0.14, 0.37] | 0.90 | 0.16[-0.13, 0.45] | 1.11 |
| Interaction | -0.02[-0.41, 0.37] | -0.11 | 0.07[-0.30, 0.44] | 0.37 | -0.16[-0.59, 0.27] | -0.74 |
| *Note.* Values in square brackets indicate the 95% confidence interval for each estimate. Candidate and referee sex were coded 0 = male, 1 = female. OE = overall experience, WR = why rated, DB = do better, WC = word count, PR = positivity of response. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 5 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *MLM results for candidate sex, referee sex, and the interaction predicting word count* |
|  |  |  |  |  |  |  |
|  | OEWC | WRWC | DBWC |
| Predictor | Estimate | *t* | Estimate | *t* | Estimate | *t* |
|  |  |  |  |  |  |  |
| Candidate sex | -3.35[-16.07, 9.38] | -0.52 | -1.26[-8.78, 6.26] | -0.33 | -4.03[-7.67, -0.40] | -2.18 |
| Referee sex | -2.37[-15.87, 11.14] | -0.35 | 0.90[-7.33, 9.14] | 0.22 | 0.41[-3.63, 4.45] | 0.20 |
| Interaction | 9.35[-9.70, 28.40] | 0.97 | -0.17[-11.79, 11.43] | -0.03 | 1.40[-4.29, 7.10] | 0.49 |
| *Note.* Values in square brackets indicate the 95% confidence interval for each estimate. Candidate and referee sex were coded 0 = male, 1 = female. OE = overall experience, WR = why rated, DB = do better, WC = word count, PR = positivity of response. |