10-2018

The profile of the 'Good Judge' in HRM: A systematic review and agenda for future research

François S. DE KOCK
Erasmus University of Rotterdam

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Marise P. BORN
Erasmus University of Rotterdam
DOI: https://doi.org/10.1016/j.hrmr.2018.09.003

Published in Human Resource Management Review, October 2018,
DOI 10.1016/j.hrmr.2018.09.003

The profile of the 'Good Judge' in HRM: A systematic review and agenda for future research

François S. De Kock[a,*,1], Filip Lievens[b,2], Marise Ph. Born[a]

[a] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, The Netherlands
[b] Lee Kong Chian School of Business, Singapore Management University, Singapore

ARTICLE INFO

Keywords:
Judgment
Accuracy
Rater
Judge
Individual differences
HRM
Machine-learning

ABSTRACT

In light of the pivotal importance of judgments and ratings in human resource management (HRM) settings, a better understanding of the individual differences associated with being a good judge is sorely needed. This review provides an overview of individual difference characteristics that have been associated with the accurate judges in HRM. We review empirical findings over > 80 years to identify what we know and do not know about the individual difference correlates of being an accurate judge. Overall, findings suggest that judges' cognitive factors show stronger and more consistent relationships with rating accuracy than personality-related factors. Specific intelligences in the social cognition domain, such as dispositional reasoning (complex understanding of traits, behaviors and a situation's potential to manifest traits into behaviors) show particular promise to help understanding what makes an accurate judge. Importantly, our review also highlights the scarcity of research on HRM context (selection vs. performance appraisal settings) and judges' motivation to distort ratings. To guide future research, we present a model that links assessor constructs to key processes required for accurate judgment and ratings in HRM contexts. The discussion suggests twenty questions for future work in this field.

1. Introduction

In human resource management (HRM), ratings play a ubiquitous (Guion & Highhouse, 2011) role. Organizations rely on them to make important decisions about selection, promotion, and performance management (Schmitt & Chan, 1998). Hence, it is easy to see why so much effort has gone into understanding how people evaluate others in organizations (see Denisi & Murphy, 2017; Graves & Karren, 1992; London, 2001; Parsons, Liden, & Bauer, 2001; Schleicher et al., 2018) and in identifying the characteristics of the 'good judge'[3] (e.g., Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005; Graves, 1993; Powell & Goffin, 2009).

This paper focuses on the latter question of individual characteristics associated with accurate judges. Accurate judges are expected to produce evaluations in HRM that show adequate rating quality, broadly defined as the degree to which a person's ratings are accurate, not only as *measures* of other people's characteristics (e.g., interview dimensions) but also as *predictors* of important

* Correspondence author at: School of Management Studies, University of Cape Town, Private Bag X3, Cape Town 7701, South Africa.
E-mail address: francois.dekock@uct.ac.za (F.S. De Kock).

[1] François S. De Kock is now at the University of Cape Town.

[2] Filip Lievens completed parts of this research as visiting honorary professor of the University of Cape Town.

[3] In this paper, the terms 'judge' and 'rater' refer to interviewers and assessors (e.g., psychologists, managers, or other trained assessors). Although in social psychology the term the "good" judge is typically used, this essentially refers to the "accurate" judge because accuracy serves in most cases as dependent variable. Therefore, we will use the term "accurate" judge throughout the paper.

outcomes (e.g., job performance criteria). However, we are currently in the dark about the profile of the accurate judge in HRM (see Jones & Born, 2008). Granted, there exists a rich body of research in social psychology on the characteristics of raters that make accurate zero/minimal acquaintance judgments. Unfortunately, the typical designs in such zero/minimal acquaintance research (e.g., Ambady, Hallahan, & Rosenthal, 1995; Borkenau & Liebler, 1992; Funder & Colvin, 1988) are not insightful for HRM because they lack the external validity to draw generalizable conclusions. For example, they often rely on brief get-to-know-you sessions and involve less motivation from both assessees and assessors (Lievens, 2017). Results of this type of research are difficult to generalize to performance ratings for example, where managers and supervisors have extensive opportunities to observe and develop acquaintance with the people they rate, whereas assessors in interviews and ACs may be, on a continuum of acquaintance, closer to the minimal/short-acquaintance end of the scale. In addition, social psychology research often ignores that various contextual effects and people's motivation to distort ratings might also play a role (as is the case in HRM).

There are at least two other reasons why a review of individual differences related to judgment accuracy in HRM is relevant and timely. First, judges (e.g., recruiters, interviewers, assessors, HR-professionals, and managers that routinely evaluate their personnel) can be screened and can be trained (Lievens, 2001; Lievens, Tett, & Schleicher, 2009; Lorenzo, 1984; Roch, Woehr, Mishra, & Kieszczynska, 2012; Stamoulis & Hauenstein, 1993; Stillman & Jackson, 2005). To screen accurate judges we need a systematic review that identifies individual difference predictors of rating quality criteria in HRM. Similary, rater training programs may be adapted to train raters by targeting these particular individual differences constructs (e.g., raters' knowledge structures) identified by such a review.

A second reason why a systematic review is relevant and timely relates to recent developments in staffing practices. In particular, the study of the accurate judge has gained new traction in light of recent social media and speed assessment applications. For example, notwithstanding caveats (Davison, Bing, Kluemper, & Roth, 2016; Roth, Bobko, Van Iddekinge, & Thatcher, 2016; Van Iddekinge, Lanivich, Roth, & Junco, 2016), recruiters are increasingly using and judging social media information of applicants to infer applicants' traits and abilities (e.g., Landers, Brusso, Cavanaugh, & Collmus, 2016). Such social media information might consist of texts, pictures, and/or short videos. Likewise, in recent HR applications such as video resumes (e.g., Waung, Hymes, & Beatty, 2014), recruiters judge people's traits on the basis of short video-based excerpts. More generally, an important issue is that the rise of machine learning for rating purposes (e.g., via the use of natural language processing; see Campion, Campion, Campion, & Reider, 2016; Speer, 2018) will not make human judges moot. Instead, the rise of algorithms goes hand in hand with an in-depth understanding of the characteristics, benefits, and drawbacks of human judges so that it becomes clear when “humans” and “machines” can work most effectively together for rating purposes. Therefore, it is nowadays critical to understand which factors relate to being an accurate judge.

1.1. The present review and organizing framework

We review empirical research on individual differences that predict rating accuracy applicable to the HRM domain. Our review draws from three primary HRM fields: interviews, assessment centres, and performance appraisal. By weighing the evidence in support of individual difference predictors of accuracy, we will outline what we know and do not yet know about the characteristics of the accurate judge in HRM. We summarize this empirical research base in Tables 1 and 2. Finally, the review outlines 20 questions (see Table 3) that hold most potential for advancing knowledge of individual differences in rating accuracy.

The organizing framework for our review integrates two seminal models, namely the Realistic Accuracy Model (RAM) (Funder, 1995, 1999, 2012) and Murphy and Cleveland's (1995) context model of judgment and rating. Accordingly, the model in Fig. 1 links rater constructs to specific RAM stages in a specific context (selection vs. performance appraisal). In the RAM, accuracy results from a four-stage social cognitive process. The target person first emits a behavior that is (1) *relevant* to the trait to be judged, in a manner where this information is (2) *available* to the perceiver. Next, the rater must (3) *detect* and correctly (4) *utilize* the information to form an impression of an applicant characteristic. Essentially, we consider the RAM as depicting the four major tasks that judges need to undertake in judging and rating others. Following a classical personnel selection paradigm (see Schmitt, Arnold, & Nieminen, 2017) the stages in the RAM thus represent the results from a ‘job analysis’ of the judge. Our review then identifies the predictors, namely the range of rater individual difference constructs that may facilitate these key judgment tasks/processes in the RAM—cue detection and cue utilization—required to achieve high-quality rating outcomes in HRM. So, a distinguishing feature of our model is that it links rater individual differences to key judgment processes (namely cue detection and cue utilization) thought to cause accuracy (RAM) (Funder, 1999). As shown in Fig. 1, these individual difference constructs that will be the focus of our review are listed below the four main RAM stages.

In line with Murphy and Cleveland (1995), our model further highlights that judgment in HRM does not occur in a vacuum. That is, an important aspect of our model is that we explicitly include the HRM context (performance appraisal vs. selection) and distinguish between an accurate judge vs. an accurate rater. That is, “judgments represent private evaluations; ratings represent public statements about ratees' performance” (Murphy & Cleveland, 1995, p. 23). Therefore, ratings follow judgments in the ‘rendering phase’ (Banks & Murphy, 1985) where the rater fills in the rating form with an assigned score(s) that might be subject to perceived consequences of ratings, political pressures, personal goals to advance interests in their department, etc. (Spence & Keeping, 2011). The distinction between judgment and ratings is especially acknowledged in a performance appraisal context (less so in a selection context) and our model introduces a more nuanced way to understand what makes an accurate judge vs. an accurate rater. For the same reason, the model includes motivation to distort (Spence et al., 2011) as a moderator between judgments and ratings.

Table 1
Research evidence[a] on individual differences predictors of judgment accuracy in HRM research (sorted alphabetically within predictor cluster).

Nr	Author(s)	*N*	Cluster	Predictor	Effect size[b]
1	Borman (1979)	146	Affect	Affect	No/negligible
2	Letzring (2008) [study 2]	138	Affect	Attachment avoidance	Medium
3	Cardy et al. (1986)	66	Affect	Liking	Medium
4	Hartog (1991)	250	Attitude	Attitudes	Not avail. (abstract only)
5	Gibson (2006)	*nr*[c]	Attitude	Life satisfaction	No/negligible
6	Lewis (2002)	149	Attitudes	Expectations	Not avail. (abstract only)
7	Schneider et al. (1953)	400	Cognitive abilities	Academic performance	Large
8	Bayroff et al. (1954)	400	Cognitive abilities	Aptitude	Not avail. (abstract only)
9	Schneider and Bayroff (1953)	400	Cognitive abilities	Aptitude	Large
10	Borman (1979)	146	Cognitive abilities	Attention span	No/negligible
11	Brtek and Motowidlo (2002)	338	Cognitive abilities	Attentiveness	Medium
12	Lewis (2002)	149	Cognitive abilities	Behavior memory	Not avail. (abstract only)
13	Sanchez and De La Torre (1996)	262	Cognitive abilities	Behavior memory	Various (small)
14	Adair (1987) [study 1]	147	Cognitive abilities	Cognitive complexity	Not avail. (abstract only)
15	Adair (1987) [study 2]	*nr*[c]	Cognitive abilities	Cognitive complexity	No/negligible
16	Bernardin et al. (1982)	72	Cognitive abilities	Cognitive complexity	Small
17	Borman (1979)	146	Cognitive abilities	Cognitive complexity	No/negligible
18	Brecker (1988)	122	Cognitive abilities	Cognitive complexity	Not avail. (abstract only)
19	Christiansen et al. (2005)	122	Cognitive abilities	Dispositional reasoning	Large
20	De Kock et al. (2015)	142	Cognitive abilities	Dispositional reasoning	Various (small to medium)
21	Janovics (2003)	410	Cognitive abilities	Dispositional reasoning	Not avail. (abstract only)
22	Powell (2008)	164	Cognitive abilities	Dispositional reasoning	Various (up to medium)
23	Powell and Bourdage (2016)	144	Cognitive abilities	Dispositional reasoning	Various (small to medium)
24	Borman and Hallam (1991)	79	Cognitive abilities	General mental ability	Medium
25	Brecker (1988)	120	Cognitive abilities	General mental ability	Not avail. (abstract only)
26	Christiansen et al. (2005)	122	Cognitive abilities	General mental ability	Small to medium
27	Davis (1999)	82	Cognitive abilities	General mental ability	Not avail. (abstract only)
28	De Kock et al. (2015)	142	Cognitive abilities	General mental ability	Small to medium
29	George (2006)	301	Cognitive abilities	General mental ability	Small (but contingent)
30	Hauenstein and Alexander (1991)	100	Cognitive abilities	General mental ability	Small to medium
31	Janovics (2003)	410	Cognitive abilities	General mental ability	Not avail. (abstract only)
32	Letzring (2008) [study 2]	138	Cognitive abilities	General mental ability	No/negligible
33	Lippa and Dietz (2000)	109	Cognitive abilities	General mental ability	Medium
34	Smither and Reilly (1987)	90	Cognitive abilities	General mental ability	Various (up to medium)
35	Powell (2008)	164	Cognitive abilities	General mental ability	Small (both neg. and pos.)
36	Borman (1979)	146	Cognitive abilities	General mental ability	Medium
37	Borman and Hallam (1991)	79	Cognitive abilities	Spatial reasoning ability	Medium
38	Adair (1987) [study 1]	147	Cognitive style/heuristics	Attribution	Not avail. (abstract only)
39	Johnson (1987)	73	Cognitive style/heuristics	Cognitive modeling	Not avail. (abstract only)
40	Lee (1988)	95	Cognitive style/heuristics	Cognitive style	Medium
41	Willis (1985)	264	Cognitive style/heuristics	Cognitive style	No/negligible
42	Cardy and Kehoe (1984)	359	Cognitive style/heuristics	Field independence	Small-to-medium
43	Clevenger (1991)	*nr*[c]	Cognitive style/heuristics	Field independence	Not avail. (abstract only)
44	Hauenstein and Alexander (1991)	100	Cognitive style/heuristics	Implicit rating theory	Medium
45	Uggerslev et al. (2008)	236	Cognitive style/heuristics	Implicit rating theory	Various (all small)
46	Borman (1979)	146	Cognitive style/heuristics	Problem-solving style	No/negligible
47	George (2006)	301	Cognitive style/heuristics	Prototypes of applicants	Not avail. (abstract only)
48	Borman (1979)	146	Complex task	Base rate estimation	No/negligible
49	Schmid Mast et al. (2011)	131	Complex task	Deception detection task	Small
50	Ambady et al. (1995)	90	Complex task	Decoding skills	Various (small)
51	Ambady et al. (1995)	90	Complex task	Non-verbal sensitivity	Various (small to medium)
52	Borman (1979)	146	Demographic	Age	No/negligible
53	Paquet (2005)	181	Demographic	Culture (Indiv-Collectiv.)	Not avail. (abstract only)
54	Ambady et al. (1995)	90	Demographic	Gender	Small (negative) to medium
55	Carney et al. (2007)	334	Demographic	Gender	Small
56	Chan et al. (2011) [study 1]	898	Demographic	Gender	Not reported
57	Christiansen et al. (2005)	122	Demographic	Gender	No/negligible
58	De Kock et al. (2015)	142	Demographic	Gender	Small
59	Letzring (2008) [study 2]	138	Demographic	Gender	Small to medium
60	Letzring (2010)	80	Demographic	Gender	Small to medium
61	Lippa and Dietz (2000)	109	Demographic	Gender	Various (No/negligible to small)
62	Schmid Mast et al. (2011)	131	Demographic	Gender	Not reported
63	Powell (2008)	164	Demographic	Gender	Small to medium
64	Vogt and Colvin (2003)	102	Demographic	Gender	Medium
65	Schmid Mast et al. (2011)	131	Demographic	Job experience	Small to medium (negative)
66	Kolk et al. (2002)	121	Demographic	Rating experience	Small to medium
67	Borman (1979)	146	Demographic	Rating experience	No/negligible
68	Borman and Hallam (1991)	79	Demographic	Rating experience	Small (negative)

