

Electronic Thesis and Dissertation Repository

8-24-2021 1:00 PM

Improving the Accuracy of References: The Effects of Felt Accountability on Personality Ratings

Cullen W. D. McCurrach, *The University of Western Ontario*

Supervisor: Goffin, Richard D., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Psychology

© Cullen W. D. McCurrach 2021

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Industrial and Organizational Psychology Commons](#)

Recommended Citation

McCurrach, Cullen W. D., "Improving the Accuracy of References: The Effects of Felt Accountability on Personality Ratings" (2021). *Electronic Thesis and Dissertation Repository*. 8044.
<https://ir.lib.uwo.ca/etd/8044>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

References are a useful preemployment assessment for gathering unique personality information about job applicants. However, despite the capacity for external observers to provide valid personality ratings, references tend to be less accurate than similar preemployment assessments. This problem is important to address, since almost all organizations use references of some kind when hiring. This study attempted to improve the accuracy of the personality ratings provided in references by using two novel interventions believed to improve referees' felt accountability to the hiring organization. These interventions included (a) a more elaborative rating format that asked participants to provide the reasons underlying each of their quantitative personality ratings, and (b) a monetary incentive that was issued at the start of the rating task. Two-hundred and eleven ($N = 211$) supervisors and managers completed a personality rating form while playing the role of a referee for four prototypical call center employees shown in a series of work-sample videos. Participants were randomly assigned to one of four conditions in a 2 X 2 factorial design with Cronbach's (1955) four accuracy components as dependent variables. The results suggest that the monetary incentive did not appear to meaningfully impact participants' felt accountability or their personality rating accuracy. The reason giving requirements also had no effects on felt accountability, but interestingly had unintended negative consequences on participants' personality rating accuracy. Possible explanations for these findings are discussed, along with study limitations and potential avenues for future research on references.

Keywords: references, personality, accountability, personality judgement, reason giving, monetary incentive

Summary for Lay Audience

References are a useful tool for gathering personality information about job applicants that is otherwise difficult to obtain. However, research suggests that references are often provided inaccurately, which reduces their overall effectiveness compared to other similar hiring tools. This problem is important to address, since nearly all organizations use some form of reference check when hiring. This study was conducted to investigate two new ways to improve the accuracy of the personality ratings provided in references. These two new ways included (a) altering the reference form so that participants had to describe the reasons underlying their personality trait ratings, and (b) altering whether or not participants received a monetary incentive as a token of appreciation for their ratings. It was predicted that each of these changes would lead participants to feel more accountable to the hiring organization, thereby improving the accuracy of their personality ratings. However, the results indicate that the monetary incentive had no significant impacts on participants' felt accountability or rating accuracy, whereas the reason giving requirements had unintended negative consequences on the accuracy of their personality ratings. This research paradoxically suggests a potential new avenue for explaining rating inaccuracies in personality references. Study limitations and potential future directions for research on references are discussed.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Richard Goffin. I have greatly appreciated Rick's insights and guidance throughout the highs and lows of graduate school, and I am extremely thankful for his numerous contributions to my professional development. Pandemic aside, Rick continued to challenge and support me in so many ways that have me thrilled to begin working more closely with him in the future.

Second, I would like to thank my supervisory committee, Dr. Alex Benson, and Dr. John Meyer. As committee members and professors, Alex and John both provided me with unique and insightful feedback that has continued to resonate throughout the execution of this project. I would like to thank them for their support in each respective domain, and I look forward learning more from them in the coming years.

Third, I would like to thank my family for supporting me throughout all of my academic pursuits. Mom and Dad, I feel each of your supports uniquely from a distance and I think about them more than you know. Granny and Papa, I cannot begin to express how fortunate I am to feel yours from an arms length.

Fourth, I would like to thank each of the organizations who supported this study. I greatly appreciate the willingness and enthusiasm of Magellan Solutions, Hipronary School and TrainerTainment, who each provided this study with such high-quality stimulus materials.

Finally, I would like to thank each of the expert raters who volunteered to assist with developing such a critical component of my thesis. Noelle Baird, Jocelyn Brown, Sarah Carver, Jose Espinoza, Zhuo Li, Jennifer Lynch, and Alexander McGregor, your support means a lot to me.

Table of Contents

Abstract	ii
Summary for Lay Audience	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vi
List of Figures	vii
List of Appendices	viii
Introduction	1
References	3
Reference Letters	3
Structured References	4
Rating Situation	4
Accountability	6
Felt Accountability	7
Reason Giving	8
Monetary Incentives	11
Favourability	13
Method	14
Participants	14
Procedure and Study Design	15
Materials	18
Results	24

Personality Rating Accuracy	24
Reason Giving	24
Monetary Incentive	27
Felt Accountability	27
Discussion	30
Reason Giving effects	31
Monetary Incentive effects	34
Limitations	36
Future Directions	37
Conclusion	39
References	40
Appendices	60
Curriculum Vitae	81

List of Tables

Table 1. Correlation Matrix for the Study Variables of Interest	54
Table 2. Means and Standard deviations for all Accuracy Components Across Conditions	55
Table 3. Univariate F-Tests for the Main Effect of Reason Giving	56
Table 4. Univariate F-Tests for the Main Effect of the Monetary Incentive	57
Table 5. Felt accountability as a Function of Reason Giving and the Monetary Incentive	58
Table 6. Mediation Analyses for the Role of Felt Accountability	59

List of Figures

Figure 1. Experimental Design	17
Figure 2. Cronbach's (1955) Accuracy Components	22
Figure 3. Significant Main Effects of Reason Giving	26
Figure 4. Hypothesized Mediation Models	29

List of Appendices

Appendix A. Ethics Approval	60
Appendix B. Cover Story	62
Appendix C. Reason Giving Manipulation	63
Appendix D. Monetary Incentive Manipulation	64
Appendix E. Personality Rating Form	65
Appendix F. Felt Accountability Scale	66
Appendix G. Affective Reaction Items	67
Appendix H. Participant Effort Items	68
Appendix I. Demographics Questionnaire	69
Appendix J. Letters of Information + Consent	71
Appendix K. Prolific Study Information	77
Appendix L. Study Debrief	80
Appendix M. Curriculum Vitae	81

Introduction

References are a ubiquitous preemployment assessment (Aamodt, 1999; Arvey, 1979; Hedricks et al., 2013; Hedricks et al., 2018) that provide incremental applicant information beyond general intelligence and previous job experience (Beason & Belt, 1976, Hedricks et al., 2018). One critical component of references is that they focus on gathering applicant personality information (Aamodt, 1999; McCarthy & Goffin, 2001; Lopresto et al., 1985; Peres & Garcia, 1962; Sleight & Bell, 1954), which has been shown to be a valid and reliable predictor of job performance (Barrick et al., 2001). Observer reports of personality in particular (e.g., those provided through references; Furnham, 2017) tend to provide incremental validity over self-reports (Oh et al., 2011), as they effectively circumvent some of the limitations associated with personality tests in high-stakes scenarios (e.g., faking and impression management; Hedricks et al., 2013; Hedricks et al., 2018; Knouse, 1987; Zimmerman et al., 2010). However, despite the fact that they are inherently dispositional (Knouse, 1987; Norman, 1963; Peres & Garcia, 1962) references tend to yield some of the lowest validity estimates among comparable preemployment assessments (Aamodt & Williams, 2005; Furnham, 2017; Hunter & Hunter, 1984; Reilly & Chao, 1982). This problem is important to address, since most organizations use some form of reference when hiring (Society for Human Resource Management, 2005), despite the paucity of research examining their validity as a preemployment assessment (Aamodt & Williams, 2005; Taylor et al., 2004).

Meta-analytic research has demonstrated that external observers can generally provide accurate personality judgements of others (Connelly & Ones, 2010), although it remains unclear why this does not occur in references. Earlier research suggested that low knowledge about applicants, general unreliability, rater idiosyncrasies and rater leniency serve to reduce the

validity of references (Aamodt et al., 1993). Yet, more recent advances have shown that people can make relatively accurate personality judgements under conditions of low familiarity (Sheppard et al., 2011; Schmidt Mast et al., 2011) and that structured, standardized references are generally reliable (Taylor et al., 2004) and are better able to control for rater and reader idiosyncrasies (McCarthy & Goffin, 2001; Hedricks et al., 2013). It is therefore surprising that little research has investigated manipulations that could improve the accuracy of referee personality ratings, aside from a small body of studies highlighting the benefits of rater confidentiality (Ceci & Peters, 1984; Knouse, 1985) and the use of forced choice (Carroll & Nash, 1972) and relative rating formats (McCarthy & Goffin, 2001).

Building upon these studies to explore novel situational interventions that could improve rating accuracy may be particularly useful, given that there is often little reason to believe that referees are motivated to make critical, accurate judgements in this context (Knouse, 1983, 1987; Miller & Rybroeck, 1988). For example, it is currently unknown if referees feel accountable to hiring organizations for providing critically accurate personality ratings, versus simply completing their references on time or in line with the applicants' best interests (e.g., Colarelli et al., 2002). As such, the goal of this research was to explore the effects of two new interventions to an online reference rating situation, with a focus on their capacity to foster referees' felt accountability to the hiring organization and improve the accuracy of their personality ratings.

The purpose of this study was to investigate the effects of Reason Giving requirements and a Monetary Incentive on referees' felt accountability to the hiring organization as well as the accuracy of their personality ratings. By manipulating the reference form to either include or not include Reason Giving requirements, and by manipulating the rating situation to either include or not include a Monetary Incentive, this study provides novel theoretical insights towards the

effects of referees' felt accountability on the accuracy of their personality ratings. By examining the effects of Reason Giving requirements and a Monetary Incentive on the accuracy of referees' personality ratings, this study provides practical evidence towards the utility of two practical interventions that could potentially be used to improve the accuracy of references.

References

Broadly, to *reference* is to mention or make use of a source of information in order to ascertain something (Merriam-Webster, n.d.). This fundamental principle takes on many forms as a preemployment assessment (i.e., a phone call, (e-)mail, fax, meeting, or online survey; SHRM, 2005), so there tends to be little consensus on the terminology used to describe references in the literature (Muchinsky, 1979). In this report, the term "reference" refers to a descriptive evaluation of an applicant that is completed by an external observer with the goal of informing a personnel selection decision (McCarthy & Goffin, 2001). The term "referee" refers to the individual who completes the reference on behalf of the applicant (e.g., a former employer, mentor, or colleague; Furnham, 2017).

Reference letters

Earlier research on references was centered on unstructured *letters of recommendation* (LORs), which brought forth a wide array of critiques towards their psychometric properties (Baxter et al., 1981; Hunter & Hunter, 1984; Mosel & Goheen, 1952, 1958; Schmidt & Hunter, 1998). These criticisms prompted scholars to question the overall utility of references, which led to the discovery that LORs from a single referee tended to be more similar for two unique applicants than those made for a single applicant taken from two referees (Baxter et al., 1981). Because narrative LORs seemingly failed to uphold the conventional metrics for reliability and validity, they were at the time deemed as an inadequate basis for personnel selection (Baxter et

al., 1981). This conclusion echoed earlier calls to increase the amount of standardization in references (Holder, 1962), and sparked the pursuit of more structured, contemporary formats.

Structured references

In contrast to narrative LORs, research on structured reference formats (i.e., Carroll & Nash, 1972; Peres & Garcia, 1962; Sleight & Bell, 1954) has shown more promising evidence of validity (Browning, 1968, Mosel & Goheen, 1959). Specifically, by pre-emptively informing referees about the personality dimensions of interest, and by collecting their trait ratings in a standardized and job-relevant way, structured references have been shown to reduce rater idiosyncrasies (Hedricks et al., 2013; McCarthy & Goffin, 2001) and yield greater reliability (Taylor et al., 2004) and criterion validity (Aamodt et al., 1993; McCarthy & Goffin, 2001; Taylor et al., 2004; Zimmerman et., 2010) compared to their unstructured counterparts. Nevertheless, despite these notable advancements, structured references are still faced with situational constraints that are not as easily controlled for by hiring organizations.

Rating situation

Scholars have long argued that referees are generally reluctant to provide critical, accurate ratings to hiring organizations (e.g., Blum & Naylor, 1968). For example, research suggests that referees may feel a heightened desire to rate inaccurately as a function of their personal relationships with applicants (Colarelli et al., 2002), or because they are often chosen under the assumption that they will provide favourable evaluations (Aamodt et al., 1993). Even for those who are less personally connected to job applicants, because there is little benefit to referees for providing negative (albeit honest) information, they may still opt to provide overly positive ratings due to the mere undesirability and/or difficulties associated with providing negative reviews (Baron, 1993; Robiner et al., 1998). These situational dynamics can be further

compounded by the fact that referees may expose themselves to serious legal consequences (e.g., defamation of character charges) in cases where applicants are made aware of receiving negative references (Woska, 2007). The reference rating situation can therefore be argued to shift the overarching interests of referees to be more in line with those of job applicants (e.g., getting hired), which may in-part explain the prevalence of inflated and inaccurate references in this context (Aamodt et al., 1993; Colarelli et al., 2002). However, a small body of studies has also found promising ways to reduce referees' embellishments and increase the accuracy of their personality ratings.

Research has shown that using forced-choice prompts, in which all possible response sets contain an equally positive valence can create ambiguity towards the hiring organizations' ideal applicant profile and improve the validity of references (Carroll & Nash, 1972). Similarly, studies have found that asking referees to engage in perspective-taking to make relative (versus absolute) ratings can ground applicants in more representative comparison groups and elicit more accurate personality ratings (e.g., Goffin et al., 2009; McCarthy & Goffin, 2001). Finally, studies have highlighted that guaranteed rater confidentiality can reduce the presence of leniency in references (Ceci & Peters, 1984; Knouse, 1985) and improve their face validity compared to open-letter conditions (Schaffer et al., 1976; Schaffer & Tomarelli, 1981).