(*continued on next page*)

Table 1 (*continued*)

Nr	Author(s)	*N*	Cluster	Predictor	Effect size[b]
69	Wood and Marshall (2008) [study 1]	194	Demographic	Rating experience	Small
70	Letzring (2010)	80	Demographic	Similarity	Small
71	Borman (1979)	146	Interests	Career interests	Various (no/negligible)
72	Borman (1979)	146	Interests	Social interest	Small to medium (negative)
73	Brtek and Motowidlo (2002)	338	Motivation	Accountability	Various (small to medium)
74	Rosenbaum (1992)	579	Motivation	Accountability	Not avail. (abstract only)
75	Strupeck (2004)	*nr*[c]	Motivation	Accountability	Not avail. (abstract only)
76	Wood and Marshall (2008) [study 1]	194	Motivation	Accountability	Medium
77	Craven (1988)	*nr*[c]	Motivation	Accuracy motivation	Not avail. (abstract only)
78	Salvemini et al. (1993)	108	Motivation	Accuracy motivation	Medium
79	Ispas (2010)	83	Motivation	Accuracy motivation	Small
80	Borman (1979)	146	Motivation	Effort	No/negligible
81	Murphy, Garcia, et al. (1982)	44	Perception	Behavior observation	Small to large
82	Borman (1979)	146	Personality	Aggression	Small
83	Christiansen et al. (2005)	122	Personality	Big 5	Small to medium
84	De Kock et al. (2015)	142	Personality	Big 5	Various (most small)
85	Gibson (2006)	*nr*[c]	Personality	Big 5	Not avail. (abstract only)
86	Janovics (2003)	410	Personality	Big 5	Not avail. (abstract only)
87	Letzring (2008) [study 1]	142	Personality	Big 5	Small to medium
88	Letzring (2008) [study 2]	138	Personality	Big 5	Small to medium
89	Lippa and Dietz (2000)	109	Personality	Big 5	Various (up to small to medium)
90	Powell (2008)	164	Personality	Big 5	Various (small to medium)
91	Borman (1979)	146	Personality	Composite (cluster)	Medium
92	Davis (1999)	82	Personality	Conscientiousness	No/negligible
93	Borman (1979)	146	Personality	Detail orientation	Small to medium
94	Borman and Hallam (1991)	79	Personality	Detail orientation	Small
95	Powell and Bourdage (2016)	144	Personality	Emotionality	Various (most small)
96	Letzring (2008) [study 2]	138	Personality	Interpersonal problems	Small to medium
97	Lippa and Dietz (2000)	109	Personality	Masculinity/femininity	Small
98	Letzring (2008) [study 2]	138	Personality	Narcissism	Small to medium
99	Davis (1999)	82	Personality	Need-to-evaluate	Size not reported[c]
100	Gibson (2006)	*nr*[c]	Personality	Need-to-evaluate	Not avail. (abstract only)
101	Borman and Hallam (1991)	79	Personality	Personal adjustment	Small
102	Vogt and Colvin (2003)	102	Personality	Psychological communion	Medium
103	Letzring (2008) [study 2]	138	Personality	Psychological well-being	Various (small to medium)
104	Human et al. (2011)	380	Personality	Psychological well-being	Various (no estimate avail.)
105	Borman (1979)	146	Personality	Self-control	Small to medium
106	Borman & Hallam, 1991	79	Personality	Self-control	No/negligible
107	Borman (1979)	146	Personality	Self-monitoring	No/negligible
108	Davis (1999)	82	Personality	Self-monitoring	Not avail. (abstract only)
109	Borman (1979)	146	Personality	Sociability	No/negligible
110	Borman (1979)	146	Personality	Tolerance	Small
111	Adams (1927)	80	Personality	Various traits (not Big 5)	Invalid (questionable method used)
112	Ambady et al. (1995)	90	Personality	Various traits (not Big 5)	Various (small to medium)
113	Borman (1979)	146	Personality	Various traits (not Big 5)	Various (up to small to medium)
114	Hjelle (1969)	72	Personality	Various traits (not Big 5)	Various (small to large)
115	Letzring (2008) [study 2]	138	Personality	Various traits (not Big 5)	Small to medium
116	Colman et al. (2017)	1153	Personality	Perspective-taking	Various
117	Colman et al. (2017)	1153	Personality	Empathic concern	Various
118	Colman et al. (2017)	1153	Personality	Fantasy	Various
119	Colman et al. (2017)	1153	Personality	Personal distress	Various
120	Brtek and Motowidlo (2002)	338	Rater behavior	Note frequency	Medium
121	Kolk et al. (2002)	121	Rater behavior	Note taking	Small
122	Middendorf and Macan (2002)	169	Rater behavior	Note taking	N/A
123	Letzring (2008) [study 1]	142	Rater behavior	Social behavior	Various (up to medium)
124	Freeberg (1969)	69	Self/other evaluations	Acquaintance	Not reported
125	Letzring et al. (2006)	180	Self/other evaluations	Acquaintance	Medium to large
126	Connelly et al. (2010) [meta-analysis]	N/A	Self/other evaluations	Acquaintance/intimacy	Various
127	Borman (1979)	146	Self/other evaluations	Assumed similarity	No/negligible
128	Zalesny et al. (1992)	83&116	Self/other evaluations	Teaching attitudes	Various (medium - to large +)
129	Davis (1999)	82	Self-evaluations	Attributional complexity	No/negligible
130	Letzring (2008) [study 2]	138	Self-evaluations	Attributional complexity	Small
131	Borman (1979)	146	Self-evaluations	Empathy	Small to medium
132	Borman and Hallam (1991)	79	Self-evaluations	Evaluative tendency	No/negligible
133	Powell (2008)	164	Self-evaluations	Interpersonal orientation	Small to medium
134	Schmid Mast et al. (2011)	131	Self-evaluations	Rating self-efficacy	No/negligible
135	Schmid Mast et al. (2011)	131	Self-evaluations	Rating self-efficacy	Medium
136	Wood and Marshall (2008) [study 1]	194	Self-evaluations	Rating self-efficacy	Medium to large

(*continued on next page*)

Table 1 (*continued*)

Nr	Author(s)	*N*	Cluster	Predictor	Effect size[b]
137	Borman (1979)	146	Self-evaluations	Self-competence	No/negligible

Notes. The table lists at least $N = 137$ distinguishable individual (or mini-sets of) effects in $k = 48$ reported studies. The actual number of individual effects is substantially larger, as some studies reported only selected results from large numbers of individual differences tested.

[a] These studies do not include work conducted outside of I–O literature. Fringe cases are discussed in our inclusion criteria.

[b] We used Cohen's (1988) guidelines to interpret effect sizes (*r*), i.e. no/trivial (0.00), small (0.10), medium (0.30) and large (0.50) effects. An effect-size interval of 0.05 around these point estimates was applied to cluster effect sizes into a description of magnitude. Effects are positive unless indicated as negative.

[c] Sample size (or other information) is not reported for some studies because it was unavailable (for instance, when results were drawn from dissertation abstracts and the original dissertation could not be sourced). More information on all these studies may be requested from the first author.

Table 2
Meta-analysis of individual difference predictors of judgment accuracy in HRM.

Rater characteristic	*N*	*k*	$\bar{r}$	SD_r	*90% CI*	*80% CV*
Cognitive variables	2789	22	0.24	0.14	0.18, 0.29	0.09, 0.38
Cognitive ability	1645	14	0.18	0.14	0.12, 0.25	0.05, 0.31
Dispositional reasoning	1144	8	0.31	0.11	0.24, 0.38	0.21, 0.41
Personality	5577	41	0.04	0.11	0.01, 0.07	−0.05, 0.14
Extraversion	1087	8	0.02	0.12	−0.06, 0.10	−0.08, 0.12
Agreeableness	1229	9	0.09	0.11	0.02, 0.16	0.00, 0.18
Conscientiousness	1087	8	0.01	0.07	−0.04, 0.05	0.01, 0.01
Emotional stability	1087	8	−0.01	0.08	−0.06, 0.05	−0.01, −0.01
Openness	1087	8	0.10	0.13	0.01, 0.19	−0.03, 0.23

Notes: N = total sample size; k = number of studies included in the analysis; $\bar{r}$ = mean observed correlation (uncorrected for indirect range restriction, unreliability, criterion unreliability); SD_r = observed standard deviation of correlations; 90% CI = 90% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$.

2. Review method

2.1. Literature search

2.1.1. Locating studies

We used four methods to locate relevant studies. First, we conducted a computer search of Web of Science and Dissertation Abstracts to retrieve research studies containing the terms *accuracy*, *assessment centre, HR, interview*, *judgment, rater*, *rating*, *performance*, and *validity*. We filtered the resulting lists according to publication field and research area. The second method was a manual search of major journals within the domain of HRM and industrial and organizational (I—O) psychology, including *Journal of Applied Psychology*, *Personnel Psychology*, *International Journal of Selection and Assessment*, *Human Resource Management Review*, *Human Performance, Human Resource Management Journal,* and others. Third, we retrieved publications within reference lists of seminal accuracy literature published both in journal articles and books. Last, we also trawled the personal research websites of five active accuracy researchers.

2.1.2. Inclusion criteria

To be included in our review, a study had to meet the following three criteria:

1. In terms of independent variable, we retained only studies that included individual difference constructs (i.e., rater characteristics such as demographic variables, personality traits, etc.) as predictors of rating criteria.
2. In terms of dependent variable, studies had to include measures of judges' rating quality. We decided to exclude measures of rating error (such as halo, leniency, etc.) because these indices show little empirical relationships with measures of judgment accuracy and validity. For example, in a meta-analysis by Murphy and Balzer (1989) the average correlation between rating error (various indices) and judgment accuracy indices was a mere 0.05 (see also Kasten & Weintraub, 1999), suggesting that the ability to correctly infer others' characteristics is relatively unrelated to tendencies to show systematic errors in one's evaluations. Therefore, we focused on judgment accuracy and rating validity (construct-related and criterion-related validity), rather than on rating bias, unfair discrimination in ratings, etc.
3. We excluded studies that were not immediately relevant to HRM due to their choice of rating tasks, target dimensions, or experimental stimuli. For example, we discarded studies that used students to judge the sexuality of other students from non-verbal behavior. Other investigations that were not easily generalizable to the HRM context focused on judging moods, emotions or affective states of others in non-work contexts (e.g., Davis & Kraus, 1997; Hall, Goh, Schmid Mast, & Hagedorn, 2015; Letzring,

Table 3
Twenty questions for future research about individual differences in rating quality in HRM.

Category	Characteristic	Research question (RQ)
Cognitive	General intelligence	RQ1. Is the relationship between intelligence and rating accuracy non-linear such that accuracy increases with intelligence, but the slope decreases at high levels of intelligence? RQ2. Do interview structure and situational complexity moderate the effect of intelligence on accuracy? RQ3. Is intelligence more important for accurately judging some dimensions (e.g., personality, interview competencies and assessment centre dimensions) than others? RQ4. Is intelligence more important for judging complex stimuli (e.g. live people) in more complex situations (assessment center tasks where different situation are activated) than less complex stimuli (e.g. videos or 'paper people') in less complex situations (one-on-one interviews)?
	Dispositional reasoning	RQ5. What is the nomological place of dispositional reasoning vis-á-vis emotional and social intelligence, and what is their relative importance in predicting judgment accuracy in HRM? RQ6. Can dispositional reasoning be developed with training and, if so, why and how does training work? RQ7. What is the relative predictiveness of the subcomponents of dispositional reasoning (i.e., induction, extrapolation and contextualization) in different judgment contexts (e.g. interviews, AC tasks, performance appraisal) and for judging different target constructs (e.g., personality, dimensions)?
	Behavior memory	RQ8. What is the comparative validity of impression-memory (i.e., memory of a dispositional or trait inference) versus behavior memory (i.e., memory of an observed behavior) for predicting judgment accuracy?
	Cognitive style/heuristics	RQ9. Do cognitive style and heuristics predict judgment accuracy in HRM ratings?
	Cognitive complexity	RQ10. How do ability-based measures of cognitive complexity predict rating accuracy as compared to self-report measures of cognitive complexity?
	Attributional complexity	RQ11. Does assessors' attributional complexity predict their rating accuracy?
Personality	Personality traits	RQ12. Do rater personality traits moderate the effect of intelligence on rating accuracy?
	Rater behaviors	RQ13. Which rater behaviors are most effective to elicit behavioral cues from targets, and do individual differences in raters' ability to elicit cues predict their rating accuracy? RQ14. Is the increased use of interviewers' behavior prompts in interviews related to higher cue availability and overall rating accuracy?
	Motivation	RQ15. How do assessors' levels of accuracy motivation affect their judgment accuracy? Is this due to enhanced cue attentiveness, better cue utilization, or to both? RQ16. How does motivation to distort affect rating quality in HRM, for example, does it moderate the relationship between judgments and ratings, or does it have a direct effect on ratings?
Specific characteristics	Behavior observation	RQ17. Do innovative measures of behavior observation ability (e.g., signal detection measures; see Lord, 1985) predict rating accuracy in conjunction with measures of behavior memory?
	Personality trait chronic accessibility	RQ18. Does assessors' personality trait chronic accessibility for various Big Five traits predict their trait judgment accuracy?
Context		RQ19. Does rating context influence rating quality, for example, are raters who are accurate in judging other people in performance appraisal ratings, also accurate judges of other people in selection settings (e.g., interviews, ACs, social media judgments)? RQ20. In which contexts and under which conditions can machine-learning replace/complement/supplement raters in making more accurate judgments and ratings?

2015; Murphy & Hall, 2011). We opted to include selected lab studies where personality traits were judged given that reliance on other-ratings of personality is an increasingly popular choice in HRM (Mount, Barrick, & Strauss, 1994; Zimmerman, Triana, & Barrick, 2010).

2.1.3. Study characteristics

A total of 54 studies adhered to all our inclusion criteria.[4] These spanned nine decades (from 1927 to 2017). Each study was coded by the first author on the following dimensions: (1) sample size, (2) type of sample (students, employees, managers, etc.), (3) target dimension/trait, (4) research design, (5) rating quality criterion measure, (6) theoretical framework, and finally, (7) observed effect size. Table 1 shows the studies reviewed in terms of selected categorization variables.

The median sample size for the studies reviewed was 143 participants (M = 215.10; SD = 209.407; Min = 44; Max = 1153). Considering the nature of target traits, study participants were most often required to judge others' job performance (k = 23; 43.4%) or personality (k = 17; 32.1%), whereas only a few studies considered raters' ability to judge interview (k = 4; 7.5%) or assessment center dimension (k = 4; 7.75%) performance. As laboratory studies cast in a HRM setting (k = 43; 81.1%) were a popular choice, only a small proportion (k = 3; 5.7%) consisted of field studies. A relative balance existed between cross-sectional (k = 28; 52.8%) and experimental (k = 24; 45.3%) research designs.

Our review revealed that a wide range of criterion measures were used. A distinction could be made between 'accuracy' criteria

[4] We focused on peer-reviewed research, although we did include both published and unpublished (e.g., dissertations and theses) studies. A few research outputs were duplicates because they were available in both dissertation and journal article format. In such cases, we removed them and retained only the journal article results.

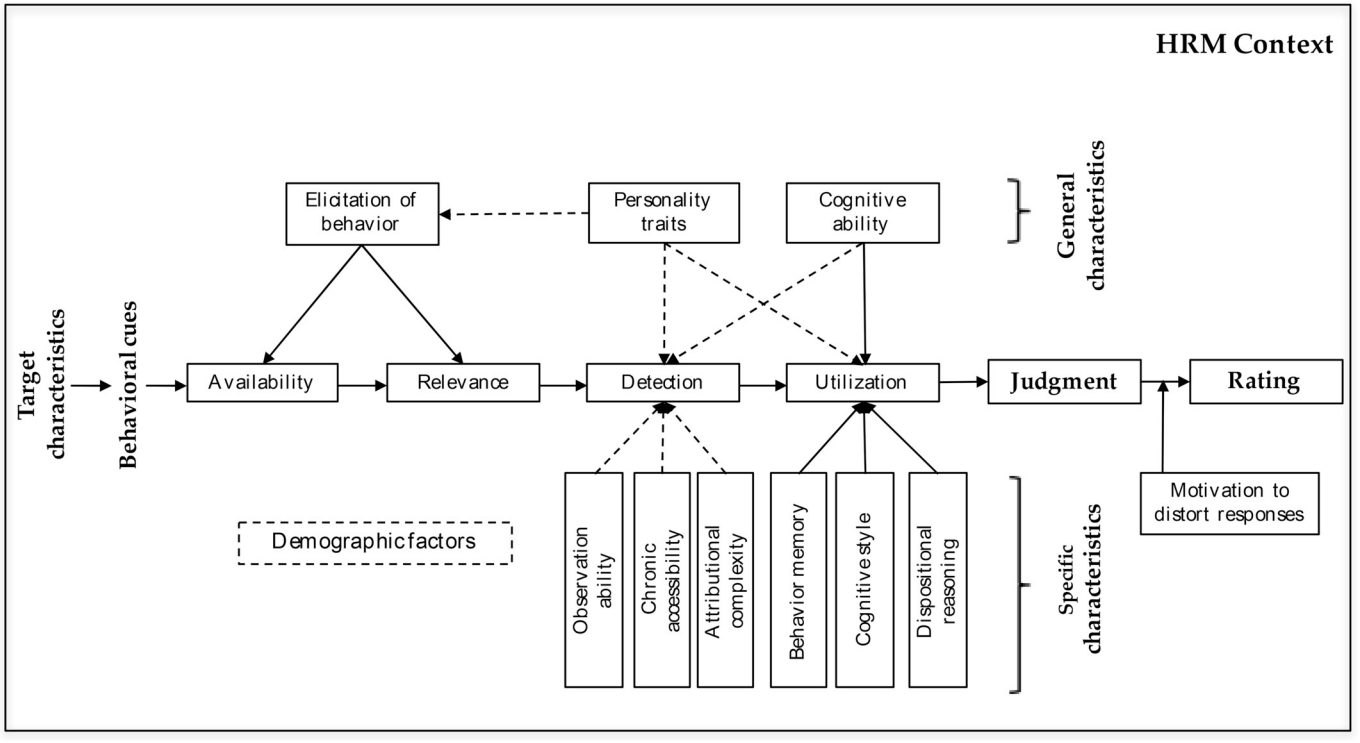

Fig. 1. A model of individual differences in judgment and rating accuracy.

(i.e., correlating judges' dimension or trait ratings of targets with some true score measure/s) and 'validity' criteria (i.e., correlating judges' ratings with an external criterion, such as performance appraisal or training performance ratings). In our review, 81.1% of the studies used accuracy criteria, whereas only 3.8% relied upon validity criteria (in 13.2% the criterion was unclear and 1.9% used both). The most often used accuracy indices were correlational measures (i.e., correlations between a judge's ratings and so-called subjective matter expert derived 'true scores'). Although some studies used multiple correlational measures (k = 10; 18.9%), Cronbach's accuracy indices (Cronbach, 1955) were popular (used in 21 studies; 40.4%), followed by simple profile correlations (with Fisher's r-to-z transformation; used in 18 studies = 34.6%). Borman's Differential Accuracy (Sulsky & Balzer, 1988) was used in only two studies. Aside from correlational measures, five studies (9.6%) employed difference score (e.g., D^2) indices. As only 3.8% used validity criteria, we do not seem to know a lot about the characteristics of judges that provide valid ratings.

Together, the type of judges and rating criteria used in prior research are relevant for our understanding of an accurate judge vs. an accurate rater. For example, the research base covers a broad spectrum of levels-of-acquaintance (from low acquaintance, in interview and AC studies, to high acquaintance, in performance appraisal studies). Moreover, studies of rating quality generally do not distinguish operationally between judgments and ratings. As such, readers should keep in mind the potential limits of generalizing findings across rating contexts (selection vs. appraisal) and rating criteria (judgments vs. ratings).

As guiding theoretical framework for their empirical investigations, an equal number of studies relied on the Realistic Accuracy Model (Funder, 1995) (k = 9; 17%) or cognitive information processing theories (k = 9; 17%). The trend was to adopt RAM in personality judgment studies, whereas information processing approaches were popular in performance appraisal studies. Although it was not always possible to identify the geographical location of the research (as most studies did not reveal this information) our initial data frame of studies (drawn from Web of Science) indicated that the majority were conducted in North America (58.5%) or European countries (37.7%)—only 3.8% of studies were done outside of these territories.

3. Review results

3.1. General characteristics: general intelligence

As judging others is a highly complex task that places a heavy information processing load on the rater (Kolk, Born, Van der Flier, & Olman, 2002; Lance, Foster, Gentry, & Thoresen, 2004) cognitive processing abilities may be an important key to producing accurate judgments (Dipboye, Macan, & Shahani-Denning, 2012; Wyer & Srull, 2014). General intelligence may affect rating quality positively because it enables effective behavior information processing, considered a key process in trait cue utilization (Funder, 1999). In their seminal review of performance rating research, Landy and Farr (1980) conclude, "in general, cognitive characteristics of raters seem to hold the most promise for increased understanding of the rating process"(p. 72). In an early review of studies on cognitive ability and accuracy, Allport (1937) observed, "Experimental studies have found repeatedly that some relationship exists between superior intelligence and the ability to judge others" (p. 514).

Overall, rater general intelligence is the most consistent predictor of rating accuracy (uncorrected[5] validity coefficients = 0.31; Borman, 1979; 0.24; Borman & Hallam, 1991; 0.25; Christiansen et al., 2005; 0.23–0.34; Hauenstein & Alexander, 1991; 0.36; Lippa & Dietz, 2000; average 0.54; Schneider & Bayroff, 1953) of all individual differences we reviewed. That said, effect sizes are often rather modest (e.g. uncorrected $0.10 < r < 0.30$) and some studies (e.g., Letzring, 2008; Powell, 2008) actually found no relationship between intelligence and accuracy.

Although substantial evidence supports the link between cognitive ability and rating accuracy, there are still unresolved questions. First, their relationship may not be as simple as being linear. For example, evidence suggests that general intelligence may be non-linearly related to raters' ability to evaluate others. In their investigation, Smither and Reilly (1987) had ninety subjects rate videotapes of simulated work performances by five ratees and found that the most intelligent raters were generally *less* accurate than moderately intelligent raters. But as expected, moderately intelligent raters were more accurate than the least intelligent raters. Their results suggest that accurate judgment may require a minimal level of information processing capacity, above which the marginal utility of increased intelligence for rating accuracy may dissipate. As such, the idea that 'more is better' may not always be the case.

Second, moderator effects[6] need a closer look. According to our model, the availability of contextual information may inhibit or promote cue detection and cue utilization. We anticipate that boundary conditions, such as interview structure (e.g., George, 2006) or situational characteristics (e.g., Brecker, 1988; Rauthmann, Sherman, & Funder, 2015) may influence the type and richness of cues available to the judge. Logically, the effect of intelligence on accuracy may increase with task complexity. Social cognitive theory suggests that intelligence can be expected to relate stronger to accuracy when it plays a greater substantive role in producing accurate judgments, such as when information processing demands are high (Ambady & Rosenthal, 1992). In this line, Lippa and Dietz (2000, p. 514) state "we suspect that intelligence will prove to correlate even more strongly with judgmental accuracy in studies that ask participants to judge personality from complex, extended information, rather than from 'thin slices' of relatively impoverished video information".

For example, cue-rich situations, as found in high-structure interviews, may place less of a cognitive demand on judges, in comparison to cue-poor situations (e.g., in low structure interviews) where little trait-relevant information is elicited (i.e., they lack

[5] Effects reported are observed correlations and have not been corrected for unreliability, nor for restriction of range, unless stated otherwise.

[6] We wish to acknowledge an anonymous reviewer that noted moderator effects tend to be very small and hard to detect, and that the search for moderators is often fruitless (Murphy & Russell, 2017).

'good information'; Funder, 2012). As such, we expect that intelligence would be a stronger predictor of accuracy in low-structure interviews (or other assessment contexts) as opposed to high-structure interviews. Likewise, in assessment centre judgments, information processing loads are higher than in interviews because multiple candidates are judged simultaneously, often on multiple dimensions, and also in varying situations (Melchers, Kleinmann, & Prinz, 2010; Melchers, Meyer, & Kleinmann, 2008). More complex judgment tasks may increase difficulty of detection and use of multiple cues. Therefore, intelligence may explain accuracy better in high-complexity tasks as compared to low-complexity tasks. Future studies could consider varying task complexity by manipulating aspects of the rating design, such as rating stimuli (e.g. vignettes, videos vs. live people, ordered from less to more complex), or number of targets rated (e.g. single, typical in interviews vs. multiple, typical in assessment centres). Researchers may thus want to explore the intelligence–accuracy link by considering variations of the rating context.

Finally, the effects of intelligence on rating quality in HR-settings might be underestimated. With few exceptions, all the studies reviewed here used college or university students, where restriction of range (see Nunnally & Bernstein, 1994) in ability-based measure scores is typical. So, the overuse of college samples might deflate the observed correlations reported. That is, intelligence may actually predict accuracy in general, non-college populations at a higher level than usually observed in college samples.

3.2. General characteristics: judges' personality, behavior, and motivation

There is a long tradition of studying the effects of broad personality traits as 'main effects' in earlier accuracy research. The conceptual arguments about how personality may affect accuracy (for an overview, see Christiansen et al., 2005; Funder, 1999) can be grouped into three streams, that is, those that consider how personality traits can directly influence perceptual processes, those that consider the actual behaviors of the judge when interacting with targets, and those where rating motivation is important. We discuss these three issues below.