Although these advances highlight the capacity of situational interventions to improve the accuracy of references, the literature has yet to move beyond these factors to investigate the effects of other potentially powerful mechanisms. For instance, studies have yet to consider the effects of referees' felt accountability in this setting, despite the fact that accountability has been previously linked to more effortful, complex, and accurate interpersonal judgements in other selection (Brtek & Motowidlo, 2002; Mero & Motowidlo, 1995; Morgan et al., 2013)

performance evaluation (Mero et al., 2007), and personality prediction contexts (Tetlock & Kim, 1987). Additionally, previous information systems research suggests that subtle, non-invasive manipulations to digital user interfaces can heighten employees' accountability and reduce their intentions to violate their access policies (Vance et al., 2015). Taken together, there is compelling evidence to suggest that situational interventions fostering referees' felt accountability to the hiring organization may improve the accuracy of their personality ratings.

Accountability

Accountability is as a universal feature of decision-making environments (Tetlock, 1985), and a foundational element of organizational behaviour (Hall et al., 2003; Hochwarter et al., 2007). Accountability refers to the "implicit or explicit expectation that ones' decisions or actions will be subject to evaluation by some salient audience(s), with the belief that there exists the potential to receive either rewards or sanctions based on these expected evaluations" (Hall et al., 2003, p. 33).

Tetlock's (1985, 1992) phenomenological perspective posits that accountability is a perceptual state (i.e., a state of mind, rather than a state of affairs; Frink & Klimoski, 1998) that can range from *implicit* feelings of potentially being called upon to explain one's choices, to *explicit* expectations that they must always be formally justified. This outlook centers on individuals' internal *felt* accountability (Frink & Klimoski, 1998; Hall et al., 2017), while acknowledging that it is in-part rooted in their perceptions of the external working environment (Hall et al., 2007; Laird et al., 2009). Because employees may not always choose to abide by explicit organizational rules, inasmuch as they may feel implicitly accountable for their behaviours without formal constituents (Desai & Kouchaki, 2015; Schlenker & Weigold, 1989), the phenomenological perspective suggests that they can be viewed as "enacting" their own

unique levels of accountability (Hall & Ferris, 2011; Lewin, 1936). Taken in context, this theory suggests that referees may provide more or less accurate ratings to hiring organizations as a function of their felt accountability towards them.

Felt accountability

Schlenker and colleagues' accountability pyramid (Schlenker, 1986; Schlenker et al., 1994) highlights how perceptions of accountability are contingent upon individual responsibility. According to this model, individual responsibility is an additive function of the links between ones' job *prescriptions* (i.e., the rules that outline the goals of their tasks, how they should be realized, and their performance criteria), *events* (i.e., their unique decisions or behaviours and their associated or anticipated consequences), and *identities* (i.e., their central images and characteristics that connect to both their job prescriptions and events). Stated differently, individuals are believed to feel responsible to the extent that clear and objective prescriptions are defined for a specific event, wherein they are bound to this event and its prescriptions through their own identities, in ways that highlight their unique control over its processes and outcomes (Schlenker 1986; Schlenker et al., 1994). By way of this responsibility, individuals are said to feel accountable when they perceive that some evaluative audience is "looking down" on their behaviours (Schlenker 1986; Schlenker et al., 1994). As such, to the extent that referees feel individually responsible for their personality ratings, they could be made to feel more implicitly accountable to hiring organizations insofar as this responsibility is made salient throughout the rating situation.

Extant research suggests that referees do feel responsible for their ratings, such that they could be made to feel more accountable to hiring organizations for accuracy. In line with Schlenker's (1986) framework, Jones and Harrison (1982) suggest that teachers and faculty may

provide more accurate references as a function of their identities, since being overly lenient may not be in the best interest of other schools and may jeopardize their credibility. Large scale studies have also demonstrated that referee response times can predict involuntary turnover in job incumbents (Hedricks et al., 2013), which suggests that they may experience internal conflicts when asked to provide accuracy-seeking organizations with ratings for incompatible job applicants. Further studies have shown that referees may even refuse to provide ratings for underperformers (Grote et al., 2001), since doing so may spark ethical dilemmas (Range et al., 1991). This evidence converges with research on accountability cross-pressures, which demonstrates how conflicting sets of accountability demands can elicit stress that leads decision-makers to delay finalizing their conclusions, and/or remove themselves from a decision situation entirely (Green et al., 2000; Tetlock & Boettger, 1994).

Taken together, it appears that referees do feel individually responsible for their ratings, such that they could be made to feel more accountable to hiring organizations through changes to their rating situation (Lerner & Tetlock, 1999; Tetlock, 1992). Extant research suggests two practical ways to facilitate these changes: (a) implementing Reason Giving requirements, which could be done through by requesting qualitative justifications of the reasons underlying referees' quantitative personality ratings (e.g., Lerner & Tetlock, 1999) and (b) emphasizing the exchange relationship between the referee and the hiring organization, which could be implemented by issuing referees a Monetary Incentive as a token of appreciation for their services (e.g., Colarelli et al., 2002). The theoretical and practical utility of both approaches is outlined below.

Reason Giving

Research suggests that Reason Giving requirements may improve referees' felt accountability for providing accurate personality ratings to hiring organizations (Lerner &

Tetlock, 1999). For instance, accountability scholars contend that asking agents to outline their decision-making processes can signal value and acceptance towards their unique perspectives, increasing their perceived impact in the decision task as well as the meaning of its accountability system (Patil et al., 2014). In line with Schlenker and colleagues' pyramid model (Schlenker, 1986; Schlenker et al., 1994), Reason Giving should serve to strengthen referees' feelings of individual responsibility, by more explicitly connecting aspects of their unique identities to the personality references they provide. Requiring referees to report on the more specific details of their task completion (e.g., by asking them to describe the reasons underlying their personality ratings) should also make them feel inherently more prone to scrutiny, thereby leading them to feel more implicitly accountable to the hiring organization for their ratings (e.g., Lerner & Tetlock, 1999). In support of this perspective, Desai & Kouchaki (2015) demonstrated that online survey respondents felt significantly more accountable to the survey administrators when they were asked to record their paid work in individual units as opposed to a broad overall cost, and that this felt accountability in-turn reduced their tendencies to overbill the survey administrators for their efforts (Desai & Kouchaki, 2015, Study 5, Study 6). This effect was then replicated in a field setting, where real auto mechanics were found to provide significantly lower cost estimates for a standardized set of brake repairs when they were asked to do so at the unit-level (Desai & Kouchaki, 2015, Study 7).

The results of Desai and Kouchaki's (2015) multimethod investigation provide support for the notion that implementing Reason Giving requirements could improve referees' felt accountability and personality rating accuracy. Specifically, they suggest that (a) more granular, descriptive reporting formats can foster agents' felt accountability, (b) these formats can lead to more accurate reporting through felt accountability, (c) individuals can feel more or less

accountable as a function of changes to an online reporting tool, and (d) the results obtained under these conditions can generalize to field settings. Requesting that referees describe the reasons underlying their personality trait ratings could therefore be particularly useful for hiring organizations seeking accurate personality ratings, since the above effects were observed in the absence of any *formal* accountability demands, for participants who did not have any interpersonal contacts with their potential constituents (Desai & Kouchaki, 2015). Although scholars have previously suggested that impersonal recruitment processes with minimal dialogues and exchanges may in-part explain why referees' interests are more closely aligned with applicants' (Aamodt et al., 1993; Colarelli et al., 2002; Nash & Carroll, 1970), the available evidence suggests that more granular reporting formats using Reason Giving requirements may serve to attenuate these effects through their capacity to foster referees' felt accountability.

Implementing Reason Giving requirements should also be highly practical, since contemporary reference formats typically contain some written sections that solicit more elaborate and unstructured accounts from referees (Hedricks et al., 2013). In these instances, research has shown that most (i.e., over 80%) referees willingly furnish descriptive, job-relevant narrations towards applicant dispositions (Hedricks et al., 2019), such that Reason Giving requirements for personality ratings may foster felt accountability in a way that is particularly non-invasive (e.g., Morgan et al., 2013; Vance et al., 2015). Moreover, Reason Giving should also improve the accuracy of referees' personality ratings by way of elaboration, which slows down the overall evaluation process and increases its control and complexity compared to the heuristic-based approaches that could otherwise occur in its absence (Fisher & Ford, 1998; Morgan et al., 2013). Although scholars have acknowledged that judgements can be simultaneously complex and inaccurate (e.g., Abelson & Levi, 1985), others suggest that this

style of processing may be a minimum precursor of accuracy in complex tasks such as personality prediction (e.g., Tetlock & Kim, 1987). As such, I predicted that the addition of Reason Giving requirements would increase referees' felt accountability to the hiring organization in a way that should also improve the accuracy of their personality ratings.

H₁: Participants in the Reason Giving conditions will report higher felt accountability to the hiring organization than those in the non-Reason Giving conditions

H₂: Participants in the Reason Giving conditions will provide more accurate personality ratings than those in the non-Reason Giving conditions

H₃: Participants in the Reason Giving conditions will provide more accurate personality ratings through their felt accountability to the hiring organization

Monetary Incentives

Similar to Reason Giving requirements, previous research suggests that the addition of a Monetary Incentive should also foster referees' accountability for providing more accurate personality ratings. As a form of economic exchange, compensation is argued to highlight the consensual elements of employee activities, creating a shared understanding of their accepted responsibilities, and thereby heightening their accountability expectations (Hershey, 1992). As a form of social exchange, compensation is believed to engender trust (i.e., a substitute for accountability; Ammeter et al., 2004) when agents are paid before their *potential* completion of a task, as this signals the mere *expectation* of a positive outcome resulting from their unique contributions (Dillman et al., 2014). In the context of online questionnaires (e.g., contemporary references; Hedricks et al., 2013, 2018) Dillman and colleagues' (2014) social exchange theory suggests that agents should more willingly comply with survey requirements when they believe that doing so will satisfy their internal cost-benefit analyses (Dillman et al., 2014). Stated

differently, this theory implies that hiring organizations issuing a Monetary Incentive to referees could motivate their accountability for accuracy by both (a) conveying trust in their unique abilities to produce ratings that facilitate the goals of the hiring system and (b) increasing the salience of the rewards associated with doing so (Dillman et al., 2014).

Research on social influence (e.g., Cialdini et al., 1981) also suggests that the provision of a Monetary Incentive should foster referees' felt accountability through the norm of reciprocity (Groves et al., 1992). This social norm highlights how individuals are generally motivated to oblige others with actions that are of equal value to those they have received (Cialdini et al., 1975). Stated differently, because referees may otherwise have little to no motivational connections to the hiring organization, the addition of a Monetary Incentive should increase the salience of their exchange relationship and re-weight their motivational attachments to be more in line with the accuracy goals of the hiring organization (Colarelli et al., 2002). In line with Schlenker and colleagues' accountability pyramid (Schlenker, 1986; Schlenker et al., 1994), the increased salience of the exchange relationship should also heighten the extent to which referees feel individually responsible for their ratings, by reinforcing (a) the importance of the accuracy prescriptions set out by hiring organization, (b) the value their ratings hold by way of their unique perspectives, and (c) the utility of their references relative to the hiring organizations' broader selection system. This should in-turn lead referees to feel more accountable given their now more desirable cost-benefit analyses while under the norm of reciprocity.

Previous research has demonstrated the beneficial effects of prepaid incentives for online survey response rates (e.g., Göritz, 2006), although scholars have more recently acknowledged that response rate is an indirect measure of survey quality (Dillman et al., 2014). Particularly

since referees are required respond to reference requirements if they are to enable applicants' chances at being selected, response *quality* should be the primary focus of incentive research on references. Although research on survey incentives has seldom operationalized response quality in terms of accuracy (Singer & Ye, 2013) previous research has highlighted how incentives can elicit favourability bias (e.g., Porter & Whitcomb, 2003), which may increase the overall *quality* (i.e., accuracy) of referee personality ratings in this context.

Favourability

Favourability is a form of biased responding that shifts the response tendencies of participants to be more in line with the goals of the survey administrators, due to the positive reactions that stem from them receiving an incentive (e.g., James & Bolstein, 1990, 1992). Stated differently (in a way that is congruent with the phenomenological perspective of accountability), favourability can be considered as an *acceptability heuristic* whereby agents respond more in line with the values of those to whom they feel accountable to more easily appease their requests (Lerner & Tetlock, 1999; Tetlock et al., 1989). For example, Porter & Whitcomb (2003) found that incentivized respondents provided significantly more ratings towards the characteristics of the universities sponsoring their surveys, despite responding similarly to the non-incentivized participants throughout the rest of their questionnaires. Thus, although the available evidence for incentives improving response accuracy may be scant (Singer & Ye, 2013), there are both theoretical and empirical advances to suggest that their provision should foster referees' felt accountability in a way that should also motivate them to shift their response patterns towards accuracy in this context. Taken together, I predicted that the provision of a Monetary Incentive would improve referees' felt accountability to the hiring organization in ways that would also increase the accuracy of their personality ratings.

H4: Participants in the Monetary Incentive conditions will report higher felt accountability to the hiring organization than those in the non-Monetary Incentive conditions

H5: Participants in the Monetary Incentive conditions will provide more accurate personality ratings than those in the non-Monetary Incentive conditions

H6: Participants in the Monetary Incentive conditions will provide more accurate personality ratings through their felt accountability to the hiring organization

Method

Participants

A total of 259 survey responses were initially collected, yet several were excluded from the reported statistical analyses. Specifically, 14 participants were excluded for failing an audio-check question that verified their abilities to play the work-sample videos. Sixteen participants were removed for failing one or more attention check items dispersed throughout the questionnaire. Two participants were removed for not identifying English as their first language, and 16 participants were excluded because they did not provide any personality ratings for one or more target employees. In total, the reported analyses were conducted on data from 211 participants.