3.2.1. Personality traits

Judges' personality may regulate their social functioning in the workplace, including aspects of interpersonal judgment (e.g., Tziner, Murphy, Cleveland, Yavo, & Hayoon, 2008). In particular, personality traits may affect one's ability to form accurate impressions of others because conceptually they might be linked to the stages of information processing in RAM (i.e., cue detection and cue utilization). For example, agreeable individuals show more concern for others' feelings (Digman, 1990) and should, therefore, be more socially attuned to other individuals with whom they interact. Extraverts are known to seek out social interactions and, because of this increased social exposure, are likely to have more opportunity to hone their interpersonal judgments through practice and feedback (Costa & McCrae, 1992). Conscientiousness manifests itself in greater detail orientation (Goldberg, 1992) generally, but it may also affect how we form impressions about others. For example, highly conscientious judges are likely to be more attentive (than low-conscientious judges) in cue detection, and also show greater consistency in cue utilization. Finally, persons higher in openness are more inquiring and frequently enjoy working with abstract ideas or concepts (Goldberg, 1992) and as such, it is logical to expect that they are also more likely to actively develop mental representations of others' traits and behavior, seek patterns of consistencies and inconsistencies, and form and test hypotheses about others' behavior (see Kihlstrom & Hastie, 1997; Kruglanski & Ajzen, 1983). Openness may also be related to the social information processing preferences of judges because judges' need for cognition has explained accuracy of performance judgments in at least one study (Palmer & Feldman, 2005).

Despite their theoretical relevance to social interaction and judgment accuracy, these hypothesized links between personality and accuracy criteria have received little support (e.g., Borman, 1979; Borman & Hallam, 1991; Hjelle, 1969; Lippa & Dietz, 2000; Powell, 2008; Vogt & Colvin, 2003). Our review shows that empirical studies of personality (for a detailed list, see Table 1) have generally shown null or inconsistent findings. Overall, it appears that the accurate judge most likely does not score higher (or lower) on certain traits. That is, no trait seems to emerge as consistent predictor of accuracy. Even in studies that report 'significant' effects, these tend to be rather small (e.g., observed correlations $0.10 < r < 0.20$). As a case in point, Christiansen et al. (2005) found that, out of the Big Five factors and using three accuracy criterion measures (interview accuracy, acquaintance accuracy, overall accuracy), only openness showed a small to medium effect ($r = 0.23$, $p < .05$) with only one of the accuracy measures, namely interview accuracy. In fact, there are even some traits that may actually be *detrimental* to accuracy, for example being domineering (-0.30), vindictive (-0.27), cold (-0.23), intrusive (-0.20) (all from Letzring, 2008), and showing aggression (-0.17; Borman, 1979). Furthermore, judges who are less sociable may be more accurate (Ambady et al., 1995) compared to sociable individuals.

To tackle these disappointing findings, we suggest future research take the following issues into account. First, we should consider bandwidth issues. Some results suggest that personality traits may be more predictive of accuracy criteria when narrow traits are considered rather than broad traits (De Vries, De Vries, & Born, 2011; Powell & Bourdage, 2016). Second, as the bulk of earlier studies considered so-called 'bright' traits, more work is needed to consider 'dark' traits (Paulhus & Williams, 2002) of raters, for example, as possible derailers of accurate judgment. Third, we urge caution when interpreting the personality–accuracy literature. Some of the studies we surveyed are often plagued by high family-wise error rates as they combine multiple personality traits and behaviors (often > 30) and various operationalizations of accuracy (e.g., by relying on permutations of 'true-score' source, accuracy index, and so forth). As a result, very large correlation matrices may become 'empirical dragnets'. Although exploratory research of this nature is common in early stages of enquiry, they should not be the norm if we wish to build a solid and replicable research base (Asendorpf et al., 2013).

3.2.2. Motivation

Rater motivation can be defined in terms of the basic goals or objectives that drive rating behavior which, in turn, is directed at

observation, storage, recall and integration of targets' behavior (Harris, 1994). Motivation is important because it can impact criterion-related validity of the predictors and the reliability of the ratings (Ispas, 2010). Rater motivation may be caused by raters' perceptions about rewards and reward probability, undesirable consequences, goals, and concerns about what others may think about one's ratings (Cleveland & Murphy, 1992; Harris, 1994; Murphy & Cleveland, 1995, 2004). From this perspective, motivation results from the perceived instrumental value of rating outcomes such as accuracy. According to RAM, when raters are motivated they should produce accurate ratings (Funder, 1999). More specifically, motivation may increase attention to behavior cues (e.g., studying others' behavior closely), a requirement for effective cue detection. Motivation may also encourage raters to assign greater cognitive resources to cue utilization (e.g., thinking deeply about what trait cues mean). This is especially important as rating occurs in a complex social context that is 'cue rich' (Levy & Williams, 2004).

However, some exceptions notwithstanding (e.g., Salvemini et al., 1993), empirical research shows that direct effects of motivation on accuracy are small or negligible (Ispas, 2010). Moreover, rating effort—a manifestation of rating motivation—does not appear to enhance accuracy (e.g., Borman, 1979). More research on motivational issues in rating is needed, though. First, the generalizability of the findings across HRM contexts is still unknown because earlier work was mostly conducted in studies of performance ratings. Therefore, findings may be different for selection and assessment ratings. Second, prior studies did not typically include direct motivation measures, but instead, relied on proxy measures (e.g., accountability; Brtek & Motowidlo, 2002; Mero & Motowidlo, 1995; Mero, Motowidlo, & Anna, 2003). A good place to start better understanding motivational influences on rating quality would be for studies to collect self-report rating motivation measures. To be useful, these measures might include aspects of accuracy motivation, rating effort, and perceptions of rewards, negative consequences, and impression management concerns (see Harris, 1994).

It is also important to note that not all judges are trying to be "good", that is, rating accuracy may not be the primary goal of raters (Spence & Keeping, 2011). Raters may be capable, but unwilling to rate accurately (Banks & Murphy, 1985). That is why a specific motivation, namely motivation to distort may enter the rating process—during the rendering phase, that is, when raters assign a rating on the appraisal form—when their attitudes to performance appraisal (Tziner & Murphy, 1999) or their idiosyncratic goals (Banks & Murphy, 1985) encourage them to assign systematically higher or lower ratings. That is why our model proposes that it is better to conceptualize motivation as a moderator between judgments and ratings, instead of a direct effect.

3.2.3. Specific rater behaviors

Rater behavior is defined as the aggregate of manifest actions to elicit, observe, classify, and evaluate information about targets (e.g., interviewees). Accurate judges are not 'passive perceivers', but actively participate in interpersonal situations when forming impressions of others (Graves, 1993). Therefore, what judges actually *do* (when evaluating other people) may be more important than who they *are* (in terms of general personality traits).

So far, few studies have sought to explore the link between judges' behaviors and their judgment accuracy. It is likely that judges' behaviors affect the availability and relevance of cues. For example, in the personality literature, Letzring (2008) conducted an experimental study using unstructured interactions in triads of previously unacquainted students and found that students' judgment accuracy of their acquaintances was related to their social skills. More specifically, accurate judges emphasized others' accomplishments, engaged in constant eye contact, compared themselves to others, expressed warmth, enjoyed the interaction, displayed ambition, seemed interested, and expressed sympathy. These results imply that judges' social behaviors during interpersonal interactions are important for creating situations within which targets are likely to reveal relevant personality cues (Letzring, 2008). In the HRM domain, interviewer research can potentially benefit from this line of work.

Interestingly, this growing area of research has urged new elements to be introduced into RAM. As interviewers' behavior relates to eliciting cues from targets, by actively taking part in the social interaction, accurate raters may elicit more and better (relevant) cues from those being judged (Lievens, Schollaert, & Keen, 2015). As such, *cue elicitation* should be considered alongside existing judgment processes (i.e., cue detection and cue utilization) in future research, as depicted in our model (Fig. 1).

Overall, studies of the accurate judge should shift their emphasis away from personality traits and towards investigating the actual behaviors of the judge. For example, an unexplored avenue for research lies in judges' use of behavior prompts to actually test or confirm initial impressions of targets. Drawing on Kruglanski's lay epistemic theory of judgment (Kruglanski, 1990), judges may evaluate others through a cyclical process of hypothesis generation and hypothesis testing of an inferred profile of the target. So, an interviewer would use verbal prompts to confirm or disconfirm an initial 'impression hypothesis'. If so, the question then becomes how do raters employ specific behaviors (e.g., verbal and non-verbal) to test these impressions?

In addition, instead of examining judges' traits and behaviors in separation we could view them as related. Given that some personality traits affect preferences for social interactions (Goldberg, 1992), we may develop our ability to read others' behavior when we expose ourselves more to social interaction. For example, some traits (e.g., extraversion) encourage increased social experience, which, in turn, affords the judge the opportunity to develop accuracy faster than judges who have less social interaction. In other words, personality may influence accuracy through the mediating role of social interaction.

3.3. Specific characteristics: dispositional reasoning

Dispositional reasoning is defined as complex knowledge of traits, behaviors and the potential of situations to elicit traits into manifest behaviors (for a recent discussion, see De Kock, Lievens, & Born, 2015). This construct was originally introduced as dispositional intelligence by Christiansen et al. (2005), who defined it as "knowledge of personality and how it manifests in behavior" (p. 139). Dispositional reasoning has three components: *trait induction* (interviewers' understanding of which traits are signalled by

particular behaviors, *trait extrapolation* (the ability to understand how traits co-vary); and *trait contextualization* (an understanding of how situations manifest trait expression in behaviors). Dispositional reasoning is hierarchically ordered, that is, its three facets are influenced by a higher order underlying general construct (De Kock, Lievens, & Born, 2017).

Dispositional reasoning may allow accurate judges to process behavioral information towards accurate trait inferences. In the context of the RAM, induction and extrapolation may facilitate more accurate cue utilization given that judges are able to correctly identify a target's likely trait levels. Further, contextualization is important to make necessary adjustments to trait inferences in light of the situational context within which behaviors are observed. So, it may help to avoid misinterpreting others' actions (e.g., some degree of anxiety is normal in a high-stakes job interview, and most likely does not indicate neuroticism).

Although research on dispositional reasoning is still in its infancy, findings are promising: Christiansen et al. (2005) used a lab study where students ($N = 122$) watched videotaped segments of individuals responding to employment interview questions and judged the personality of the video interviewees. They also rated acquaintances who later completed self-report personality inventories. Results showed that dispositional reasoning was the best predictor of various accuracy indices (with r ranging from 0.41 to 0.52), in fact, better than general mental ability and personality. In a similar study that included a training component, Powell (2008) found that dispositional reasoning correlated with Cronbach's differential accuracy scores in both the control group (0.34) as well as the training group (0.22). A recent partial replication (Powell & Bourdage, 2016) revealed that dispositional reasoning predicted (0.22) students' ability to infer the personality profiles of applicants depicted in video-taped interviews. Delving into the role of its components in judges' ability to produce quality ratings, De Kock et al. (2015) evaluated the three facets of dispositional reasoning as predictors of interviewers' accuracy for judging interview dimensions in high-structure interviews. Results showed evidence of differential prediction of the components: trait extrapolation (0.33), trait contextualization (0.26), and trait induction (0.14). Furthermore, the components incremented general cognitive ability to predict accuracy, indicating that they explain something about accuracy that is not only related to general intelligence. All four of these studies support the view that judges' dispositional reasoning may be an important determinant of people's accuracy of judging others' personality traits. Taken together, as compared to other assessor constructs, dispositional reasoning shows the highest (and most consistent) criterion-related validity (to predict accuracy outcomes).

In light of these findings, future accuracy studies should consider including measures of dispositional reasoning (e.g., the Revised Interpersonal Judgment Inventory; De Kock et al., 2015, 2017). Dispositional reasoning is especially promising to advance our understanding of what makes the an accurate judge given that it may help to uncover not only how judges process cues about targets (their behaviors and traits), but also the *situations* within which behaviors occur, as well as the interaction between persons and situations (De Kock, 2017; Lievens, 2017).

3.4. Specific characteristics: cognitive style and heuristics

Cognitive style is another more specific construct that refers to the unique ways in which raters may perceive or process behavioral stimuli (Witkin, Dyk, Faterson, Goodenough, & Karp, 1962). Earlier research has identified at least three classes of individual differences in cognitive style. First, selective attention is defined as "the ability to separately attend to the features of multidimensional stimuli" (Cardy & Kehoe, 1984; p. 589) and it is often measured with tests of field dependence-independence (e.g., the Hidden Figures Test; Thurstone, 1938). Second, raters may also differ in their cognitive heuristics, such as their implicit theories of performance (Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Hauenstein & Alexander, 1991), which are related to 'personal constructs' (Borman, 1987) that "are likened to performance schemata and 'folk theories' of job performance"(p. 387). Third, the way that raters think about and evaluate the behavior of others— for example, when they employ idiosyncratic decision processes to evaluate information about applicants (Arvey & Campion, 1982; Graves & Karren, 1992; Ostroff & Ilgen, 1992) or differentially weight pieces of information about them (Dougherty, Ebert, & Callender, 1986; Kinicki, Lockwood, Hom, & Griffeth, 1990; Sackett & Hakel, 1979; Zedeck & Kafry, 1977)—may affect accuracy (Brehmer, 1994). Taken together, in the context of the RAM, elements of cognitive style are thought to influence accuracy through their effect on cue detection (e.g., through selective attention ability) and cue utilization (e.g., through differences in how information about others is assimilated on the basis of implicit performance theories or relative cue weighting).

The empirical evidence on cognitive style and heuristics is scant and has become dormant in recent times. Raters high on selective attention ability (Cardy & Kehoe, 1984; Lee, 1988) and those possessing a normative implicit theory of performance (i.e., their beliefs about required behaviors concurred with those contained in formal rating criteria; Hauenstein & Alexander, 1991) tend to be more accurate in their performance ratings. Clearly, we need a more solid empirical research base before firm conclusions may be drawn about the usefulness of cognitive style and heuristics to predict rating quality. One area for future research is to consider how different types of rater groups may be distinguished in terms of their cognitive style and heuristics. For example, managers and psychologists may differ not only in their implicit theories of performance, but also the relative importance they assign to particular behavior cues. This in turn may influence their relative ability to detect and use cues. Managers may have better developed implicit theories of performance, whereas psychologists may have superior abilities to attend to the right behaviour cues, for example. Studies that compare managers and psychologists on cognitive heuristics like selective perception ability and implicit theories of performance might advance our understanding of how cognitive styles and heuristics influence processes required to reach accuracy.

3.5. Specific characteristics: schema complexity

Schema complexity represents a further specific characteristic that may affect rating quality. Cognitive complexity (Bieri, 1955),

defined as “the degree to which a person possesses the ability to perceive behavior in a multidimensional manner”(Schneier, 1977, p. 541) draws upon personal construct theory (Kelly, 1955) and suggests that high-complexity raters prefer to differentiate more between people (and their dimensions) than low-complexity raters. This may facilitate cue utilization as it allows raters to identify the unique characteristics of targets, rather than seeing them as generally alike. Research has mostly failed to support the notion that cognitive complexity may influence accurate judgment (Adair, 1987; Bernardin, Cardy, & Carlyle, 1982; Borman, 1979; Gerber, 2013). However, measurement issues have plagued this line of work (Guion, 2011; Woehr, Miller, & Lane, 1998). For example, cognitive complexity measures are typically based on repertory-grid style measures where raters specify a few people they know well and evaluate these targets on a few dimensions. A ‘complexity’ score captures the degree to which they tend to differentiate between persons (across dimensions) and dimensions (across persons). The resulting indices may not have validity as indicators of *actual* complexity, as they may tap into typical evaluative tendencies, rather than the ability to perceive behavior in a multidimensional manner. As such, ability-based measures of cognitive complexity should be explored in future studies to explain differences in rater accuracy. Similar issues are encountered for a second type of schema complexity, so-called attributional complexity,[7] which is defined as the tendency to engage in complex social information processing and inferential reasoning (Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986). The research base for attributional complexity as a predictor of accuracy is scant. One study (Fletcher, Grigg, & Bull, 1988) found that high-complexity raters made more accurate judgments of traits and attitudes, but others (e.g., Letzring, 2008; $-11 < r < 0.07$) did not replicate these findings (see also Davis, 1999). Similar to cognitive complexity measures, issues of operationalization prevent progress in this area as measures of attributional complexity are not ability-based measures, but rely on self-reports.

3.6. Specific characteristics: chronic accessibility

The chronic accessibility of constructs can be defined as the degree to which individuals differ in the readiness with which particular constructs are utilized in information processing of behavioral stimulus input (Higgins, King, & Mavin, 1982). For example, an interviewer with conscientiousness as a chronically accessible trait would more readily employ it to identify and categorize others' behaviors than using other traits (e.g., extroversion, if it is not equally ‘accessible’).

In the context of the RAM, construct accessibility may influence both the detection and utilization of cues. When raters evaluate others, chronically accessible constructs act as perceptual filters that influence which behavior cues are detected and perceived. As a result, raters are more likely to process and retain behaviors (e.g., which they see in an interview) that are related to their accessible constructs, compared to behaviors that are related to inaccessible constructs (Srull & Wyer, 1979). In other words, construct accessibility affects the storage, encoding and retrieval of behavioral information (Bargh & Thein, 1985; Srull, 1981, 1983). Construct accessibility may affect accuracy because it influences perceptual selection (Higgins, et al., 1982) as individuals with accessible constructs are more sensitive (than individuals with inaccessible constructs) to stimuli associated with those constructs (Bargh & Pratto, 1986).

Overall, chronic accessibility is a relatively unexplored predictor of rating quality in HRM. Woehr (1992) showed that chronically accessible constructs may not only affect the degree to which performance-related dimensions are accessible for use, but also that performance ratings will be more accurate if the performance dimensions are accessible to a rater. Construct accessibility deserves more research attention, especially in the domain of personnel selection. Such investigations hold potential practical benefits. For example, interviewers may be trained to become aware of their chronically inaccessible traits, because these may act as perceptual ‘blind spots’—traits for which interviewers easily fail to detect corresponding behavior cues. As another application, interviewers with a particular chronically accessible trait may be employed as a ‘trait expert’ to rate specific traits in an interview.

3.7. Specific characteristics: behavior observation ability and behavior memory

Behavior observation ability denotes the ability to detect behavior cues as soon as they are emitted, whereas behavior memory refers to the capacity to recall observed behaviors following the rating task. Although most studies of ‘behavior accuracy’ (e.g., Lewis, 2002; Middendorf & Macan, 2002; Murphy & Balzer, 1986; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Sanchez & De La Torre, 1996; Sulsky & Balzer, 1988) seem to confound behavior detection with behavior recall, these are two distinct abilities that each may help to deepen our understanding of judgment processes as proposed by RAM. To be accurate, raters must first detect manifestations of traits (Funder, 1999) (or other target dimensions) to pass this information on to cue utilization functions. Observation of ratees' behavior is the first task in producing judgments about performance (Borman & Hallam, 1991). More specifically, observation ability is similar to cue detection in RAM (Funder, 1995). Some raters may have greater sensitivity to detect verbal and non-verbal stimuli when these occur, whereas other raters may be oblivious to subtle cue signals as they happen.

Behavior observation ability is important because it affects encoding of behaviors—and subsequent storage and recall— into memory. In this way, behavior observation ability largely determines the quality of information cues available to judges (for cue utilization). In turn, behavior memory depends on the effective storage and recall of information about targets (e.g., see person-memory models proposed in Srull & Wyer, 1989). Together, both behavior observation ability and behavior memory may facilitate

[7] Although related, attributional complexity focuses on attributions for others' behavior (e.g., thinking of reasons why a colleague sent a rude email), while cognitive complexity focuses on raters' evaluative tendencies (e.g., a manager distinguishing between target persons on a given dimension).

cue detection and the availability of *good information* to the judge (Funder, 1999).

Surprisingly little research has studied behavior observation ability as a predictor of judgment accuracy, as earlier approaches used measures confounding detection and memory. These studies of 'behavior accuracy' show mixed support as some (e.g., Denisi & Peters, 1996; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Rush, Phillips, & Lord, 1981) found positive effects, whereas others (e.g., Lewis, 2002; Middendorf & Macan, 2002; Murphy & Balzer, 1986; Sanchez & De La Torre, 1996; Sulsky & Balzer, 1988) reported trivial or no effects.

To advance work on behavior observation and memory, we propose two avenues that both seek to refine available measurement approaches. First, more thought should be given in future accuracy studies to use tasks that differentiate between behavior observation and behavior recall (Murphy, Martin, & Garcia, 1982). This is because some raters may be able to detect cues, but fail to remember them, whereas others can remember all cues (but failed to detect every cue objectively presented to them). For instance, pure tests of behavior observation ability may require raters to 'tag' behaviors (e.g., using a clicker) as they occur within a live stream of cues. These may be presented in video stimuli pre-coded by expert raters. Alternatively, raters may be asked to provide verbal protocols (e.g., with 'think-out-loud studies') while observing videotaped or live applicants. On their part, behavior memory tests could consist of showing video-clips of single behavior displays (including verbal and non-verbal content). Following a time delay (e.g., a 30-min delay may be typical in actual interviews) raters are then asked to list the behaviors they are able to recall. Future studies should distinguish between measures that evaluate raters' ability to detect behaviors from those that test their ability to recall having seen these behaviors.

Second, we suggest splitting memory measures based on *content type,* given that memories about people may be about their actual behaviors, and/or their abstract personality traits or dispositions (Srull & Wyer, 1989). By illustration, interviewers may recall things interviewees said (i.e., behavior memories), as well as the impressions about the applicant (i.e., trait or disposition memories) they recall. These memories may not overlap completely. In sum, we look forward to new research examining the differential and incremental validity (to predict rating quality) of the nuanced measures we propose here.

3.8. Demographic characteristics: rater gender

As men and women may differ in their ability to evaluate others, rater gender has been the most often-studied demographic variable predictor of rating quality. Hypotheses about gender differences in accuracy (e.g., that female judges are more accurate) have been driven by findings that show gender disparities in constructs that are thought to facilitate accuracy, for example, interpersonal sensitivity (Hall & Bernieri, 2001), a potentially important component in both cue detection and utilization.

However, research findings are not clear-cut. In some studies, female judges were more accurate than male judges (small-to-medium effects; Ambady et al., 1995; Carney, Colvin, & Hall, 2007; De Kock et al., 2015; Letzring, 2010; Schmid Mast, Bangerter, Bulliard, & Aerni, 2011; Vogt & Colvin, 2003), whereas other studies (e.g., Christiansen et al., 2005) showed no gender differences. Mixed findings in this area are also common (e.g., Chan, Rogers, Parisotto, & Biesanz, 2011; Letzring, 2008).

These inconsistencies may suggest the presence of moderator effects. For example, the trait being rated may moderate the effect of rater gender on accuracy outcomes. Female judges may outperform male judges at rating particular traits (e.g., extraversion and positive affect, Ambady et al., 1995; neuroticism, Lippa & Dietz, 2000; Schmid Mast et al., 2011; vulnerability to stress, Powell, 2008). Second, as women are generally more accurate as judges of non-verbal expressions of emotions (Hall & Schmid Mast, 2008), judgment stimuli may also moderate gender-effects on accuracy.

Future studies should attempt to develop stronger theories and hypotheses for gender-related effects on judgment processes and/or outcomes, rather than relying on surface-level demographic characteristics (e.g., gender) alone. For example, gender-related personality traits (e.g., measures of rater masculinity-femininity) should be explored as predictors of accuracy of certain but not other traits (Lippa & Dietz, 2000). Likewise, to the extent that masculinity-femininity affects sensitivity to others' behaviour cues, this gender-related trait may contribute to the ability to produce quality ratings (Hall & Schmid Mast, 2008).

3.9. Demographic characteristics: rating experience

Drawing from the model of work experience of Quiñones, Ford, and Teachout (1995), we define rating experience as the amount, time, and type of experience in rating other people at varying levels of specificity in the organization (e.g., task dimensions, job dimensions, team dimensions, organizational competencies, etc.). The ability to judge others in day-to-day life, and organizations in particular, should increase with age and experience (Fiske & Taylor, 2013), as repeated 'trial-and-error' in judging others may help us to refine our judgment schemas and heuristics. In the context of the RAM, correct cue utilization might be reinforced every time a judge makes a correct judgment (about somebody else's actions), with the result that schemas used for cue utilization are continuously shaped and refined using the ongoing feedback from observing behavioral outcomes (that eventually follow). In short, rating experience is derived from repeated practice.

Empirical studies showed that accuracy may be higher for judges with more experience (Kolk et al., 2002) but these effects may be rather small (e.g., uncorrected validitiy = 0.18; Wood & Marshall, 2008) or negligible (e.g., Borman, 1979). In one study, observational accuracy was actually lower for judges with more experience (-0.16; Borman & Hallam, 1991). Thus, the link between rater experience and accuracy is therefore not straightforward. In future studies, it may be important to take the type of experience (e.g., rating experience vs. job experience, see Quinones et al.) into account. Most prior studies have examined rating experience, rather than the actual job experience, as predictor of rating quality measures. In terms of RAM, managers with many years of on-the-job experience in a particular functional role (i.e., they have high job tenure) may be better at spotting the right cues (i.e., cue

detection) because they know what to look for, and are more likely to correctly use these cues to make dispositional inferences. In addition, experienced managers/psychologists may have developed relevant tacit knowledge about the most important predictors of performance within a particular job role. When they evaluate candidates against normative assessment criteria (i.e., interview dimensions contained in the rating materials), they are likely to employ their implicit theories developed through experience when evaluating candidates. So, we urge more research to uncover how particular types of experience (e.g., rating vs. job) may enhance different types of judgments (e.g., trait judgments vs. expectations about performance).

3.10. Demographic characteristics: culture/ethnicity

Raters' shared values, beliefs, and norms (i.e., culture) and/or ethnic affiliation may potentially affect rating quality in organizations, although it is an understudied area (e.g., Albright et al., 1997). This is relevant in an increasingly multicultural workplace where managers and employees routinely have to judge others across cultural/ethnic lines.

According to RAM, the effect of culture and ethnicity on accuracy processes—as is the case with other surface-level characteristics, like gender and age—may be mediated by deeper-level constructs (rather than showing direct effects in themselves). For example, collectivism (vs. individualism) is characterized by greater awareness of, and valuing of, relationships and interactions with others, which in the context of the RAM may allow for increased attention to cues emitted by other people. However, collectivism may also reduce accuracy as it implies less awareness of differences between individuals. Finally, dyadic similarity (in terms of culture/ethnicity) between judges and targets may lead to greater familiarity with the target's verbal and non-verbal behaviors, thereby facilitating cue detection and utilization. In fact, when rater-ratee pairs are matched in terms of gender and ethnicity, accuracy of personality judgments may be higher (Letzring, 2010).