211 participants were recruited through Prolific. Pre-screen questions were applied to ensure that the sample contained only supervisors or managerial personnel working a minimum of 21 hours per week. This was done to increase the generalizability of this study to actual reference rating scenarios, where ratings are most commonly provided by applicants' superiors (Hedricks et al., 2019). Participants were middle aged ($M = 41.33$, $SD = 10.04$) supervisors or managers that identified as either male ($N = 102$) or female ($N = 109$). An other category for gender was provided, yet all participants reported identifying as cisgender. Participants were

primarily Caucasian (92%), English speaking (100%) residents of the United Kingdom. They had been working for over 20 years on average ($M = 21.17$, $SD = 9.58$), and had spent close to half of their careers working in a supervisory role ($M = 9.42$, $SD = 7.07$). Participants reported having an average longest tenure of 11 years ($M = 11.00$, $SD = 6.35$), and providing approximately 13 references to hiring organizations in the past ($M = 13.17$, $SD = 14.93$).

Procedure

This study was conducted entirely online through Prolific. Participants were invited to participate in a study examining the usefulness of references in call centers for a total compensation of £4. Participants began the study by providing their informed consent and by completing a brief demographics questionnaire. They were then presented with a cover story (see Appendix B) describing how a new call center recently contacted an ostensible “Human Resources Lab” with concerns regarding the utility of supervisory references for their upcoming hiring system. Participants were then led to believe that this call center had contracted the Human Resources lab to provide evidence-based suggestions towards how they could best use supervisory references when hiring. More specifically, participants were told that by asking them to provide references for a set of prototypical call center employees, the Human Resources lab could better understand how supervisors make their personality references in call centers in order to help determine their overall utility for hiring. They were then randomly assigned to one of four experimental conditions, in which they were asked to assume the role of a referee for four call center employees shown in a set of work-sample videos. After watching each work-sample video, participants were asked to complete a personality rating form for each target as if they were their actual supervisors providing the Human Resources Lab with real reference evaluations

for these employees. Additional measures were also included in the Qualtrics questionnaire that were not the focus of the current study.

Participants were asked to rate a common set of personality traits across targets, but their rating scenarios were each randomly manipulated using a 2x2 factorial design. These manipulations were conducted so that all possible conditions either contained or did not contain (a) Reason Giving requirements accompanied by textboxes that were presented under each quantitative rating scale (see Appendix C), or (b) prompts at the start of the rating task outlining that participants would be issued a £6 Monetary Incentive as a token of appreciation for their efforts (see Appendix D). This design is depicted in Figure 1.

		<u>Reason Giving:</u>	
		Yes	No
<u>Monetary Incentive:</u>	Yes	$N = 52$	$N = 52$
	No	$N = 51$	$N = 56$

Figure 1. *Experimental Design.*

Materials

Work-sample videos. A set of four work-sample videos was sourced from three transnational training firms that condoned their inclusion in the current study. An initial and comprehensive web search led the student researcher to collect a sample of eight useable work-sample videos, of which were later refined with the principal investigator to yield a more homogeneous subset congruent to the study goals. After selecting four usable work-samples, each individual publisher was contacted to obtain written permission for their use.

This set of four work-sample videos were presented in a random order to each study participant. These videos were approximately 4 minutes in length on average and depicted unique male call center employees handling typical customer service calls at their desks. The scope of each call varied, but the entire set of videos showcased a set of common customer service inquiries that are typically handled through call centers (i.e., account inquiries, placing orders for new products and services, and handling post-sales troubleshooting issues).

Past research suggests that two to three minutes of exposure to an individual is sufficient for accurate trait judgments (Ambady & Rosenthal, 1992; Sheppard et al., 2011; Shmidt Mast et al., 2011) and that trait observations from performance-relevant domains yield similar validity to those from longer acquaintanceships (Colvin & Funder, 1991). Thus, the chosen set of work-sample videos was believed to have provided an adequate basis for participants to form their personality references. Personality has also been shown to be an important predictor of job performance in call center employees (e.g., Konradt et al., 2003), which is a job sector that is continuously expanding in the global economy (Aksin et al., 2007; Bradshaw et al., 1999). Thus, the use of personality references in call centers is both ecologically valid and generalizable to field settings. Finally, since modern-day call centers typically record their customer interactions

for training, development, and quality assurance purposes, and because call center employees primarily work unsupervised, it is more likely that actual supervisors in these organizations may draw on similar work-sample data to inform their references. The chosen video set therefore enabled this research to explore the effects of Reason Giving requirements and the Monetary Incentive in a highly practical context through an ecologically valid approach.

Employee personality. This study used an adapted version of a personality rating form developed by Powell and Goffin (2009) using the trait definitions from the Big Five model of personality (NEO-FFI; Costa & McCrae, 1992). This model contains five global personality traits (i.e., Emotional Stability, Openness to Experience, Agreeableness, Conscientiousness and Extraversion; Costa & McCrae, 1992), of which four were included in the current study. Openness to Experience was excluded due to its general lack of criterion validity (Barrick et al., 2001) and studies suggesting greater difficulties towards accurately perceiving this trait (e.g., Powell & Bourdage, 2016). The resulting rating form (see Appendix E) contained four questions that each defined one of the traits in the model (e.g., *Agreeableness*) at both high (e.g., *Trusting of others; helpful; cooperative; accepts criticism and blame; avoids confrontations and conflicts, humble*) and low (e.g., *Sceptical of others' intentions; enjoys arguments; stubborn; easily annoyed; takes offence easily; avoids apologizing*) levels on its respective continuum. The rating form asked participants to rate each employees' personality using 7-point Likert-type Agree-Disagree scales while referring to these descriptions.

True score ratings. True score estimates for the personality ratings were determined using Borman's (1977) approach. Specifically, 9 subject matter experts (SMEs) (i.e., one professor and eight graduate students with expertise in personality and industrial / organizational psychology) were asked to watch each of the four work-sample videos and make ratings towards

the employees' broad-level personality traits using the above rating form. These SMEs had infinite access to the work-sample videos in order to better inform their true score ratings, and they were encouraged to take all of the necessary steps to arrive at their most accurate ratings of the employees (e.g., pausing, rewinding, fast-forwarding, and replaying the videos; Borman, 1977). Averages of these trait ratings then served as an estimate of the true personality profiles of the targets, which served as a criterion for rating accuracy across the experimental conditions.

Interrater reliability for the expert judges was calculated using both ICC₍₂₎ and the Spearman-Brown formula. ICC₍₂₎ reliability was shown to be .97. Using the Spearman-Brown formula, the mean interrater correlation was first computed then stepped up by 9 to yield an overall estimate of the Alpha reliability of the raters (e.g., Jelley & Goffin, 2001; Powell & Goffin, 2009). The estimated alpha interrater reliability was .99, indicating that the expert judges showed high levels of agreement across the four targets in this study.

Accuracy. Cronbach's (1955) accuracy components are a set of deviation scores that are useful for judging rater accuracy when multiple targets are rated on at least two dimensions by several raters. Figure 1 depicts the formulas for each of the four accuracy components described below. Because these statistics are deviation scores, lower values indicate greater accuracy. Accuracy values in the current study ranged from .01 to 1.65.

Elevation (EL) represents the differential grand mean. Inaccuracy on this component depicts raters' tendency to rate either too high or too low on the rating scale, averaged across all rates and items. This component of accuracy has important implications for hiring organizations, as it depicts referees' tendencies to be overly severe or lenient as a rater.

Differential Elevation (DE) captures the differential main effect of the targets. Accuracy in this component represents the rater's ability to correctly rank order the targets, averaged

across all items while controlling for their elevation. This component of accuracy has implications for the administrative usage of references, as it demonstrates referees' accuracy in discriminating between applicants based on their aggregated personality ratings.

Stereotype Accuracy (SA) reflects the differential main effect of the items. Not to be confused with the stereotype accuracy commonly investigated in social psychology (i.e., the extent to which a group member is accurately perceived as conforming to some stereotypical attribute of their ingroup; Ryan, 2013), accuracy in this component outlines the rater's accuracy towards distinguishing the average individual personality trait levels in a group of targets. This component of accuracy has implications for hiring organizations as it highlights referees' capacity to determine if a given group of ratees are higher on one job-relevant trait compared to others, which could inform how they tailor their references.

Differential Accuracy (DA) describes the target x item interactions that occur within targets. Accuracy in this component demonstrates raters' abilities to identify the unique trait profiles of individual targets, which most closely captures the concept of interpersonal sensitivity (Borman, 1977). This component of rating accuracy encompasses the variance not explained by a rater's overall rating tendency (EL), or by the effects of the individual ratees (DE) or items on the rating form (SA). This component of accuracy has critical implications in contexts where decisions may be made based on the ratings provided (e.g., in preemployment assessment; Murphy & Cleveland, 1995).

$$\begin{aligned}
 EL &= \sqrt{(\bar{x}_G - \bar{t}_G)^2} \\
 DE &= \sqrt{\frac{1}{n} \sum_i [(\bar{x}_i - \bar{x}_G) - (\bar{t}_i - \bar{t}_G)]^2} \\
 SA &= \sqrt{\frac{1}{k} \sum_j [(\bar{x}_j - \bar{x}_G) - (\bar{t}_j - \bar{t}_G)]^2} \\
 DA &= \sqrt{\frac{1}{kn} \sum_i \sum_j [(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}_G) - (t_{ij} - \bar{t}_i - \bar{t}_j + \bar{t}_G)]^2}
 \end{aligned}$$

n: number of rates.

k: number of items.

\bar{x}_G : rater grand mean (across items and rates).

\bar{t}_G : true score grand mean (across items and rates).

\bar{x}_i : rater mean for the ratee i (across items).

\bar{t}_i : true score mean for ratee i (across items).

\bar{x}_j : rater mean for item j (across rates).

\bar{t}_j : true score mean for item j (across rates).

x_{ij} : rating for ratee i on item j.

t_{ij} : true score for ratee i on item j.

Figure 2. *Cronbach's (1955) accuracy components*

Felt accountability. Participants' felt accountability was measured using an adapted 5-item version of Hochwarter and colleagues (2007) felt accountability scale (Appendix F). This scale contained adaptations similar to those of Desai and Kouchaki's (2015) 3-item measure that was specifically designed for MTurk use (e.g., *The researchers paying me for this task hold me accountable for my descriptions*), and asked participants to respond to each statement using a 7-point Likert type Agree-Disagree scale. Previous research has shown good reliability for both the 3-item ($\alpha = .74$, Desai & Kouchaki, 2015) and 8-item versions of this scale ($\alpha = .74$; Hochwarter et al., 2007), and that the 8-item version correlates meaningfully with relevant workplace outcomes (e.g., organizational citizenship behaviours; Hall & Ferris, 2011).

Affective reactions. Participants' confidence in their ratings (e.g., *I feel confident about the accuracy of my descriptions*) as well as their perceived value (e.g., *I feel that I provided valuable insights for this research*) were each assessed using single-item measures on a 5-point Likert-type Agree-Disagree scale (Appendix G). These variables were included to provide further insights on the effects of felt accountability in the reference rating context (e.g., Tetlock & Kim, 1987; Patil et al., 2014).

Effort. In line with Funder's (1995) Realistic Accuracy Model of personality perception, participants' efforts while both watching (e.g., *I put effort into watching the videos - I tried to identify and understand the employees' characteristics as best as I could*) and rating the targets (e.g., *I put effort into my descriptions – I tried to provide careful and accurate descriptions of the employees to the best of my abilities*) were each assessed using single item measures on a 5-point Likert-type Agree-Disagree scale (Appendix H).

Demographics. Participants were asked to complete a demographics questionnaire at the beginning of the study (see Appendix I). This questionnaire asked participants questions related

to their age, gender, ethnicity, and primary language, and also asked them to report on their past educational, job, supervisory and reference-related experiences.

Results

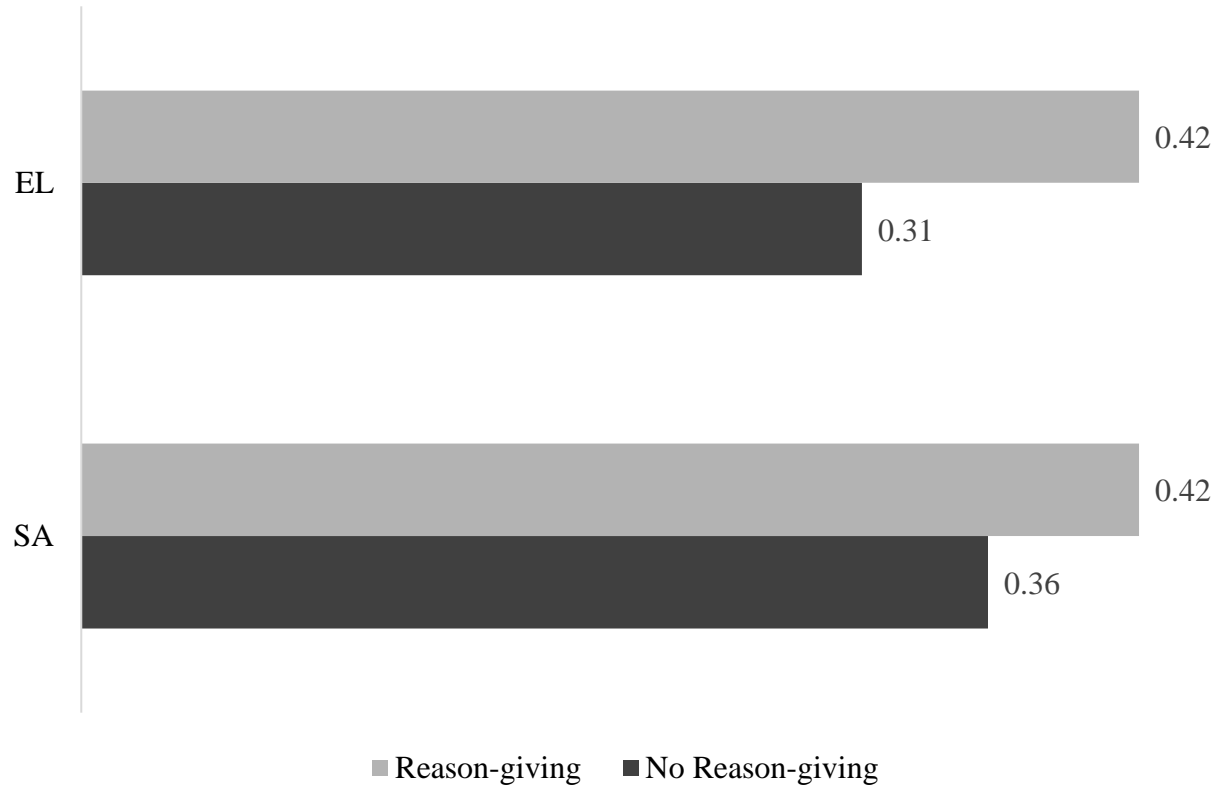
Personality Rating Accuracy

Correlations for the study variables of interest are reported in Table 1. The multivariate results for personality rating accuracy are presented in Table 2. The current study predicted that both Reason Giving requirements (H_2) and the Monetary Incentive (H_5) would improve the accuracy of referees' personality ratings. To test these hypotheses, a completely randomized multivariate analysis of variance (MANOVA) with two independent variables (Reason Giving requirements and the Monetary Incentive) was conducted with the four accuracy components (EL, DE, SA & DA) as dependent variables. The multivariate main effect of Reason Giving was significant, Pillai's trace = .075, $F(4,189) = 3.847$, $p = .005$, whereas the multivariate main effect of the Monetary Incentive was not significant, Pillai's trace = .007, $F(4,189) = .343$, $p = .849$.