Note that the emphasis in research on culture/ethnicity effects (e.g., Ng, Koh, Ang, Kennedy, & Chan, 2011) and dyadic similarity in culture/ethnicity (e.g., Sacco, Scheu, Ryan, & Schmitt, 2003) on rating outcomes has traditionally been on rating bias, with only a few studies that explored their effects on rating accuracy or validity. In one laboratory investigation of cultural factors and accuracy (Paquet, 2005) students evaluated the lecturing skills of teaching assistants. Results showed that students' level of collectivistic orientation was related to lower accuracy. If these findings involving student evaluations at universities generalize to the field of HRM (e.g., in 360-degree evaluations), they may have important consequences for rating quality in organizations with multicultural workforces. As this topic area is ripe for more research, we recommend therefore to also scrutinize the effects on rating accuracy outcomes.

3.11. Demographic characteristics: rater age

Rater age may contribute to rating accuracy processes and outcomes for the same reasons as experience. Given that interpersonal judgment accuracy tends to develop across the lifespan (see Fiske & Taylor, 2013) it is not clear why empirical findings so far show that raters' chronological age may be unrelated to accuracy (e.g., Borman, 1979). The paucity of empirical research in this area means that more research is needed before firm conclusions are possible.

3.12. Other characteristics

Various other individual difference constructs have been explored in rating quality research, although findings are not promising. For example, vocational interests (e.g., Holland's six interest types, 1973) were poor predictors of accuracy criteria in one study (e.g., Borman, 1979). Likewise, findings on rater attitudes (e.g., life satisfaction) produced inconsistent results in two unpublished dissertation studies (Gibson, 2006; Hartog, 1991).

3.13. Meta-analytic summary

We conducted bare-bones psychometric meta-analysis (as described by Schmidt & Hunter, 2014) for effects reported for variables that had a sizable number of samples, namely cognitive variables (general cognitive ability and dispositional reasoning) and Big 5 personality traits. Meta-analytic results are shown in Table 2. Results showed that cognitive variables were more strongly related to rating quality measures ($\bar{r}$ = 0.24, 80% credibility interval [CV] 0.09, 0.38) than Big 5 personality traits ($\bar{r}$ = 0.04, CV -0.05, 0.14). The following rater individual differences showed 90% CI that did not include zero, in order of their criterion-related validity (to predict rating quality criteria): dispositional reasoning ($\bar{r}$ = 0.31), cognitive ability ($\bar{r}$ = 0.18), openness to experience ($\bar{r}$ = 0.10), and agreeableness ($\bar{r}$ = 0.09).

4. Discussion

In light of the pivotal importance of judgments and ratings in traditional and recent HRM areas (interviews, assessment centres and performance evaluations, video resumes, social media evaluations, etc.), a better understanding of the individual difference constructs associated with an accurate judge is needed. Unfortunately, the typical designs used in social psychological research lack the external validity to draw generalizable conclusions. Therefore, one objective was to review the available body of HRM research. We synthesized the literature into a model (see Fig. 1) that answers earlier calls (see Jones & Born, 2008) to explain how assessor constructs facilitate specific judgment processes. Distinguishing features of our model were that (1) it linked rater individual

differences to key judgment processes (namely cue detection and cue utilization) thought to result in accuracy (RAM) (Funder, 1999) and (2) included the notion that due to the HRM context and raters' motivation to distort, their judgements might not always converge with their ratings. This is important because studies of rating quality generally do not distinguish operationally between judgments and ratings. As a second objective, we also aim to promote new research avenues in the field of individual differences related to judging and rating in HRM. In this final section, we therefore outline various implications for future research and propose 20 research questions (see Table 3).

4.1. The importance of cognitive factors

Overall, our meta-analytic estimates show that cognitive factors seem to play an important role in rating quality. For example, the accurate judge is generally more intelligent (than less accurate judges) – one of the more consistent findings in this area of research. Our review shows that effect sizes for cognitive factors are moderate and these appear to be relatively stable in laboratory studies. By virtue of better processing of behavioral information (i.e., encoding, storage, and recall) accurate judges are able to form accurate impressions of targets (e.g., interview applicants, AC candidates, employees).

In addition to having higher levels of general intelligence, accurate judges may also show more developed specific abilities. A growing stream of literature on dispositional reasoning suggests that accurate raters are adept at dealing with social information in particular, that is, they have well-developed schemata about behaviors, underlying traits, and the role of situations in trait expression. In our view, specific abilities of the accurate judge, like dispositional reasoning, hold great potential to help us better understand accuracy.

4.2. The personality paradox

The accurate judge does not seem to score significantly higher or lower than others on any particular personality trait. None of the broad Big Five traits are consistent predictors of accuracy in HRM, according to our review and meta-analytic summary. Effects tend to be trivial in studies where these reach significance and results suggest that, with the exception of openness to experience and agreeableness (which in our meta-analysis both showed 'small' effects; Cohen, 1988), traits are not important to shape rating quality. It seems therefore ironic that personality traits of the judge do not appear to be influential in rating outcomes, but rather, their *understanding* of personality (i.e., in the form of dispositional reasoning) may be important.

Future work on personality predictors should delve deeper into areas where closer conceptual alignment exists between a judge's personality and rating tasks in HRM. For example, narrow traits (as opposed to broad traits) that are socially-oriented may be useful as predictors of accuracy, because narrow traits have the advantage of higher fidelity for predicting closely matched criteria (Soto & John, 2017). In particular, a fruitful approach may be to shift attention away from understanding the role of general traits of the judge in ratings, to explore more closely their specific behaviors when evaluating others in the rating context. For example, how do accurate judges elicit better information from targets in various stages of a selection interview? Furthermore, what do accurate judges do to ensure better detection of cues displayed by the candidate? For example, do they demonstrate various identifiable behaviors not yet studied in earlier research (e.g., showing particular gaze patterns to scan for diagnostic non-verbal behavior cues, listening strategies, and note-taking strategies)?

Another direction consists of investigating how personality may *moderate* the influence of other individual differences constructs on rating quality. For example, Christiansen et al. (2005) demonstrated that interviewers' conscientiousness and agreeableness moderated the relationship between dispositional intelligence and acquaintance accuracy. When interviewers' elevation on these two traits was high, dispositional intelligence predicted acquaintance accuracy better than when elevation on these traits was low. Openness to experience also tends to correlate with measures of cognitive ability (Ackerman & Heggestad, 1997), a characteristic which promotes higher accuracy.

4.3. Motivation to distort and HRM context effects

A further key direction is whether being an accurate judge is a stable individual difference, or whether there are also situational factors at play so that a particular individual might be good in some judgment tasks and bad in others. The context variable in our model accounts for these situational factors and deserves much more attention in future research on the accurate judge. As noted, in some contexts (performance appraisal), accuracy might not be the primary consideration of a judge (supervisor). Accordingly, (s)he might be a good interviewer and assessor but not so in performance appraisals. One reason for these contextual effects might be motivation to distort ratings. For example, supervisors who wish to advance their own political goals may intentionally inflate ratings to ratees under their supervision (Spence & Keeping, 2011). Unfortunately, so far few studies have been conducted about the "motivation to distort" variable. Therefore, more work is needed to determine its prevalence and understand the conditions under which rating distortion occurs.

Note that, so far, social and personality research (in which due to the low stakes context the motivation to distort plays little role) has shown that the ability to judge others' emotions may generalize to the ability to judge others' personality traits (Hall, Gunnery, Letzring, Carney, & Colvin, 2016). Therefore, one area that deserves particular attention is whether or not judges who are good at evaluating people in traditional contexts (e.g., interviews, ACs, performance appraisal) are also good at evaluating other people from social media information.

4.4. Are emotional and social intelligence 'missing in action'?

We encourage more work on other rater characteristics that show strong conceptual overlap with the difficult task of seeking and assimilating behavioral information when evaluating people in HRM. These include emotional intelligence and social intelligence (see Lievens & Chan, 2010). Given that interpersonal interaction in the work environment makes it inherently social, it follows that the ability to interpret social information may facilitate understanding others' behavior as they occur within situational contexts (Lievens, 2017).

Despite their potential importance, so far both constructs have been conspicuously absent from rating quality research. Hence, we call for research into their effects. In addition, we need to explore their discriminant validity[8] (and incremental validity) related to established predictors of rating quality (e.g., dispositional reasoning and general mental ability). Overall, we suggest that these three constructs (dispositional reasoning, emotional intelligence, and general intelligence) should be given attention in combination in future theory development, because their conceptual linkage with the judgment processes in RAM is compelling. Other promising individual differences in the domain of emotional and social intelligence include the tendency to perspective-take and empathic concern (Colman, Letzring, & Biesanz, 2017).

4.5. The judge as an active cue elicitator

Our review has identified potential extensions of Funder's RAM. RAM was built on the basis of models that describe how people perceive physical objects (e.g., Brunswik's Lens model; Brunswik, 1956) and, as such, it implies a view of the judge as a more passive observer, waiting to pick up on behavior signals to be used in impression formation. In contrast, a growing line of research (e.g., Letzring, 2008; Lievens et al., 2015) suggests that when interacting with targets, good interviewers actively *elicit* good behavioral cue information, by encouraging the target (interviewee) to express useful trait-relevant information. To this end, they employ interpersonal skills (for example, active listening or non-verbal communication) to put the interviewee at ease, draw out more information, reflect non-verbal signals, etc. Lievens et al. (2015) demonstrated that high accuracy in assessment center ratings was due to role-players that were effective at both eliciting and evaluating candidate behavior, suggesting that cue elicitation may work in tandem with other judgment processes (e.g., cue detection and cue utilization). These findings suggest the need to include *cue elicitation* in the RAM (as demonstrated in Fig. 1) for enhancing the availability and relevance of cues available to the judge. In future research, we should determine how interviewers and assessors manage the interpersonal interaction (both in the relationship-building and questioning stage of the interview) to elicit useful behavioral data for their judgments. Experimental studies that consider the main and interactive effects of interviewers' cue elicitation, cue detection, and cue utilization (see Funder, 1999) are useful here, given that these processes were mostly treated in isolation in earlier investigations.

4.6. Towards better understanding of cue detection

As the traditional focus has been on how judges use cues, rather than how (and if) they are able to detect them, assessor constructs that may enhance cue detection represent an area that is ripe for more study. In line with the notion of 'garbage in, garbage out', the quality of cue utilization (and resulting judgments) is predetermined by the quality of cues detected by the judge. We expect that judges that are cue sensitive are able to pick up on both verbal (e.g., detecting fine variations in applicants' speech patterns and tone of voice) and non-verbal (e.g., reading micro-expressions on AC candidates' faces) stimuli as soon as they occur. To study these issues in HRM settings, we could draw on other fields where non-verbal cues (and what they may mean about underlying dimensions) are often studied, for example judging emotions and affect (e.g., Davis & Kraus, 1997; Hall, Andrzejewski, & Yopchick, 2009) or judging personality from facial expressions (Borkenau, Brecke, Möttig, & Paelecke, 2009). Conceptually, rater constructs that could support better cue detection may include oral comprehension (e.g., to understand a high load of complex verbal stimuli), conscientiousness (e.g., to remain attentive throughout lengthy interviews), and non-verbal visual sensitivity (e.g., to identify subtle body-language and facial expressions).

4.7. Implications for HRM practice

Our review generates two promising ways to advance rater selection and rater training. First, our results suggest that organizations might consider using cognitive ability measures to select raters (e.g., interviewers, assessors and performance evaluators) because these measures predict rating quality and are therefore 'job-relevant'. In addition, dispositional reasoning shows promising results as single and incremental predictors of accuracy. The validities in some studies approach those for predicting job performance (Schmidt & Hunter, 1998). Therefore, dispositional reasoning and the readily available inventories for measuring it provide organizations with a straightforward approach for rater selection. That said, practical constraints need to be acknowledged. For example, in performance appraisal "accuracy" might not always be the primary objective of supervisors (Spence & Keeping, 2011) and we often cannot choose judges on the basis of these individual differences because there may only be one person (supervisor) able to provide ratings.

[8] As an anonymous reviewer pointed out there is overlap between components of dispositional reasoning and components of emotional intelligence (Mayer, Salovey, & Caruso, 2004).

As a second practical implication, one should consider targeting the constructs that predict rating quality with training. The dominant approach to rater training is frame-of-reference (FOR) training, which seeks to impose a common evaluation standard and reduce rater idiosyncracy by shaping rater schemas about the dimensions and effectiveness levels to be judged (Roch et al., 2012). However, FOR can also potentially be used to develop the dispositional reasoning components. Yet, the viability of this practical implication depends on whether dispositional reasoning components are malleable. This speaks to the key question as to whether or not the accurate judge is born, or made? So far, attempts to enhance one of the components of dispositional reasoning, to understand behavior-trait links ('induction'), have been unsuccessful (Powell & Bourdage, 2016; Powell & Goffin, 2009). So, before trainings in organizations to develop dispositional reasoning can be recommended, evidence is required to show that it can be developed.

Finally, a broader question is whether accuracy should be the ultimate criterion in settings such as performance appraisal, given the role of appraisals in HRM to help organizations improve performance and create and maintain competitive advantage (DeNisi & Pritchard, 2006; Denisi & Sonesh, 2011). Appraisal ratings should also meet other important needs, including to enhance relationships between supervisors and ratees, be acceptable to ratees, and help facilitate better decisions (Schleicher et al., 2018). Although judgment accuracy is not a *sufficient* condition for these additional outcomes, we see it as one of the *necessary* conditions to create the context in which appraisal ratings can better serve their purpose. Along these lines, future research should determine in which stage of our model machine-learning (algorithms) is most useful to avoid potential biases and improve rating quality. If human judges and "machines" are going to work together for rating purposes in the future, we need to examine whether a substitute, complementary or supplementary approach is the best and under which conditions and in which contexts.

5. Conclusion

Through the ebbs and flows in the HR domain over the last century, the question of 'what makes the good judge?' has endured. Our model (portrayed in Fig. 1) integrates important rater individual differences into a framework that explains how these characteristics may drive key judgment processes that influence rating quality. Cognitive factors (general intelligence and dispositional reasoning) related to the accurate judge showed stronger and more consistent relationships with rating accuracy than personality-related factors. Importantly, our review highlights the scarcity of research on HRM context (selection vs. performance appraisal settings) and judges' motivation to distort ratings. We invite HRM researchers and practitioners to join the search for accurate judges because it holds a lot of potential to enhance rating quality via better rater screening and training. It might also pave the way for integrating human and algorithm-based approaches to judging and rating people.

Acknowledgement

This work was supported by the Andrew W. Mellon Foundation, New York, USA, and the National Research Foundation, Pretoria, South Africa.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219–245. https://doi.org/10.1037//0033-2909.121.2.219.

Adair, F. A. (1987). *The effects of ratee job experience, performance variability, and rater cognitive complexity on performance rating accuracy*. Ann Arbor: Louisiana State University and Agricultural & Mechanical College. Retrieved from http://search.proquest.com/docview/303606504?accountid=14500 (Ph.D. ProQuest Dissertations & Theses A&I database) .

Adams, H. F. (1927). The good judge of personality. *The Journal of Abnormal and Social Psychology, 22*, 172–181. https://doi.org/10.1037/h0075237.

Albright, L., Malloy, T. E., Dong, Q., Kenny, D. A., Fang, X., Winquist, L., & Yu, D. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology, 72*, 558–569. https://doi.org/10.1037/0022-3514.72.3.558.

Allport, G. W. (1937). The ability to judge people. In G. W. Allport (Ed.). *Personality: A psychological interpretation* (pp. 499–522). New York: Henry Holt.

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69*, 518–529. https://doi.org/10.1037/0022-3514.69.3.518.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256–274. https://doi.org/10.1037/0033-2909.111.2.256.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281–322. https://doi.org/10.1111/j.1744-6570.1982.tb02197.x.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119. https://doi.org/10.1002/per.1919.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38*, 335–345. https://doi.org/10.1111/j.1744-6570.1985.tb00551.x.

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology, 22*, 293–311. https://doi.org/10.1016/0022-1031(86)90016-8.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology, 49*, 1129–1146. https://doi.org/10.1037/0022-3514.49.5.1129.

Bayroff, A. G., Haggerty, H. R., & Rundquist, E. A. (1954). Validity of ratings as related to rating techniques and conditions. *Personnel Psychology, 7*(1), 93–113. https://doi.org/10.1111/j.1744-6570.1954.tb02262.x.

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? *Journal of Applied Psychology, 67*, 151–160. https://doi.org/10.1037/0021-9010.67.2.151.

Bieri, J. (1955). Cognitive complexity-simplicity and predictive behavior. *Journal of Abnormal and Social Psychology, 51*, 263–268. https://doi.org/10.1037/h0043308.

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality, 43*, 703–706. https://doi.org/10.1016/j.jrp.2009.03.007.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*, 645–657. https://doi.org/10.1037/0022-3514.62.4.645.

Borman, W. C. (1979). Individual difference correlates of rating accuracy using behavior scales. *Applied Psychological Measurement, 3*, 103–115. https://doi.org/10.1177/014662167900300111.
Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307–322. https://doi.org/10.1016/0749-5978(87)90018-5.
Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-differences correlates. *Journal of Applied Psychology, 76*, 11–18. https://doi.org/10.1037/0021-9010.76.1.11.
Brecker, N. (1988). *The effects of rater training, environmental complexity, cognitive complexity and rater intelligence on performance appraisal accuracy.* Ann Arbor: Stevens Institute of Technology (8817333 Ph.D., ProQuest Dissertations & Theses A&I database).
Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137–154. https://doi.org/10.1016/0001-6918(94)90048-5.
Brtek, M. D., & Motowidlo, S. J. (2002). Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology, 87*, 185–191. https://doi.org/10.1037//0021-9010.87.1.185.
Brunswik, E. (1956). *Perception and the representative design of experiments.* Berkeley: University of California Press.
Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*, 958–975. https://doi.org/10.1037/apl0000108.
Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology, 60*, 197–205. https://doi.org/10.1111/j.2044-8325.1987.tb00253.x.
Cardy, R. L., & Kehoe, J. E. (1984). Rater selective attention ability and appraisal effectiveness: The effect of a cognitive style on the accuracy of differentiation among ratees. *Journal of Applied Psychology, 69*, 589–594. https://doi.org/10.1037/0021-9010.69.4.589.
Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004.
Chan, M., Rogers, K. H., Parisotto, K. L., & Biesanz, J. C. (2011). Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality, 45*, 117–120. https://doi.org/10.1016/j.jrp.2010.11.001.
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*, 123–149. https://doi.org/10.1207/s15327043hup1802_2.
Clevenger, J. P.. Field dependence, group composition and rating accuracy in a simulated integration discussion. Ph.D., Colorado State University, Ann Arbor. (1991). Retrieved from http://search.proquest.com/docview/303950924?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. *Research in Personnel and Human Resources Management, 10*, 121–185.
Colman, D. E., Letzring, T. D., & Biesanz, J. C. (2017). Seeing and feeling your way to accurate personality judgments: The moderating role of perceiver empathic tendencies. *Social Psychological and Personality Science, 8*, 1–10. https://doi.org/10.1177/1948550617691097.
Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092–1122. https://doi.org/10.1037/a0021212.
Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Craven, C. L.. Rater accuracy training: An examination of students rating their instructor. 8826002 Ph.D., DePaul University, Ann Arbor. (1988). Retrieved from http://search.proquest.com/docview/303641628?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177–193. https://doi.org/10.1037/h0044919.
Davis, M. E. (1999). *Influence of assessor individual differences on rating errors and rating accuracy in assessment centers.* Ann Arbor: The University of Nebraska - Lincoln. 9952674 Ph.D., Retrieved from http://search.proquest.com/docview/304513121?accountid=14500.
Davis, M. H., & Kraus, L. (1997). Personality and empathic accuracy. In M. Davis, L. Kraus, & W. Ickes (Eds.). *Empathic accuracy* (pp. 144–168). New York/London: Guilford Press.
Davison, H. K., Bing, M. N., Kluemper, D. H., & Roth, P. L. (2016). Social media as a personnel selection and hiring resource: reservations and recommendations. In R. N. Landers, & G. B. Schmidt (Eds.). *Social media in employee selection and recruitment: Theory, practice and current challenges* (pp. 15–42). Springer. https://doi.org/10.1007/978-3-319-29989-1.
De Kock, F. S. (2017). Contextualizing personality judgment: Reading people in (and) their situations. [Peer commentary on “Assessing personality–situation interplay in personnel selection: Toward more integration into personality research,” by F. Lievens]. *European Journal of Personality, 31*, 441–502. https://doi.org/10.1002/per.2119.
De Kock, F. S., Lievens, F., & Born, M. P. (2015). An in-depth look at dispositional reasoning and interviewer accuracy. *Human Performance, 28*, 1–23. https://doi.org/10.1080/08959285.2015.1021046.
De Kock, F. S., Lievens, F., & Born, M. P. (2017). A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups. *International Journal of Selection and Assessment, 25*, 240–252. https://doi.org/10.1111/ijsa.12176.
De Vries, A., De Vries, R. E., & Born, M. P. (2011). Broad versus narrow traits: Conscientiousness and honesty–humility as predictors of academic criteria. *European Journal of Personality, 25*, 336–348. https://doi.org/10.1002/per.795.
Denisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology, 102*, 421–433. https://doi.org/10.1037/apl0000085.
DeNisi, A. S., & Pritchard, R. D. (2006). *Performance Appraisal, Performance Management and Improving Individual Performance: A Motivational Framework. Management and Organization Review. 2*, 253–277. https://doi.org/10.1111/j.1740-8784.2006.00042.x.
Denisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*, 717–737. https://doi.org/10.1037/0021-9010.81.6.717.
Denisi, A. S., & Sonesh, S. (2011). The appraisal and management of performance at work. In S. Zedeck (Vol. Ed.), *Selecting and developing members for the organization: . Vol. 2. APA handbooks in psychology. APA handbook of industrial and organizational psychology* (pp. 255–279). Washington, DC, US: American Psychological Association. https://doi.org/10.1037/12170-009.
Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221.
Dipboye, R. L., Macan, T., & Shahani-Denning, C. (2012). The selection interview from the interviewer and applicant perspectives: Can't have one without the other. In N. Schmitt (Ed.). *The Oxford handbook of personnel assessment and selection* (pp. 323–352). New York, NY: Oxford University Press.
Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*, 9–15. https://doi.org/10.1037/0021-9010.71.1.9.
Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
Fletcher, G. J. O., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology, 51*, 875–884. https://doi.org/10.1037/0022-3514.51.4.875.
Fletcher, G. J. O., Grigg, F., & Bull, V. (1988). The organization and accuracy of personality impressions: Neophytes versus experts in trait attribution. *New Zealand Journal of Psychology, 17*, 68–77. 1989-29364-001.
Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology, 53*, 518–524.
Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670. https://doi.org/10.1037/0033-295X.102.4.652.
Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego: Academic Press.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*, 177–182. https://doi.org/10.1177/0963721412445309.
Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality & Social Psychology, 55*, 149–158. https://doi.org/10.1037//0022-3514.55.1.149.
George, E. (2006). *Interviewer accuracy across levels of structure in the employment interview.* Ann Arbor: Colorado State University. 3226127 Ph.D. Retrieved from http://search.proquest.com/docview/305357339?accountid=14500 (ProQuest Dissertations & Theses A&I database) .
Gerber, T. (2013). *Cognitive complexity in dimensional rating accuracy – A useless concept or poor operationalization?* Maastricht: Maastricht University (Master of Psychology).
Gibson, J. E. M. (2006). *Interpersonal perception: Don't worry, be happy.* Ann Arbor: University of Victoria (Canada). NR27667 Ph.D., Retrieved from http://search.proquest.com/docview/304983390?accountid=14500 (ProQuest Dissertations & Theses A&I database) .
Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*, 26–42. https://doi.org/10.1037/1040-3590.4.1.26.
Graves, L. M. (1993). Sources of individual differences in interviewer effectiveness: A model and implications for future research. *Journal of Organizational Behavior, 14*, 349–370. https://doi.org/10.1002/job.4030140406.
Graves, L. M., & Karren, R. J. (1992). Interviewer decision processes and effectiveness: An experimental policy-capturing investigation. *Personnel Pyschology, 45*, 313–340. https://doi.org/10.1111/j.1744-6570.1992.tb00852.x.
Guion, R. M. (2011). *Assessment, measurement and prediction for personnel decisions* (2nd ed.). New York, NY: Taylor & Francis Group.
Guion, R. M., & Highhouse, S. (2011). *Essentials of personnel assessment and selection.* Malwah, NJ: Routledge.
Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33*, 149–180. https://doi.org/10.1007/s10919-009-0070-5.
Hall, J. A., & Bernieri, F. J. (Eds.). (2001). *Interpersonal sensitivity: Theory and measurement.* Mahwah, NJ: Lawrence Erlbaum Associates.
Hall, J. A., Goh, J. X., Schmid Mast, M., & Hagedorn, C. (2015). Individual differences in accurately judging personality from text. [Advance online publication]. *Journal of Personality*. https://doi.org/10.1111/jopy.12170.
Hall, J. A., Gunnery, S. D., Letzring, T. D., Carney, D. R., & Colvin, C. R. (2016). Accuracy of judging affect and accuracy of judging personality: How and when are they related? *Journal of Personality,* 1–10. https://doi.org/10.1111/jopy.12262.
Hall, J. A., & Schmid Mast, M. (2008). Are women always more interpersonally sensitive than men? Impact of goals and content domain. *Personality and Social Psychology Bulletin, 34*, 144–155. https://doi.org/10.1177/0146167207309192.
Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737–756. https://doi.org/10.1016/0149-2063(94)90028-0.
Hartog, S. B. (1991). *A systematic evaluation of the components of frame-of-reference training and their effects on rating error, accuracy, and individual cognitive processes.* Ann Arbor: City University of New York. 9130321 Ph.D., Retrieved from http://search.proquest.com/docview/303937997?accountid=14500 (ProQuest Dissertations & Theses A&I database) .
Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes, 50*, 300–323. https://doi.org/10.1016/0749-5978(91)90024-N.
Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology, 43*, 35–47. https://doi.org/10.1037/0022-3514.43.1.35.
Hjelle, L. A. (1969). Personality characteristics associated with interpersonal perception accuracy. *Journal of Counseling Psychology, 16*, 579–581. https://doi.org/10.1037/h0028439.
Human, L. J., & Biesanz, J. C. (2011). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology, 100*, 349–364. https://doi.org/10.1037/a0021850.
Ispas, D. (2010). The role of rater motivation in personnel selection validation studies. University of South Florida; Ph.D., Retrieved from http://scholarcommons.usf.edu/etd/3473.
Janovics, J. E.. Knowing thyself: The influence of dispositional intelligence on self-rating accuracy. Ph.D., Central Michigan University, Ann Arbor. (2003). Retrieved from http://search.proquest.com/docview/305221786?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Johnson, R. L., Jr.. The influence of direct versus indirect observation, candidate report format, and assessor training on the accuracy of assessor ratings. 8805581 Ph.D., Old Dominion University, Ann Arbor. (1987). Retrieved from http://search.proquest.com/docview/303636970?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Jones, R. G., & Born, M. P. (2008). Assessor constructs in use as the missing component in validation of assessment center dimensions: A critique and directions for research. *International Journal of Selection and Assessment, 16*, 229–238. https://doi.org/10.1111/j.1468-2389.2008.00429.x.
Kasten, R., & Weintraub, Z. (1999). Rating errors and rating accuracy: A field experiment. *Human Performance, 12*, 137–153. https://doi.org/10.1207/s15327043hup1202_3.
Kelly, G. A. (1955). *The psychology of personality constructs.* New York, NY: Norton.
Kihlstrom, J. F., & Hastie, R. (1997). Mental representations of persons and personality. In S. R. Briggs, R. Hogan, & W. H. Jones (Eds.). *Handbook of personality psychology* (pp. 711–735). San Diego, CA: Academic Press.
Kinicki, A. J., Lockwood, C. A., Hom, P. W., & Griffeth, R. W. (1990). Interviewer predictions of applicant qualifications and interviewer validity: Aggregate and individual analyses. *Journal of Applied Psychology, 75*, 477–486. https://doi.org/10.1037/0021-9010.75.5.477.
Kolk, N. J., Born, M. P., Van der Flier, H., & Olman, J. M. (2002). Assessment center procedures: Cognitive load during the observation phase. *International Journal of Selection and Assessment, 10*, 271–278. https://doi.org/10.1111/1468-2389.00217.
Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological Inquiry, 1*, 181–197. https://doi.org/10.1207/s15327965pli0103_1.
Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology, 13*, 1–44. https://doi.org/10.1002/ejsp.2420130102.
Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89*, 22–35. https://doi.org/10.1037/0021-9010.89.1.22.
Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven Web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods, 21*, 475–492. https://doi.org/10.1037/met0000081.
Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107. https://doi.org/10.1037//0033-2909.87.1.72.
Lee, J. A. (1988). The effects of cognitive style on rating accuracy with an overall rating scale. *Human Performance, 1*, 261–271. https://doi.org/10.1207/s15327043hup0104_3.
Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91*, 111–123. https://doi.org/10.1037/0022-3514.91.1.111.
Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42*, 914–932. https://doi.org/10.1016/j.jrp.2007.12.003.
Letzring, T. D. (2010). The effects of judge-target gender and ethnicity similarity on the accuracy of personality judgments. *Social Psychology, 41*, 42–51. https://doi.org/10.1027/1864-9335/a000007.
Letzring, T. D. (2015). Observer judgmental accuracy of personality: Benefits related to being a good (normative) judge. *Journal of Research in Personality, 54*, 51–60. https://doi.org/10.1016/j.jrp.2014.05.001.
Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*, 881–905. https://doi.org/10.1016/j.jm.2004.06.005.
Lewis, C. F. (2002). *The effects of consensus process expectations and rater training strategies on rater accuracy, interrater agreement, and behavior recall in an assessment center simulation.* Saint Louis, Ann Arbor: University of Missouri. 3037975 Ph.D., Retrieved from http://search.proquest.com/docview/305450431?accountid=14500 (ProQuest Dissertations & Theses A&I database) .
Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255–264.