To examine how the two independent variables affected each of the four accuracy components individually, separate univariate analyses of variance (ANOVAs) were executed with EL, DE, SA, and DA as the dependent variables. This practice was justified given the nature of Cronbach's (1955) accuracy components as distinct, yet somewhat interrelated elements of raters' overall accuracy (e.g., Jelley & Goffin, 2001).

Reason Giving. The univariate results for Reason Giving are presented in Table 3. Results for the main effect of Reason Giving revealed a significant effect on EL [$F(3, 207) = 7.027$, $p = .006$, $\eta^2 = .033$], such that participants who received the form containing Reason Giving requirements were significantly less accurate in terms of rating too low across all targets

and items ($M = .419, SD = .309$) compared to those who did not ($M = .313, SD = .273$). The Reason Giving requirements also had a significant effect on SA [$F(3, 205) = 5.320, p = .022, n^2 = .025$], such that those in the Reason Giving conditions were significantly less accurate in terms of rating the average level of each trait across the targets ($M = .418, SD = .206$) compared to those who were not ($M = .357, SD = .172$). Graphical representations of these effects are shown in Figure 2. Both results would be considered as small to medium effects as per Cohen's (1988) benchmarks. No significant effects were observed for DE or DA.



Rating accuracy on a 7-point Likert-type scale. Lower values indicate greater accuracy

Figure 3. *Significant main effects of Reason Giving.*

Monetary Incentive. The univariate results for the Monetary Incentive are presented in Table 4. The Monetary Incentive did not have any significant effects towards any of Cronbach's (1955) accuracy components. Thus, participants who received the form containing notices about receiving an additional £6 payment did not make significantly more accurate personality ratings than those who did not, across all possible combinations of the independent and dependent variables.

Felt Accountability

The results concerning felt accountability are reported in Table 5. The current study predicted that both Reason Giving requirements (H_1) and the Monetary Incentive (H_4) would improve referees' felt accountability to the hiring organization. However, the results of a 2x2 Analysis of Variance (ANOVA) indicated that participants felt accountability was not significantly different across the four experimental conditions $F(3, 207) = 1.166, p = .324, \eta^2 = .017$. This result provides no support for Hypotheses 1 and 4 and suggests that neither Reason Giving ($\eta^2 = .000$) nor the Monetary Incentive ($\eta^2 = .118$) or their interaction ($\eta^2 = .005$) significantly influenced participants' felt accountability to the hiring organization.

Mediation model

The results of the mediation analyses are presented in Table 6. The current study predicted that both the Reason Giving requirements (H_3) and the Monetary Incentive (H_6) would improve the accuracy of referees' personality ratings through their felt accountability to the hiring organization. This model is depicted in Figure 3. These analyses were conducted using bootstrap-based mediation with 5000 iterations in PROCESS for SPSS (Preacher & Hayes, 2004), testing the model four times for each independent variable using each of the accuracy components in-turn as the dependent variable. Across all eight models, none of the indirect

effects of felt accountability on personality rating accuracy were significant, providing no support for Hypotheses 3 and 6. This suggests that the accuracy-related effects of Reason Giving requirements and the Monetary Incentive did not occur through referees' felt accountability to the hiring organization in the current study.

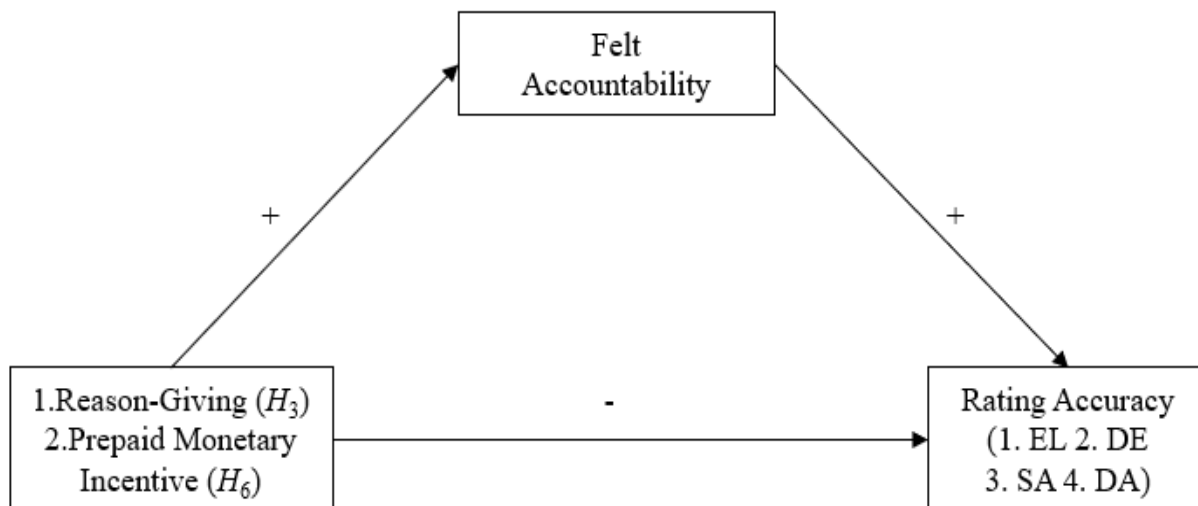


Figure 4. *Hypothesized mediation models.*

Discussion

References are a ubiquitous preemployment assessment (SHRM, 2005) that have the capacity to produce incremental personality information that can improve hiring decisions (Aamodt, 1999; McCarthy & Goffin, 2001; Hedricks et al., 2013; Hedricks et al., 2018). However, referees who might otherwise be capable of providing both valid and reliable personality assessments (e.g., Connelly & Ones, 2010) tend to provide relatively inaccurate references (Aamodt & Williams, 2005; Furnham, 2017; Hunter & Hunter, 1984; Reilly & Chao, 1982). The current study was conducted to examine two new ways to improve the accuracy of referee personality ratings, by adding both Reason Giving requirements and a Monetary Incentive to the reference rating situation.

Drawing on accountability theory (Schlenker, 1986; Schlenker et al., 1994; Tetlock, 1985, 1992), the current study argued that referees could be made to feel more implicitly accountable to the goals of hiring organizations by reinforcing their individual responsibility in the rating situation. It was then argued that this felt accountability would improve the accuracy of their personality ratings, since referees might otherwise have little reason to be motivated towards the accuracy goals of hiring organizations in this context (Colarelli et al., 2002; Knouse, 1983, 1987; Miller & Rybroeck, 1988).

By signalling value and acceptance towards their unique perspectives and by reinforcing scrutiny (Lerner & Tetlock, 1999; Patil et al., 2014), the current study predicted that Reason Giving requirements would foster referees' felt accountability to the hiring organization. By leading them to make more integratively complex and controlled evaluations of job applicants (Fisher & Ford, 1988; Mero & Motowidlo, 1995; Morgan et al., 2013; Tetlock & Kim, 1987), the current study predicted that Reason Giving requirements would improve referees' personality

rating accuracy. This line of reasoning converged with studies showing that participants accountable for Reason Giving made more accurate interpersonal judgements across other selection and performance evaluation contexts (Brtek & Motowidlo, 2002; Mero & Motowidlo, 1995; Mero et al., 2007; Morgan et al., 2013; Tetlock & Kim, 1987).

By engendering trust (Dillman et al., 2014) and reciprocity (Cialdini et al., 1975; Groves et al., 1992), the current study predicted that the Monetary Incentive would foster referees' felt accountability as a form of social exchange. By shifting their response tendencies to be more in line with the goals of the hiring organization (Colarelli et al., 2002; James & Bolstein, 1990, 1992), the current study predicted that the Monetary Incentive would improve the accuracy of referees' personality ratings. This line of reasoning converged with studies demonstrating the positive effects of incentives for motivating survey responding (Görizt et al., 2006; Jobber et al., 2014; Singer & Ye, 2013) as well as those highlighting the capacity for incentives to elicit favorable response tendencies towards survey administrators (e.g., Porter & Whitcomb, 2003). Nevertheless, the current study revealed that neither the Reason Giving requirements nor the Monetary Incentive tended to influence referees' felt accountability, whereas the Reason Giving requirements had unexpected negative consequences for the accuracy of referee personality ratings. This was indeed surprising in light of the above rationale, but nonetheless provides useful contributions to the existing literature on references.

Reason Giving effects

Perhaps most notably, the current study revealed that adding Reason Giving requirements to an online personality reference form did not improve participants' felt accountability and resulted in unintended negative consequences for their personality rating accuracy. This finding is surprising, as it contradicts previous theoretical advances suggesting that Reason Giving

should (a) signal value and acceptance towards agents' unique perspectives and heighten the meaning of their accountability system (Patil et al., 2014), (b) make them inherently more prone to scrutiny (Lerner & Tetlock, 1999), and (c) make them feel more accountable for their responses (Morgan et al., 2013). Further, these findings oppose the empirical evidence suggesting that Reason Giving requirements can improve rating accuracy, from studies demonstrating their capacity to reduce bias in evaluations of job applicants (Morgan et al., 2013), improve personality prediction (Tetlock & Kim, 1987), and increase the accuracy of interview (Brtek & Motowidlo, 2002) and performance ratings (Mero & Motowidlo, 1995; Mero et al., 2007).

Although the results observed in this study do not align with those described above, they may provide support for the potential moderating effect of the communication method used for reason giving. For example, much of the aforementioned research on Reason Giving requirements was conducted in laboratory settings, whereas this study took place entirely online. To this end, Mero and colleagues (2007) acknowledged that raters' identifiability can vary saliently across different methods of communication (Schlenker, 1986), such that reason giving should more strongly amplify rating processes in situations where identifiability is high (e.g., when Reason Giving face-to-face) compared to when it is lower (e.g., when Reason Giving in writing; Klimoski & Inks, 1990; Mero et al., 2007; or to an external party; Tetlock et al., 1989). Although there was good reason to believe that Reason Giving requirements had the capacity to elicit greater *felt* accountability in referees, in a way could also lead them to report more accurate personality ratings (e.g., Desai & Kouchaki, 2015; Tetlock & Kim, 1987), it could be that the distal nature of online reference tools removes the identifiability necessary for Reason Giving requirements to significantly improve rating accuracy. Stated differently, it may be that Reason

Giving requirements have the capacity to improve rating accuracy, but potentially only in situations where agents feel that they can be more *explicitly* held accountable for their ratings by way of their identifiability (e.g., in internal performance appraisal settings; Mero et al., 2007). Otherwise, in situations where raters are more far-removed from those to whom they may feel *implicitly* accountable, the current study suggests that Reason Giving requirements may serve to paradoxically reduce personality rating accuracy. As explained next, the accountability literature (e.g., Tetlock & Boettger, 1989) suggests that this finding may be attributable to a dilution effect.

Although much of the accountability literature converges on the positive effects of holding agents accountable through Reason Giving requirements (Lerner & Tetlock, 1999; Patil et al., 2014), other studies have found that they can reduce judgement accuracy by way of dilution (e.g., Nisbett et al., 1981). Specifically, research suggests that when agents must outline the reasons underlying their decisions, they tend to engage in more complex integrations of the available information, becoming less attuned to its overall relevance and ultimately incorporating more invalid cues into their judgements (e.g., Siegel-Jacobs & Yates, 1996, Study 1; Tetlock & Boettger, 1989; Tetlock et al., 1996). For example, Tetlock and Boettger (1989) demonstrated that participants held accountable for Reason Giving made more moderate and integratively complex predictions towards the GPAs of college students when presented with both diagnostic (e.g., the number of hours they spent studying per week) and non-diagnostic information (e.g., that they played tennis once a week), but made more extreme judgements towards these same targets when only diagnostic information was presented. Similarly, Siegel-Jacobs and Yates (1996) found that judges accountable for Reason Giving were not any more discriminative than those in confidential, non-disclosure conditions as they were shown to devote more attention to both the valid and non-valid cues presented towards the focal attitude of interest.

Although accountability scholars have acknowledged the need for some degree of integrative complexity in tasks such as personality prediction (Tetlock & Kim, 1987) it may be that Reason Giving requirements lead referees to draw on judgement strategies that are overly complex for rating applicant personality, perhaps by incorporating more trait-*irrelevant* information into their ratings. Indeed, Reason Giving has been theorized to create a “socially safe” space for decision-makers so long as their procedures are properly outlined and defensible (Patil et al., 2014), which may have reduced referees’ apprehension towards uncertainty when determining the trait-relevance of certain cues and/or behaviours (e.g., Siegel-Jacobs & Yates, 1996). To these ends, the Reason Giving referees in the current study (who reported feeling accountable to the hiring organization; $M = 5.98$, $SD = .80$) spent over twice as long evaluating each of the target employees ($M = 10:24$ versus $4:18$ in the non-Reason Giving conditions), despite making significantly less accurate personality ratings.

Monetary Incentive effects

Second, the current study revealed that the addition of the Monetary Incentive did not significantly impact referees’ felt accountability to the hiring organization, nor did it affect the accuracy of their personality ratings. These findings oppose the literature on survey incentives and social influence, which suggest that Monetary Incentives should engender feelings of trust (Dillman et al., 2014) and reciprocity (Cialdini, et al., 1975; Groves et al., 1992) in respondents, that lead them to shift their response tendencies to be more in line with the goals of the survey administrators (e.g., favourability; James & Bolstein, 1990 1992; Porter & Whitcomb, 2003) by way of their exchange relationships (Colarelli et al., 2002).

Although much meta-analytic evidence suggests that the provision of monetary incentives can improve responding in online and mail-based surveys (e.g., Göritz et al., 2006;

Jobber et al., 2004), the majority of this literature is rooted in questionnaires that require markedly less expertise from respondents (e.g., undergraduate non-admission surveys; Porter & Whitcomb, 2003) completing tasks that are arguably not as cognitively demanding as forming personality references (e.g., online consumer surveys; Enander & Sajti, 1999). Thus, even though individuals tend to be generally familiar with the personality perception process because of its importance for everyday life (Funder, 1995, 2012), it may be that the favourable response tendencies evoked by incentives in more typical survey contexts do not generalize to referees completing personality references. To this end, Funder's (1995) Realistic Accuracy Model of personality perception suggests that any variable influencing personality judgement accuracy must influence raters' *detection* (i.e., devoting sufficient attention to observing the cues emitted by a target) and/or *utilization* (i.e., acknowledging that the observed information is related to a given personality dimension and ascribing it as such). However, participants in the current study reported near-identical efforts when both watching ($M = 4.98$ vs. 4.97) and rating ($M = 4.97$ vs. 4.95) the target employees, regardless of if they received the Monetary Incentive.

Opposingly, participants receiving the Monetary Incentive reported feeling accountable to the hiring organization ($M = 6.06$, $SD = .77$), expressed confidence in their ratings ($M = 4.79$, $SD = .41$) and believed that they brought value to the study goals ($M = 4.71$, $SD = .52$). They also each received a sizeable incentive that increased their overall compensation for participation by 250%. Thus, it could be that the inaccuracies observed in referee personality ratings are less a function of their motivation as was initially hypothesized, and that they may be more strongly explained by other individual difference variables (e.g., Dispositional Intelligence; Christansen et al., 2005). Particularly since the Monetary Incentive was issued in a way that was unbeknownst to participants prior to arriving at the rating task, to elicit a novelty effect (e.g., Rose et al., 2007)

similar to what actual referees may experience in field settings, this line of reasoning can be stated with greater confidence.

Limitations

While the current research provides both practical and theoretical insights towards the effects of Reason Giving requirements and the Monetary Incentive on referees' felt accountability and their personality rating accuracy, it does not do so without some notable limitations. First, the current study controls for the presence of relational biases such as leniency (i.e., the tendency for raters to systematically inflate the scores they assign to a target stimulus; Hoyt, 2000), which have otherwise been shown to plague references in field settings (e.g., Aamodt et al., 1993; Aamodt & Williams, 2005). By removing the potential for referees to have had previous working relationships with the targets (Colarelli et al., 2002) and by controlling for the possibility that they could experience ramifications from those receiving negative reviews (e.g., Bernardin & Buckley, 1981), the current study likely reduced the extent to which participants were motivated to convey the targets in an overly positive light. This therefore limits the generalizability of the observed effects of Reason Giving requirements and the Monetary Incentive to field settings. Although the current study does mirror the frequent situation where referees are asked to provide ratings to an external party (Knouse, 1983, 1987; Miller & Rybroeck, 1988) for applicants that they may be relatively unfamiliar with (Aamodt et al., 1993; Blum & Naylor, 1968; Daniel, 1990; Knouse, 1994), it is unable to determine how the accuracy-related effects of Reason Giving requirements and the Monetary Incentive fare against other potential competing motivations for personality rating *inaccuracy* in references.

Second, as with most fleeting laboratory designs, the current study does not capture the multifaceted nature of felt accountability found in field settings. For instance, the notion of

embeddedness, which highlights how employees in organizations are each uniquely embedded into broader webs of accountability demands (e.g., Frink & Klimoski, 1998, 2004) cannot be captured by the current study since participants never had the opportunity to interact with the targets or the researchers. Although the Reason Giving requirements and Monetary Incentive were implemented to foster participants' *felt* accountability in a way that was both practical and non-invasive, it is likely that referees in field settings also experience other competing accountability demands when providing their personality references to hiring organizations (e.g., whether or not they were chosen by the applicant under the assumption that they would provide positive ratings; Aamodt et al., 1993). As such, although this study provides potentially useful evidence towards the effects of Reason Giving requirements and the Monetary Incentive on referees' felt accountability to the hiring organization, it does not do so in a way that can delineate these effects from those that they might also have on referees' felt accountability towards job applicants in the field.

Finally, there is the potential that some of the factors held constant across the experimental conditions may be considered limitations in the current study. For example, every study participant received a smaller initial compensation for their participation, and each read a carefully crafted cover story at the beginning of the rating task that was designed to solicit purposeful responding. Nevertheless, any empirical questions towards how these factors may have influenced the observed results cannot be examined directly due to the current study design.

Future Directions

Given the results of this study, one useful direction for future research would be to investigate the effects of other accountability manipulations that could more saliently reinforce referees' identifiability in the rating situation (e.g., Lerner & Tetlock, 1999). For instance, Vance

and colleagues (2015) found that displaying the name, photograph, and user ID of the person accessing a company information system both (a) improved employees' felt accountability for their actions while logged in and (b) reduced their intentions to violate its access policies. In the context of self-report personality testing, Dwight and Donovan (2003) also found that warnings of identifiability reduced test-takers' mean scores and the presence of faking behaviours. To these ends, it may be useful for future studies to explore the effects of other potential identifying and/or individuating mechanisms that could be seamlessly implemented into the reference rating situation (e.g., using more personalized cover letters and/or page headings) to reduce the potentially de-identifying nature of contemporary reference formats. Changes of this nature should serve to amplify referees' rating processes (i.e., Mero et al., 2007), which could improve the accuracy of their personality ratings.

A second direction for future research would be to revisit Funder's (1995) Realistic Accuracy Model to explore other interventions that could potentially improve referees' detection and/or utilization of job applicants' personality information. For example, although the Reason Giving requirements in this study were shown to reduce the accuracy of referee personality ratings, it could be that other rating formats elicit more channelled efforts in referees' detection and/or utilization that might similarly improve the accuracy of their personality ratings. One such approach might be the use of structured free recall, which requires raters to outline both positive and negative behavioural examples relevant to a specific dimension prior to rating it (e.g., Baltes et al., 2007). This approach has been shown to decrease the influence of gender stereotypes on performance rating accuracy in past research (Bauer & Baltes, 2002), so it may serve as a promising avenue for improving the accuracy of referees' personality ratings in line with Funder's (1995) model. Moreover, because studies have shown that most referees (i.e., over

80%) willingly speak towards applicants' strengths and weaknesses in writing (Hedricks et al., 2018), it could also be a practical way to improve referees' personality rating accuracy if found successful.

Conclusion

The goal of this study was to examine the effects of two novel interventions to an online reference form, in attempts to increase referees' felt accountability to the hiring organization and improve the accuracy of their personality ratings. The results presented contribute to the literature on references in a number of useful ways. First, they highlight a novel, and perhaps paradoxical avenue for understanding rating *inaccuracies* in references, such that open-ended Reason Giving requirements may be a practical, yet counterproductive response format for the accuracy of referee personality ratings. Second, this study showcases that referees may not be motivated by receiving monetary incentives from hiring organizations for their services, and that attempts to foster favourable responding through trust and reciprocity may not be a useful pathway for improving their personality rating accuracy in this context.

From a practical standpoint, this research contributes to the larger body of studies highlighting how structural and situational factors can influence the accuracy of references. Additionally, it answers the longstanding call to investigate the effects of monetary incentives in references. Theoretically, this study contributes to the growing literature on accountability by examining a novel source of felt accountability in a novel preemployment assessment context. It also adds to our understanding of the theoretical linkages between felt accountability and personality rating accuracy. Ultimately, it is hoped that this study can stimulate further research on references, which remain a widely used, yet relatively understudied preemployment assessment.

References

- Aamodt, M. G. (1999). *Applied industrial organizational psychology* (3rd ed.). London: Brooks/Cole.
- Aamodt, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting performance with letters of recommendation. *Public Personnel Management*, 22, 81–90.
- Aamodt, M. G., & Williams, F. (2005). *Reliability, validity, and adverse impact of references and letters of recommendation*. Paper presented at the 20th annual meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.
- Abelson, R. P., & Levi, A. (1985). Decision making and decision theory In Lindzey G and Aronson E, (Eds.): *Handbook of Social Psychology* (Vol. 1, pp. 231–209).
- Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6), 665-688.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2), 256-274.
- Ammeter, A. P., Douglas, C., Ferris, G. R., & Goka, H. (2004). A social relationship conceptualization of trust and accountability in organizations. *Human Resource Management Review*, 14(1), 47-65.
- Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86(4), 736-765.
- Baltes, B. B., Bauer, C. B., & Frensch, P. A. (2007). Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism?. *Journal of Applied Psychology*, 92(1), 151-164.

- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and Assessment*, 9(1-2), 9-30.
- Baron, R. A. (1993). Criticism (informal negative feedback) as a source of perceived unfairness in organizations: Effects, mechanisms, and countermeasures. In R. Cropanzano (Ed.), *Justice in the workplace: Approaching fairness in human resource management* (pp. 155–170). Lawrence Erlbaum Associates, Inc.
- Bauer, C. C., & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles*, 47(9), 465-476.
- Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology*, 66(3), 296-301.
- Beason, G. M., GM, B., & JA, B. (1976). VERIFYING APPLICANT'S BACKGROUNDS. *Personnel Journal*, 55, 345-348.
- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205-212.
- Blum, M. L., Naylor, J. C., & Naylor, J. C. (1968). *Industrial psychology; Its theoretical and social foundations*. Harper & Row.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20(2), 238-252.
- Bradshaw, D., Wood, S., & Delaney, J. (1999). *Next generation call centres: Cti voice and the web*. Ovum.
- Browning, R. C. (1968). Validity of reference ratings from previous employers. *Personnel Psychology*, 21(3), 389–393.

- Brtek, M. D., & Motowidlo, S. J. (2002). Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology, 87*(1), 185-191.
- Carroll, S. J., & Nash, A. N. (1972). Effectiveness of a forced-choice reference check. *Personnel Administration, 35*(2), 42-46.
- Ceci, S. J., & Peters, D. (1984). Letters of reference: A naturalistic study of the effects of confidentiality. *American Psychologist, 39*(1), 29-31.
- Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance, 18*(2), 123-149.
- Cialdini, R. B., Petty, R. E., & Cacioppo, J. T. (1981). Attitude and attitude change. *Annual Review of Psychology, 32*(1), 357-404.
- Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D., & Darby, B. L. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology, 31*(2), 206-215.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Colarelli, S. M., Hechanova-Alampay, R., & Canali, K. G. (2002). Letters of recommendation: An evolutionary psychological perspective. *Human Relations, 55*(3), 315-344.
- Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology, 60*(6), 884-894.
- Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092-1122.

- Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*(1), 5-13.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity.". *Psychological Bulletin, 52*(3), 177-193.
- Daniel, D. (1990). Validity of a standardized reference checklist. *Applied HRM Research, 1*(2), 51-66.
- Desai, S. D., & Kouchaki, M. (2015). Work-report formats and overbilling: how unit-reporting vs. cost-reporting increases accountability and decreases overbilling. *Organizational Behavior and Human Decision Processes, 130*, 79-88.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81-106.
- Enander, J., & Sajti, A. (1999, February). Online survey of online customers. Value-added market research through data collection on the Internet. In *Proceedings of the ESOMAR Worldwide Internet Conference Net Effects* (Vol. 2, pp. 35-51).
- Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology, 51*(2), 397-420.
- Frink, D. D., & Klimoski, R. J. (1998). Toward a theory of accountability in organizations and human resource management. In G. R. Ferris (Ed.), *Research in personnel and human resources management*, (Vol. 16, pp. 1-51). Elsevier Science/JAI Press.

- Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review, 14*(1), 1-17.
- Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychological Review, 102*(4), 652-670.
- Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177-182.
- Furnham, A. (2017). The Contribution of Others' Methods in Recruitment and Selection. *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention, 202*.
- Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 48*(2), 251-268.
- Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science, 1*(1), 58-70.
- Green, M. C., Visser, P. S., & Tetlock, P. E. (2000). Coping with accountability cross-pressures: Low-effort evasive tactics and high-effort quests for complex compromises. *Personality and Social Psychology Bulletin, 26*(11), 1380-1391.
- 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*(6), 655.

- Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56(4), 475-495.
- Hall, A. T., Bowen, M. G., Ferris, G. R., Royle, M. T., & Fitzgibbons, D. E. (2007). The accountability lens: A new way to view management issues. *Business Horizons*, 50(5), 405-413.
- Hall, A. T., & Ferris, G. R. (2011). Accountability and extra-role behavior. *Employee Responsibilities and Rights Journal*, 23(2), 131-144.
- Hall, A. T., Frink, D. D., Ferris, G. R., Hochwarter, W. A., Kacmar, C. J., & Bowen, M. G. (2003). Accountability in human resources management. *New Directions in Human Resource Management*, 29-63.
- Hall, A. T., Frink, D. D., & Buckley, M. R. (2017). An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *Journal of Organizational Behavior*, 38(2), 204-224.
- 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.
- Hedricks, C. A., Rupayana, D., Puchalski, L., & Robie, C. (2018). Content of qualitative feedback provided during structured, confidential reference checks. *Personnel Assessment and Decisions*, 4(1), 41-54.
- Hedricks, C. A., Rupayana, D. D., Fisher, P. A., & Robie, C. (2019). Factors affecting compliance with reference check requirements. *International Journal of Selection and Assessment*, 27(2), 139-151.

- Hershey, N. (1992). Compensation and accountability: The way to improve peer review. *Quality Assurance and Utilization Review*, 7(1), 23-29.
- Hochwarter, W. A., Ferris, G. R., Gavin, M. B., Perrewé, P. L., Hall, A. T., & Frink, D. D. (2007). Political skill as neutralizer of felt accountability—job tension effects on job performance ratings: A longitudinal investigation. *Organizational Behavior and Human Decision Processes*, 102(2), 226-239.
- Holder, W. B. (1962). Letters of Recommendation. *American Psychologist*, 17(7), 506–507.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods*, 5(1), 64-86.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96(1), 72-98.
- James, J. M., & Bolstein, R. (1990). The effect of Monetary Incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly*, 54(3), 346-361.
- James, J. M., & Bolstein, R. (1992). Large Monetary Incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56(4), 442-453.
- Jelley, R. B., & Goffin, R. D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology*, 86(1), 134-144.
- Jobber, D., Saunders, J., & Mitchell, V. W. (2004). Prepaid Monetary Incentive effects on mail survey response. *Journal of Business Research*, 57(1), 21-25.
- Jones, A., & Harrison, E. (1982). Prediction of performance in initial officer training using reference reports. *Journal of Occupational Psychology*, 55(1), 35-42.

- Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes*, 45(2), 194-208.
- Knouse, S. B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology*, 36(2), 331-341.
- Knouse, S. B. (1985). The letter of recommendation: Confidentiality and unfavorability of information. *Proceedings of the Eastern Academy of Management*, 22, 97-99.
- Knouse, S. B. (1987). Confidentiality and the letter of recommendation: A new approach. *The Bulletin of the Association for Business Communication*, 50(3), 6-8.
- Knouse, S. B. (1994). Improving the letter of recommendation. *Journal of Employment Counseling*, 31(3), 105-109.
- Konradt, U., Hertel, G., & Joder, K. (2003). Web-based assessment of call center agents: Development and validation of a computerized instrument. *International Journal of Selection and Assessment*, 11(2-3), 184-193.
- Laird, M. D., Perryman, A. A., Hochwarter, W. A., Ferris, G. R., & Zinko, R. (2009). The moderating effects of personal reputation on accountability-strain relationships. *Journal of Occupational Health Psychology*, 14(1), 70-83.
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255-275.
- Lewin, K. 1936. *Principles of topological psychology*. New York: McGraw-Hill.
- Lopresto, R., Mitcham, D. E., & Ripley, D. E. (1985). *Reference Checking Handbook*. Alexandria, VA: ASPA.

- McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three standardized reference forms. *Military Psychology, 13*(4), 199-222.
- Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. *Journal of Management, 33*(2), 223-252.
- Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*(4), 517-524.
- Miller, R. K., & Van Rybroek, G. J. (1988). Internship letters of recommendation: Where are the other 90%?. *Professional Psychology: Research and Practice, 19*(1), 115-117.
- Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation. *Journal of Applied Social Psychology, 43*(11), 2297-2306.
- Mosel, J. N., & Goheen, H. W. (1952). Agreement among replies to an employment recommendation questionnaire. *American Psychologist, 7*, 365-366.
- Mosel, J. N., & Goheen, H. W. (1958). The Validity of the Employment Recommendation Questionnaire in Personnel Selection: I. Skilled Traders 1. *Personnel Psychology, 11*(4), 481-490.
- Mosel, J. N., & Goheen, H. W. (1959). The Employment Recommendation Questionnaire: III. Validity of Different Types of References 1. *Personnel Psychology, 12*(3), 469-478.
- Muchinsky, P. M. (1979). The use of reference reports in personnel selection: A review and evaluation. *Journal of Occupational Psychology, 52*(4), 287-297.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.

- Nash, A. N., & Carroll Jr, S. J. (1970). A hard look at the reference check: Its modest worth can be improved. *Business Horizons*, 13(5), 43-49.
- Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13(2), 248-277.
- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574-583
- Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. *Journal of Applied Psychology*, 96(4), 762.
- Patil, S. V., Vieider, F., & Tetlock, P. E. (2014). Process versus outcome accountability. *The Oxford handbook of public accountability*, 69-89.
- Peres, S. H., & Garcia, J. R. (1962). Validity and dimensions of descriptive adjectives used in reference letters for engineering applicants. *Personnel Psychology*, 15(3), 279-286.
- Porter, S. R., & Whitcomb, M. E. (2003). The impact of lottery incentives on student survey response rates. *Research in Higher Education*, 44(4), 389-407.
- Powell, D. M., & Bourdage, J. S. (2016). The detection of personality traits in employment interviews: Can “good judges” be trained?. *Personality and Individual Differences*, 94, 194-199.
- Powell, D. M., & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. *Human Performance*, 22(5), 450-465.

- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36(4), 717-731.
- 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(5), 389-392.
- Reference. (n.d.). In *Merriam-Webster.com dictionary*. Retrieved from <https://www.merriam-webster.com/dictionary/reference>
- Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35(1), 1-62.
- Robiner, W. N., Saltzman, S. R., Hoberman, H. M., Semrud-Clikeman, M., & Schirvar, J. A. (1998). Psychology supervisors' bias in evaluations and letters of recommendation. *The Clinical Supervisor*, 16(2), 49-72.
- Rose, D. S., Sidle, S. D., & Griffith, K. H. (2007). A penny for your thoughts: Monetary incentives improve response rates for company-sponsored employee surveys. *Organizational Research Methods*, 10(2), 225-240.
- Ryan, C. (2003). Stereotype accuracy. *European Review of Social Psychology*, 13(1), 75-109.
- Schlenker, B. R. (1986). Personal accountability: Challenges and impediments in the quest for excellence. (Technical report for the Navy Personnel Research and Development Center). San Diego, CA: Navy Personnel Research and Development Center.
- Schlenker, B. R., & Weigold, M. F. (1989). Self-identification and accountability. In R. A. Giacalone & P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 21–43). Lawrence Erlbaum Associates, Inc.

- Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632-652.
- Schmid Mast, M., Bangerter, A., Bulliard, C., & Aerni, G. (2011). How accurate are recruiters' first impressions of applicants in employment interviews?. *International Journal of Selection and Assessment*, *19*(2), 198-208.
- 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*(2), 262-274.
- Shaffer, D. R., Mays, P. V., & Etheridge, K. (1976). Who shall be hired: A biasing effect of the Buckley Amendment on employment practices?. *Journal of Applied Psychology*, *61*(5), 571-575.
- Shaffer, D. R., & Tomarelli, M. (1981). Bias in the Ivory Tower: An unintended consequence of the Buckley Amendment for graduate admissions?. *Journal of Applied Psychology*, *66*(1), 7-11.
- Sheppard, L. D., Goffin, R. D., Lewis, R. J., & Olson, J. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology*, *10*, 24-33.
- Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, *65*(1), 1-17.
- Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112-141.
- Sleight, R. B., & Bell, G. D. (1954). Desirable content of letters of recommendation. *Personnel Journal*, *32*, 421-422.

- Society for Human Resource Management. (2005). *2004 Reference and Background Checking Survey Report*. Alexandria, VA: Society for Human Resource Management.
- Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology, 57*(3), 745-772.
- Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. *Research in Organizational Behavior, 7*(1), 297-332.
- Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57*(3), 388-398.
- Tetlock, P. E., & Boettger, R. (1994). Accountability amplifies the status quo effect when change creates victims. *Journal of Behavioral Decision Making, 7*(1), 1-23.
- Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52*(4), 700-709.
- Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57*(4), 632-640.
- Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology, 25*(3), 331-76.
- Vance, A., Lowry, P. B., & Eggett, D. (2015). Increasing Accountability Through User-Interface Design Artifacts. *Mis Quarterly, 39*(2), 345-366.
- Woska, W. J. (2007). Legal issues for HR professionals: Reference checking/background investigations. *Public Personnel Management, 36*(1), 79-89.

Zimmerman, R. D., Triana, M. D. C., & Barrick, M. R. (2010). Predictive criterion-related validity of observer ratings of personality and job-related competencies using multiple raters and multiple performance criteria. *Human Performance*, 23(4), 361-378.

Table 1.

Correlation matrix for the study variables of interest

Item	<i>M (SD)</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Felt accountability	5.97 (.79)	(.65)													
2. Reason Giving	-.02 (1.00)	.02													
3. Monetary Incentive	-.01 (1.00)	.11	.02												
4. Elevation	.36 (.30)	.13	.18	.08											
5. Differential Elevation	.67 (.29)	.15	.04	.01	.08										
6. Stereotype Accuracy	.39 (.19)	-.06	.16	.01	.10	-.23									
7. Differential Accuracy	.67 (.22)	.05	-.08	.03	.15	.12	.20								
8. Rater confidence	4.80 (.41)	.28	-.01	-.03	.12	.13	-.13	-.01							
9. Rater perceived value	4.73 (.51)	.42	-.06	-.04	.14	.11	-.05	.08	.52						
10. Study Time	43.23 (18.40)	.03	.66	.04	.17	.08	.15	-.01	-.01	.01					
11. Age	41.33 (10.04)	-.04	-.03	.02	-.11	.15	-.18	-.04	-.02	-.03	.11				
12. Gender	1.52 (.50)	.15	-.04	-.13	.04	.06	.02	.17	.04	.05	.05	.04			
13. Work Experience	21.17 (9.58)	-.02	.03	.05	-.06	.15	-.16	.01	.00	-.03	.11	.91	.04		
14. Supervisory Experience	9.42 (7.07)	-.07	-.05	-.01	-.09	.06	-.07	-.02	-.04	-.05	.07	.67	-.02	.71	
15. Past references	13.17 (14.93)	.06	.02	-.02	.11	.11	-.10	.02	.09	.06	-.01	.37	-.07	.40	.51

Note. $N = 211$. Alpha reliabilities where available are reported in brackets on the diagonal. Correlations of .14 or greater are significant at $p = .05$. Correlations of .19 or greater are significant at $p = .01$. Reason Giving coded as 1 = present or -1 = absent; Monetary Incentive coded as 1 = present or -1 = absent; Gender coded as 1 = Male, 2 = Female.

Table 2.

Means and Standard deviations for all accuracy components across conditions

		Monetary Incentive					
		<u>Yes</u>		<u>No</u>		<u>Overall</u>	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<u>Reason Giving</u>							
<u>Yes</u>							
	<i>EL</i>	.467	.309	.336	.312	.420	.313
	<i>DE</i>	.691	.321	.635	.300	.665	.311
	<i>SA</i>	.396	.198	.437	.203	.415	.200
	<i>DA</i>	.662	.265	.629	.181	.647	.229
<u>No</u>							
	<i>EL</i>	.305	.289	.320	.263	.313	.273
	<i>DE</i>	.619	.274	.686	.263	.654	.269
	<i>SA</i>	.374	.180	.341	.165	.357	.172
	<i>DA</i>	.682	.222	.681	.207	.682	.214
<u>Overall</u>							
	<i>EL</i>	.382	.307	.340	.284		
	<i>DE</i>	.653	.298	.664	.279		
	<i>SA</i>	.384	.188	.382	.187		
	<i>DA</i>	.673	.242	.659	.197		

Note. $N = 196$. Lower mean scores indicate greater accuracy. EL = Elevation; DE = Differential Elevation; SA = Stereotype Accuracy; DA = Differential accuracy.

Table 3.

Univariate F-tests for the Main Effect of Reason Giving

Dependent Variable	<i>df</i>	<i>MS</i>	<i>F</i>	<i>n</i> ²
<i>EL</i>	3, 207	.591	7.027*	.033
<i>DE</i>	3, 205	.035	.423	.002
<i>SA</i>	3, 205	.191	5.320*	.025
<i>DA</i>	3, 192	.063	1.282	.007

Note. *N* = 211. *MS* = Mean Square; *EL* = elevation; *DE* = differential elevation; *SA* = stereotype accuracy; *DA* = differential accuracy; * = significant at $p < .05$.

Table 4.

Univariate F-tests for the Main Effect of the Monetary Incentive

Dependent Variable	<i>df</i>	<i>MS</i>	<i>F</i>	<i>n</i> ²
EL	3, 207	.108	1.285	.006
DE	3, 205	.004	.046	.000
SA	3, 205	.001	.014	.000
DA	3, 192	.014	.281	.001

Note. $N = 211$. MS = Mean Square; EL = elevation; DE = differential elevation; SA = stereotype accuracy; DA = differential accuracy; * = significant at $p < .05$.

Table 5.

Mean scores on felt accountability as a function of Reason Giving and the Monetary Incentive

Reason Giving	Prepaid Monetary Incentive		Overall
	Yes	No	
Yes	6.12 (.76)	5.83 (.83)	5.98 (.80)
No	5.99 (.77)	5.93 (.81)	5.96 (.79)
Overall	6.06 (.77)	5.87 (.82)	5.97 (.79)

Note. $N = 211$. Scores are based on a 7-point Likert-type scale. Values in brackets are standard deviations.

Table 6.

Mediation Tests for the Role of Felt Accountability

Model Number	IV	DV	Effect of IV on FA (a)	Unique effect of FA (b)	Indirect effect (ab)	95% CI	
						Lower	Upper
1.	<i>RG</i>	<i>EL</i>	.013 (.055)	.047 (.025)	.001 (.003)	-.006	.007
2.	<i>RG</i>	<i>DE</i>	.006 (.055)	.005 (.020)	.000 (.003)	-.006	.007
3.	<i>RG</i>	<i>SA</i>	.015 (.055)	-.014 (.017)	.000 (.001)	-.003	.003
4.	<i>RG</i>	<i>DA</i>	.026 (.057)	.015 (.020)	.000 (.002)	-.002	.004
5.	<i>MI</i>	<i>EL</i>	.085 (.055)	.046 (.025)	.004 (.003)	-.002	.012
6.	<i>MI</i>	<i>DE</i>	.088 (.055)	.054 (.025)	.005 (.004)	-.001	.014
7.	<i>MI</i>	<i>SA</i>	.088 (.055)	-.014 (.017)	-.001 (.002)	-.006	.002
8.	<i>MI</i>	<i>DA</i>	.067 (.056)	.013 (.020)	.001 (.002)	-.002	.006

Note. $N = 196$. *RG* = Reason Giving; *MI* = Monetary Incentive; *EL* = elevation; *DE* = differential elevation; *SA* = stereotype accuracy; *DA* = differential accuracy.

Appendix A: Ethics Approval



Date: 13 July 2021

To: Prof. Richard Goffin

Project ID: 119145

Study Title: Call Center References

Short Title: Call Center References

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: August 6 2021

Date Approval Issued: 05/Jul/2021 19:46

REB Approval Expiry Date: 13/Jul/2022

Dear Prof. Richard Goffin

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. **All other required institutional approvals and mandated training must also be obtained prior to the conduct of the study.**

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Interpersonal Attraction Questionnaire	Online Survey		
Felt Accountability Scale	Online Survey		
Personality Reference Rating Form	Online Survey		
Reference Rating Instructions	Online Survey		
Demographic Questionnaire	Online Survey		
Final Scale	Online Survey		
Single Item Recommendation	Online Survey		
Monetary Incentive Manipulation	Online Survey		
Rating Explanation Manipulation	Online Survey		
Media Check Question	Other Data Collection Instruments		
Prolific Voice Check	Other Data Collection Instruments		
Employee videos	Other Data Collection Instruments		
Prolific Study Description - NonExclusive	Recruitment Materials		
Prolific Study Description - Male	Recruitment Materials		
Prolific Study Description - Female	Recruitment Materials		
Letter of Information - NonExclusive	Implied Consent/Assent		
Letter of Information - Female	Implied Consent/Assent		
Letter of Information - Male	Implied Consent/Assent		
Debrief Form	Debriefing Letter		
Cover Story	Online Survey		

No deviations from, or changes to the protocol should be initiated without prior written approval from the NMREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Kelly Patterson , Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Appendix B: Cover Story

Study background:

As you may know, nearly all organizations use supervisory reference evaluations when hiring.

However, some organizations may be concerned about their usefulness for hiring call center workers, since these employees mainly work alone and unsupervised.

A new call center company planning to open at the end of this year recently contacted us (the Human Resources Lab) to express their concerns about using supervisory reference evaluations.

They have since hired our Lab to provide evidence-based suggestions for how they can best use supervisory reference evaluations in their upcoming mass-hiring system.

The Human Resources Lab now requires your assistance to help inform our suggestions for this call center's new hiring system.

To ensure that our suggestions will be valid, we are asking you to provide supervisory reference evaluations for four call center employees whom you will soon see in a series of videos. These employees are similar to those who may actually be hired by this organization.

You have indicated that you currently have some level of supervisory duties at work, so we believe that your perspectives are particularly valuable to our research goals.

Your help in this research will ensure that our suggestions are based on the insights of real supervisory professionals, which will improve the accuracy and fairness of this call center's hiring system.

Appendix C: Reason Giving Manipulation

Please now review the following characteristics and provide a rating towards how each of them relates to your employee, (employee name).

Please also provide brief explanations of your descriptions in the textboxes below, just as if you were providing us with a real reference evaluation of (employee name).

2. Extraversion:

High: Enjoys friendship of others; confident and comfortable in social situations; tries to control environment and influence or direct people; likes to have an audience and to be the centre of attention.

Low: Has little interest in meeting new people or initiating conversation; avoids expressing opinions or assuming positions of power; prefers to go unnoticed or remain anonymous.

I would describe (employee name)'s level of Extraversion as:

1. Very low, 2. Low, 3. Somewhat low, 4. Moderate, 5. Somewhat high, 6. High, 7. Very high

Please outline some of the reasons you considered when selecting from the above descriptions (using at least 30 words). For example, you may have thought about certain cues or behaviours that you believe showed the presence (or absence) of Extraversion in (employee name).

Word count: 0

Appendix D: Monetary Incentive Manipulation

Because we deeply appreciate you sharing your unique perspectives with us, we will be providing you with an additional £6 payment as a token of our appreciation for your efforts.

This means that you will now be receiving £10 in your Prolific account for your assistance with this study instead of the £4 you were initially promised. We thank you in advance for your help!

Appendix E: Personality Rating Form

Please review each of the following characteristics and describe how you think they apply to (employee name) using the scales provided.

1. Agreeableness:

High: Trusting of others; helpful; cooperative; accepts criticism and blame; avoids confrontations and conflicts, humble.

Low: Skeptical of others' intentions; enjoys arguments; stubborn; easily annoyed; takes offence easily; avoids apologizing.

I would describe (employee name)'s level of Agreeableness as:

1. Very low, 2. Low, 3. Somewhat low, 4. Moderate, 5. Somewhat high, 6. High, 7. Very high

2. Extraversion:

High: Enjoys friendship of others; confident and comfortable in social situations; tries to control environment and influence or direct people; likes to have an audience and to be the center of attention.

Low: Has little interest in meeting new people or initiating conversation; avoids expressing opinions or assuming positions of power; prefers to go unnoticed or remain anonymous.

I would describe (employee name)'s level of Extraversion as:

1. Very low, 2. Low, 3. Somewhat low, 4. Moderate, 5. Somewhat high, 6. High, 7. Very high

Please be careful when providing your descriptions; they will help improve our hiring system.

3. Conscientiousness:

High: Efficient; neat and organized; careful; plans before acting; aspires to reach challenging goals; takes a serious approach to life; self-disciplined.

Low: Tends to act on the spur of the moment; not concerned with neatness or organization; gives up quickly and loses drive over time; enjoys doing things for pleasure rather than achievement; has an easy-going attitude toward life.

I would describe (employee name)'s level of Conscientiousness as:

1. Very low, 2. Low, 3. Somewhat low, 4. Moderate, 5. Somewhat high, 6. High, 7. Very high

4. Emotional Stability:

High: Does not worry about things; is comfortable with themselves; tends to be relaxed and remains calm under pressure; rarely gets irritated.

Low: Prone to worrying; often down in the dumps; has frequent mood swings; becomes stressed out often or panics easily.

I would describe (employee name)'s level of Neuroticism as:

1. Very low, 2. Low, 3. Somewhat low, 4. Moderate, 5. Somewhat high, 6. High, 7. Very high

The descriptions you provide here are important; they will help improve our hiring system.

Appendix F: Felt Accountability Scale

(adapted from Desai & Kouchaki, 2015; Hochwarter et al., 2007)

Please read each of the following statements and decide how much you agree or disagree with them, drawing on your experiences evaluating Andrew, Emmanuel, Jordan and Sam.

1. I feel accountable for my descriptions of the employees shown in the videos

Strongly disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly agree

2. The researchers paying me for this task hold me accountable for my descriptions

Strongly disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly agree

3. If I am not careful when making my descriptions, I will hear about it from the researchers paying me to provide them

Strongly disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly agree

4. The hiring systems of future call center workers depend on the carefulness of my descriptions

Strongly disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly agree

5. The people paying me for my descriptions closely scrutinize my efforts

Strongly disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly agree

Appendix G: Affective Reaction Items

1. I feel confident about the accuracy of my descriptions

Strongly disagree, Disagree, Neutral, Agree, Strongly agree

2. I feel that I provided valuable insights for this research

Strongly disagree, Disagree, Neutral, Agree, Strongly agree

Appendix H: Participant Effort Items

1. I put effort into watching the videos - I tried to identify and understand the employees' characteristics as best as I could

Strongly disagree, Disagree, Neutral, Agree, Strongly agree

2. I put effort into my descriptions – I tried to provide careful and accurate descriptions of the employees to the best of my abilities

Strongly disagree, Disagree, Neutral, Agree, Strongly agree

Appendix I: Demographics Questionnaire

This first section contains questions that will help us understand some of your characteristics.

Please answer honestly, all of your responses throughout this survey will be kept strictly confidential.

1. What is your gender?

Female, Male, You do not have an option that applies to me. I identify as (please state):

2. What is your ethnicity?

Asian (e.g., Chinese, Japanese, Korean, Vietnamese), Aboriginal/First Nations, Black North American/African, East Asian (e.g., Indian, Pakistani), Hispanic, White/White European, Other (please specify):

3. What is your primary language?

English, Other (please specify):

4. Please select your age on the slider:

Age:

5. Have you ever had a job? Please choose the option that describes your highest level of employment:

No, I have never had a job, Yes, I have had a full-time job (25 or more hours per week), Yes, I have had a part-time job (10-24 hours per week), Yes, I have had a part-time job (9 hours or fewer per week),

6. What type of work have you done (choose all that apply):

Retail sales, Cashier, Office Clerk, Food Service/Food Preparation, Nurse, Nursing Assistant, Orderly or Personal Service Worker, Waiter or Waitress, Customer Service Representative, Material Mover (truck driver, truck loader, loading dock worker, baggage handler), Janitor, Stock Clerk or Order Filler (Working in storage facilities, warehouses, or shipping/receiving), Management, Education, Manufacturing, Other (please specify):

7. Approximately how long were you employed in the job that you remained in the longest? Please select the total number of years using the slider:

Number of years:

8. Approximately how many years of past work experience have you had in total in your life? Please select the total number of years using the slider:

Total years:

9. Approximately how many years of past supervisory experience have you had in total in your life? Please select the total number of years using the slider:

Total years:

**10. Approximately how many references have you previously provided in total in your life?
Please select the total number using the slider:**

Total references:

11. What is the highest level of school you have completed or the highest degree you have received?

Less than secondary school degree, Secondary school degree or equivalent (e.g., GED), Some college without degree completion, some university without degree completion, Skilled trade certification (e.g., plumber, electrician, carpenter, etc.), College degree or certification, Undergraduate degree (e.g., B.A., B.Sc.), Graduate degree (e.g., M.A., M.Sc., Ph.D.), Other (please specify):

Appendix J: Letters of Information and Consent

General

Project Title: Call Center References

Principal Investigator: Richard Goffin, Ph.D., Western University

Additional Research Staff: Cullen McCurrach, Western University

1. Invitation to Participate - You are being invited to participate in this research study about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, and (c) you are currently employed for at least 20 hours per week.

2. Why is this study being done? The purpose of this study is to learn more about the usefulness of supervisory references in call centers. More specifically, we are investigating how supervisors and managerial personnel such as yourself make evaluations of their past employees in call centers. We are focusing on call center workers in particular because this is one of the most prevalent jobs in the economy and because these employees often work unsupervised.

3. How long will you be in this study? It is expected that the study will take 45 minutes of your time to complete, in a single session. You will not be asked to complete additional sessions or follow-up studies.

4. What are the study procedures? If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls. You will then be asked to evaluate some of their job-relevant characteristics according to how they behaved during the calls. Because we are interested in understanding how call center supervisors form evaluations of their employees, you may also be asked to provide written explanations of your chosen descriptions.

5. What are the risks and harms of participating in this study? There are no known or anticipated risks or discomforts associated with participating in this study.

6. What are the benefits of participating in this study? The possible benefits to participants could be the opportunity to learn more about supervisory reference evaluations of employees in call centers. The possible benefits to society may be a greater understanding of how to use supervisory reference evaluations when hiring. References such as those collected in this study play an important role in determining who does and does not get hired, so it is important for us to understand how they are made if we are to be as fair as possible to job applicants.

7. Can participants choose to leave the study? Yes, you can withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving the study will still be used. No new information will be collected without your permission. If you prefer to completely withdraw all of your data from this study, instructions on how to do that will be provided at the end of the survey along with a unique completion ID number that will be required for that purpose.

Male

Project Title: Call Center References

Principal Investigator: Richard Goffin, Ph.D., Western University

Additional Research Staff: Cullen McCurrach, Western University

1. Invitation to Participate - You are being invited to participate in this research study about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, (c) you are currently employed for at least 20 hours per week, and (d) you are registered with Prolific as a male, as we currently need responses from males in this research in order to get equal numbers of males and females.

2. Why is this study being done? The purpose of this study is to learn more about the usefulness of supervisory references in call centers. More specifically, we are investigating how supervisors and managerial personnel such as yourself make evaluations of their past employees in call centers. We are focusing on call center workers in particular because this is one of the most prevalent jobs in the economy and because these employees often work unsupervised.

3. How long will you be in this study? It is expected that the study will take 45 minutes of your time to complete, in a single session. You will not be asked to complete additional sessions or follow-up studies.

4. What are the study procedures? If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls. You will then be asked to evaluate some of their job-relevant characteristics according to how they behaved during the calls. Because we are interested in understanding how call center supervisors form evaluations of their employees, you may also be asked to provide written explanations of your chosen descriptions.

5. What are the risks and harms of participating in this study? There are no known or anticipated risks or discomforts associated with participating in this study.

6. What are the benefits of participating in this study? The possible benefits to participants could be the opportunity to learn more about supervisory reference evaluations of employees in call centers. The possible benefits to society may be a greater understanding of how to use supervisory reference evaluations when hiring. References such as those collected in this study play an important role in determining who does and does not get hired, so it is important for us to understand how they are made if we are to be as fair as possible to job applicants.

7. Can participants choose to leave the study? Yes, you can withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving the study will still be used. No new information will be collected without your permission. If you prefer to completely withdraw all of your data from this study, instructions on how to do that will be provided at the end of the survey along with a unique completion ID number that will be required for that purpose.

Female

Project Title: Call Center References

Principal Investigator: Richard Goffin, Ph.D., Western University

Additional Research Staff: Cullen McCurrach, Western University

1. Invitation to Participate - You are being invited to participate in this research study about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, (c) you are currently employed for at least 20 hours per week, and (d) you are registered with Prolific as a female, as we currently need responses from females in this research in order to get equal numbers of males and females.

2. Why is this study being done? The purpose of this study is to learn more about the usefulness of supervisory references in call centers. More specifically, we are investigating how supervisors and managerial personnel such as yourself make evaluations of their past employees in call centers. We are focusing on call center workers in particular because this is one of the most prevalent jobs in the economy and because these employees often work unsupervised.

3. How long will you be in this study? It is expected that the study will take 45 minutes of your time to complete, in a single session. You will not be asked to complete additional sessions or follow-up studies.

4. What are the study procedures? If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls. You will then be asked to evaluate some of their job-relevant characteristics according to how they behaved during the calls. Because we are interested in understanding how call center supervisors form evaluations of their employees, you may also be asked to provide written explanations of your chosen descriptions.

5. What are the risks and harms of participating in this study? There are no known or anticipated risks or discomforts associated with participating in this study.

6. What are the benefits of participating in this study? The possible benefits to participants could be the opportunity to learn more about supervisory reference evaluations of employees in call centers. The possible benefits to society may be a greater understanding of how to use supervisory reference evaluations when hiring. References such as those collected in this study play an important role in determining who does and does not get hired, so it is important for us to understand how they are made if we are to be as fair as possible to job applicants.

7. Can participants choose to leave the study? Yes, you can withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving the study will still be used. No new information will be collected without your permission. If you prefer to completely withdraw all of your data from this study, instructions on how to do that will be provided at the end of the survey along with a unique completion ID number that will be required for that purpose.

Appendix K: Prolific Study Information

General

Project Title: Call Center References

You are being invited to participate in this research project about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, and (c) you are currently employed for at least 20 hours per week. The purpose of this study is to learn more about the usefulness of supervisory references in call centers, because this one of the most prevalent jobs in the economy and because call center employees often work unsupervised.

If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment that is free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls, and you will then be asked to evaluate some of their job-relevant characteristics according to how they behaved. Because we are interested in understanding how call center supervisors form reference evaluations of their employees, you may also be asked to provide written explanations of your chosen ratings. Completing this survey in a calm environment is strongly recommended. Using headphones is also encouraged.

It is expected that the study will take 45 minutes of your time to complete in a single session. You will be compensated £4 for your participation, and you will not be asked to complete additional sessions or follow-up studies. Your rewards will be provided on the first Saturday following your survey completion.

You can choose to withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving will still be used. If you prefer to completely withdraw all of your data from this study, instructions on how to do so will be provided at the end of the survey along with a unique completion ID that will be required for that purpose. No names or identifying information will be collected, so your responses will always be anonymous. However, this data may be shared with others for the purposes of collaboration or publication.

If you have questions about this research, please contact Cullen McCurrach, Graduate Student, University of Western Ontario
Faculty, University of Western Ontario
If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Human Research Ethics

Male

Project Title: Call Center References

You are being invited to participate in this research project about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, (c) you are currently employed for at least 20 hours per week, and (d) you are registered with Prolific as a male, as we currently need responses from males in this research in order to get equal numbers of males and females. The purpose of this study is to learn more about the usefulness of supervisory references in call centers, because this one of the most prevalent jobs in the economy and because call center employees often work unsupervised.

If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment that is free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls, and you will then be asked to evaluate some of their job-relevant characteristics according to how they behaved. Because we are interested in understanding how call center supervisors form reference evaluations of their employees, you may also be asked to provide written explanations of your chosen ratings. Completing this survey in a calm environment is strongly recommended. Using headphones is also encouraged.

It is expected that the study will take 45 minutes of your time to complete in a single session. You will be compensated £4 for your participation, and you will not be asked to complete additional sessions or follow-up studies. Your rewards will be provided on the first Saturday following your survey completion.

You can choose to withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving will still be used. If you prefer to completely withdraw all of your data from this study, instructions on how to do so will be provided at the end of the survey along with a unique completion ID that will be required for that purpose. No names or identifying information will be collected, so your responses will always be anonymous. However, this data may be shared with others for the purposes of collaboration or publication.

If you have questions about this research, please contact Cullen McCurrach, Graduate Student, University of Western Ontario
Faculty, University of Western Ontario
You may also contact Dr. Richard Goffin,
If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Human Research Ethics

Female

Project Title: Call Center References

You are being invited to participate in this research project about call center references because: (a) you have a Prolific Academic account, (b) you have managerial or supervisory duties at work, (c) you are currently employed for at least 20 hours per week, and (d) you are registered with Prolific as a female, as we currently need responses from females in this research in order to get equal numbers of males and females. The purpose of this study is to learn more about the usefulness of supervisory references in call centers, because this one of the most prevalent jobs in the economy and because call center employees often work unsupervised.

If you agree to participate in this study, you will be asked to complete an online survey on your own in a quiet environment that is free of distractions. First, you will be asked to complete some demographic questions. Then, you will be asked to assume the role of a call center supervisor who has agreed to provide reference evaluations for four of their hypothetical past employees. In this part of this study, you will be shown pre-recorded videos of these employees handling typical customer service calls, and you will then be asked to evaluate some of their job-relevant characteristics according to how they behaved. Because we are interested in understanding how call center supervisors form reference evaluations of their employees, you may also be asked to provide written explanations of your chosen ratings. Completing this survey in a calm environment is strongly recommended. Using headphones is also encouraged.

It is expected that the study will take 45 minutes of your time to complete in a single session. You will be compensated £4 for your participation, and you will not be asked to complete additional sessions or follow-up studies. Your rewards will be provided on the first Saturday following your survey completion.

You can choose to withdraw from the study at any time simply by closing your browser window. However, if you decide to withdraw from the study, the information that was collected prior to you leaving will still be used. If you prefer to completely withdraw all of your data from this study, instructions on how to do so will be provided at the end of the survey along with a unique completion ID that will be required for that purpose. No names or identifying information will be collected, so your responses will always be anonymous. However, this data may be shared with others for the purposes of collaboration or publication.

If you have questions about this research, please contact Cullen McCurrach, Graduate Student, University of Western Ontario
Faculty, University of Western Ontario
If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Human Research Ethics
ethics@uwo.ca

Appendix L: Study Debrief

Project Title: Call Center References

Principal Investigator: Richard Goffin, Ph.D., Western University,

Additional Researcher: Cullen McCurrach, Western University,

Thank you for your participation in our research. Here's some further details about this research that you might find interesting.

Often, supervisors providing references respond inaccurately when asked to evaluate job applicants. This inaccurate responding reduces the effectiveness of their references for hiring organizations. This study is being done to learn more about how certain elements of the evaluation situation may influence references' response tendencies when describing the personality of job applicants. For instance, we are assessing two unique approaches that may reduce the tendency for references to provide inaccurate trait descriptions to hiring organizations. More specifically, this research responds to calls from various researchers (Colarelli et al., 2002; Morgan et al., 2013; Hall et al., 2017; Hedricks et al., 2013, 2018) to investigate new sources of accountability in organizations, and to assess how certain interventions may reduce the presence of inaccurate references and lead to fairer hiring decisions.

This research is important because using references when hiring can lead to greater productivity and better overall fit for organizations. But, to the extent that references provide inaccurate evaluations of job applicants, their capacity to facilitate higher productivity and fit is ultimately reduced.

Although we initially mentioned that your descriptions would be used to develop the hiring system of a new call center organization, this is not actually the case. In reality, your ratings will be used to explore the effects of two unique approaches to improving the accuracy of references' personality ratings, which we will attempt to publish after this study has been completed. So, although the descriptions you provided do have the capacity to inform future hiring systems, they may actually only do so indirectly through the results that will potentially be published from this research. Creating this cover story was necessary to replicate actual reference scenarios where supervisors' evaluations are often provided to an external organization to whom they may feel accountable.

If you have any further questions about this study, you may contact Cullen McCurrach,
at _____ or the project supervisor, Dr. Richard Goffin at _____

For any questions or concerns about the conduct of this study or your rights as a participant, please contact:

Appendix M: Curriculum Vitae

CULLEN W. D. MCCURRACH**Industrial/Organizational Psychology***Curriculum Vitae**August 2020***EDUCATION**

Western University London, Ontario	Master of Science (Industrial/Organizational Psychology) Anticipated Graduation October 2021
---------------------------------------	--

University of Guelph London, Ontario	Bachelor of Arts, Honours (Psychology) Graduated June 2018
---	--

HONOURS AND AWARDS

Western Graduate Research Scholarship Western University	2019 – present
--	----------------

Distinction University of Guelph Awarded to graduates with a cumulative degree average above 80%.	2018 – 2019
--	-------------

Guelph Campus Co-Operative Scholarship University of Guelph Awarded annually to a member of the co-operative with strong academic standing.	2017
--	------

Dean's Honour List University of Guelph Awarded for maintaining a cumulative average above 80% with a full course-load.	2015 – 2017
--	-------------

RESEARCH AND TEACHING INTERESTS

*accountability; faking; personality; personnel selection; psychometrics; virtual teamwork.***RESEARCH EXPERIENCE**

MSc. Thesis Student The Goffin Lab Western University, Supervisor: Dr. Richard Goffin	9/2019 – present
---	------------------

Research Assistant, Honor's Thesis Student 9/2016 – 4/2017
 Guelph Selection Lab
 University of Guelph, **Supervisor:** Dr. Deborah Powell

Research Intern 9/2016 – 12/2016
 Guelph Organizational Trust Lab
 University of Guelph, **Supervisor:** Dr. Harjinder Gill

Research Intern 4/2016 – 8/2016
 Centre for Workers' Health and Well-being
 University of Guelph, **Supervisor:** Dr. Peter Hausdorf

Research Intern 1/2016 – 4/2016
 Guelph Selection Lab
 University of Guelph, **Supervisor:** Dr. Deborah Powell

PROFESSIONAL EXPERIENCE

Research & Development Intern 6/2021 – present
 SIGMA Assessment Systems
 London, ON

Research Assistant 6/2017 – 8/2019
 Organization and Management Solutions
 University of Guelph

Undergraduate Research Assistant (URA) 4/2017 – 8/2017
 Guelph Organizational Trust Lab
 University of Guelph, **Supervisor:** Dr. Harjinder Gill

Undergraduate Research Assistant (URA) 4/2016 – 8/2016
 Department of Marketing and Consumer Studies
 University of Guelph, **Supervisor:** Dr. Brent McKenzie

VOLUNTEER EXPERIENCE

I/O Department Liaison 7/2020 – present
 Psychology Graduate Student Association (PGSA)
 Western University

Volunteer Research Assistant 4/2018 – 4/2019
 Guelph Selection Lab
 University of Guelph

Project Lead 4/2017 – 8/2018
 Centre for Workers' Health and Well-being
 University of Guelph

Volunteer Research Assistant 4/2016 – 4/2018
 Guelph Organizational Trust Lab
 University of Guelph

CONFERENCE PRESENTATIONS

Carver, S., Goffin, R. D., & **McCurrach., C.** (2021, April). *Should Faking be Reviled or Revered in Pre-employment Personality Testing?*. Paper presented (virtually) at the 2021 SIOP Annual Conference, New Orleans, Louisiana.

McCurrach, C., & Leonard, C. M. (2020, May). *Introducing eSports: Bridging the gaps between sport psychology and virtual team research*. Oral symposium at the 80th Canadian Psychological Association Annual Convention, Montreal, Canada. (Conference cancelled)

Kearney, J. J., **McCurrach, C.,** Provencher, Y., & Hausdorf, P. A. (2018, June) *Into the mystic: Understanding the relation between music, mood and personality*. Paper presented at the 29th Annual International Congress of Applied Psychology, Montreal, QC.

Cassidy, S., Lech, O., **McCurrach, C.,** Miller, R., Zaroski, O., Gill., H. (2017, June) *Making the most of your virtual team: A thematic analysis of effective and maladaptive virtual team processes in rich- and lean- media environments*. Paper presented at the 78th annual Canadian Psychological Association National Convention, Toronto, ON.

McCurrach, C., Provencher, Y., Kearney, J. J., & Hausdorf, P. A. (2017, June) *Music and Mood: Does a Consistent Relationship Exist?* Paper presented at the 78th annual Canadian Psychological Association National Convention, Toronto, ON.

McCurrach, C., & Powell, D. M. (2017, May) *Deceptive Impression Management and Interview Performance: Machiavellianism as a Moderator*. Paper presented at the 47th Annual Ontario Psychology Honours Thesis Conference, Ottawa, ON.

MEMBERSHIPS

Canadian Psychological Association – Student Affiliate	2017 – Present
Society for Industrial Organizational Psychology – Student Member	2020 – Present