https://doi.org/10.1037/0021-9010.86.2.255.
Lievens, F. (2017). Assessing personality–situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality, 31*, 424–440. https://doi.org/10.1002/per.2111.
Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. L. Farr, & N. T. Tippins (Eds.). *Handbook of employee selection* (pp. 339–360). New York, NY: Routledge/Taylor and Francis Group.
Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in iassessment center exercises. *Journal of Applied Psychology, 100*, 1169–1188. https://doi.org/10.1037/apl0000004.
Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio, & H. Liao (Eds.). *Research in personnel and human resources management* (pp. 99–152). Bingley: JAI Press.
Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*, 25–43. https://doi.org/10.1023/A:1006610805385.
London, M. (Ed.). (2001). *How people evaluate others in organizations*. Mahwah, N.J: Lawrence Erlbaum.
Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*, 66–71. https://doi.org/10.1037/0021-9010.70.1.66.
Lorenzo, R. V. (1984). Effects of assessorship on manager's proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology, 37*, 617–634. https://doi.org/10.1111/j.1744-6570.1984.tb00529.x.
Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional intelligence: Theory, findings, and implications. *Psychological Inquiry, 15*, 197–215. https://doi.org/10.1207/s15327965pli1503_02.
Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment center candidates on rating quality. *International Journal of Selection and Assessment, 18*, 329–341. https://doi.org/10.1111/j.1468-2389.2010.00516.x.
Melchers, K. G., Meyer, M., & Kleinmann, M. (2008). Cognitive load and rating accuracy during the observation of an assessment center group discussion. *International Journal of Psychology, 43*, 110.
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*, 517–524. https://doi.org/10.1037/0021-9010.80.4.517.
Mero, N. P., Motowidlo, S. J., & Anna, A. L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology, 33*, 2493–2514. https://doi.org/10.1111/j.1559-1816.2003.tb02777.x.
Middendorf, C. H., & Macan, T. H. (2002). Note-taking in the employment interview: Effects on recall and judgments. *Journal of Applied Psychology, 87*, 293–303. https://doi.org/10.1037/0021-9010.87.2.293.
Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology, 79*, 272–280. https://doi.org/10.1037/0021-9010.79.2.272.
Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71*, 39–44. https://doi.org/10.1037/0021-9010.71.1.39.
Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619–624. https://doi.org/10.1037/0021-9010.74.4.619.
Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives.* Thousand Oaks, CA: Sage.
Murphy, K. R., & Cleveland, J. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*, 158–164. https://doi.org/10.1037/0021-9010.89.1.158.
Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*, 320–325. https://doi.org/10.1037/0021-9010.67.3.320.
Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67*, 562–567. https://doi.org/10.1037/0021-9010.67.5.562.
Murphy, N. A., & Hall, J. A. (2011). Intelligence and interpersonal sensitivity: A meta-analysis. *Intelligence, 39*, 54–63. https://doi.org/10.1016/j.intell.2010.10.001.
Ng, K.-Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K.-Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology, 96*, 1033–1044. https://doi.org/10.1037/a0023368.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.
Ostroff, C., & Ilgen, D. (1992). Cognitive categories of raters and rating accuracy. *Journal of Business and Psychology, 7*, 3–26. https://doi.org/10.1007/bf01014340.
Palmer, J. K., & Feldman, J. M. (2005). Accountability and need for cognition effects on contrast, halo, and accuracy in performance ratings. *Journal of Psychology, 139*, 119–137. https://doi.org/10.3200/JRLP.139.2.119-138.
Paquet, S. L. (2005). *A cultural look at performance appraisals: The role of individualism and collectivism in rating accuracy*. Ann Arbor: University of Calgary (Canada). Retrieved from http://search.proquest.com/docview/305029729?accountid=14500 (NR03880 Ph.D., ProQuest Dissertations & Theses A&I database) .
Parsons, C. K., Liden, R. C., & Bauer, T. N. (2001). Person perception in employment interviews. In M. London (Ed.). *How people evaluate others in organizations* (pp. 67–90). Mahwah, N.J: Lawrence Erlbaum.
Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and Psychopathy. *Journal of Research in Personality, 36*, 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6.
Powell, D. M. (2008). *Assessing personality in the employment interview: The impact of rater training and individual differences on rating accuracy.* Ann Arbor: The University of Western Ontario (Canada).
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. https://doi.org/10.1016/j.paid.2016.01.009.
Powell, D. M., & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. *Human Performance, 22*, 450–465. https://doi.org/10.1080/08959280903248450.
Quiñones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology, 48*, 887–910. https://doi.org/10.1111/j.1744-6570.1995.tb01785.x.
Rauthmann, J. F., Sherman, R. A., & Funder, D. (2015). Principles of situations research: Towards a better understanding of psychological situations. *European Journal of Personality, 29*, 363–381. https://doi.org/10.1002/per.1994.
Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*, 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x.
Rosenbaum, A. L.. The effects of personal accountability and severity of rating consequence on evaluative judgments: Implications for performance rating accuracy. 9315130 Ph.D., Texas A&M University, Ann Arbor. (1992). Retrieved from http://search.proquest.com/docview/304018100?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. *Journal of Management, 42*, 269–298. https://doi.org/10.1177/0149206313503018.
Rush, M. C., Phillips, J. S., & Lord, R. G. (1981). Effects of a temporal delay in rating on leader behavior descriptions: A laboratory investigation. *Journal of Applied Psychology, 66*, 442–450.
Sacco, J. M., Scheu, C. R., Ryan, A. M., & Schmitt, N. (2003). An investigation of race and sex similarity effects in interviews: A multilevel approach to relational demography. *Journal of Applied Psychology, 88*, 852. https://doi.org/10.1037/0021-9010.88.5.852.
Sackett, P. R., & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. *Organizational Behavior and Human Performance, 23*, 120–137. https://doi.org/10.1016/0030-5073(79)90051-5.
Sanchez, J. I., & De La Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81*, 3–10. https://doi.org/10.1037/0021-9010.81.1.3.
Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational*

Behavior and Human Decision Processes, 55, 41–60. https://doi.org/10.1006/obhd.1993.1023.
Schleicher, D. J., Baumann, H. M., Sullivan, D. W., Levy, P. E., Hargrove, D. C., & Barros-Rivera, B. A. (2018). Putting the system into performance management systems: A review and agenda for performance management research. *Journal of Management, 44*, 2209–2245. https://doi.org/10.1177/0149206318755303.
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*, 198–208. https://doi.org/10.1111/j.1468-2389.2011.00547.x.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. https://doi.org/10.1037/0033-2909.124.2.262.
Schmidt, F. L., & Hunter, J. E. (2014). *Methods of Meta-Analysis: Correcting error and bias in research findings* (3rd ed.). London: Sage Publications.
Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Sage Publications.
Schmitt, N. W., Arnold, J. D., & Nieminen, L. (2017). Validation strategies for primary studies. *Handbook of employee selection* (pp. 51–71). Taylor and Francis. https://doi.org/10.4324/9780203809808.
Schneider, D. E., & Bayroff, A. G. (1953). The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology, 37*, 278–280. https://doi.org/10.1037/h0062458.
Schneier, C. E. (1977). Operational utility and psychometric characteristics of Behavioral Expectation Scales: A cognitive reinterpretation. *Journal of Applied Psychology, 62*, 541–548. https://doi.org/10.1037/0021-9010.62.5.541.
Smither, J., & Reilly, R. (1987). True intercorrelation among job components, time-delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 40*, 369–391. https://doi.org/10.1016/0749-5978(87)90022-7.
Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*, 117–143. https://doi.org/10.1037/pspp0000096.
Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71*, 299–333. https://doi.org/10.1111/peps.12263.
Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review, 21*, 85–95. https://doi.org/10.1016/j.hrmr.2010.09.013.
Srull, T. K. (1981). Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 440–463. https://doi.org/10.1037/0278-7393.7.6.440.
Srull, T. K. (1983). Organizational and retrieval processes in person memory: An examination of processing objectives, presentation format, and the possible role of self-generated retrieval cues. *Journal of Personality and Social Psychology, 44*, 1157–1170. https://doi.org/10.1037/0022-3514.44.6.1157.
Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660–1672. https://doi.org/10.1037/0022-3514.37.10.1660.
Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96*, 58–83. https://doi.org/10.1037/0033-295x.96.1.58.
Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*, 994–1003. https://doi.org/10.1037/0021-9010.78.6.994.
Stillman, J. A., & Jackson, D. R. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of Occupational and Organizational Psychology, 78*, 581–594. https://doi.org/10.1348/096317905X26147.
Strupeck, S. A.. Assessment center ratings in a social context: The effect of accountability on assessor ratings. 3150401 Ph.D., The University of Tulsa, Ann Arbor. (2004). Retrieved from http://search.proquest.com/docview/305135061?accountid=14500 (ProQuest Dissertations & Theses A&I database).
Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497–506. https://doi.org/10.1037/0021-9010.73.3.497.
Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.
Tziner, A., & Murphy, K. (1999). Additional evidence of attitudinal influences in performance appraisal. *Journal of Business and Psychology, 13*, 407–419. https://doi.org/10.1023/A:1022982501606.
Tziner, A., Murphy, K., Cleveland, J. N., Yavo, A., & Hayoon, E. (2008). A new old question: Do contextual factors relate to rating behavior: An investigation with peer evaluations. *International Journal of Selection and Assessment, 16*, 59–67. https://doi.org/10.1111/j.1468-2389.2008.00409.x.
Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*, 711–719. https://doi.org/10.1037/0021-9010.93.3.711.
Van Iddekinge, C. H., Lanivich, S. E., Roth, P. L., & Junco, E. (2016). Social media for selection? Validity and adverse impact potential of a Facebook-based assessment. *Journal of Management, 42*, 1811–1835. https://doi.org/10.1177/0149206313515524.
Vogt, D. S., & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of Personality, 71*, 267–295. https://doi.org/10.1111/1467-6494.7102005.
Waung, M., Hymes, R. W., & Beatty, J. E. (2014). The effects of video and paper resumes on assessments of personality, applied social skills, mental capability, and resume outcomes. *Basic and Applied Social Psychology, 36*, 238–251. https://doi.org/10.1080/01973533.2014.894477.
Willis, R. P.. Cognitive style, training format, and rating situation as determinants of halo and accuracy in performance ratings (dogmatism). 8527905 Ph.D., University of South Florida, Ann Arbor. (1985). Retrieved from http://search.proquest.com/docview/303387176?accountid=1450 (ProQuest Dissertations & Theses A&I database).
Witkin, H. A., Dyk, R., Faterson, H. F., Goodenough, D. R., & Karp, S. A. (1962). *Psychological differentiation.* New York: Wiley.
Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior, 13*, 357–367. https://doi.org/10.1002/job.4030130404.
Woehr, D. J., Miller, M. J., & Lane, J. A. S. (1998). The development and evaluation of a computer-administered measure of cognitive complexity. *Personality and Individual Differences, 25*, 1037–1049. https://doi.org/10.1016/s0191-8869(98)00068-3.
Wood, R. E., & Marshall, V. (2008). Accuracy and effectiveness in appraisal outcomes: The influence of self-efficacy, personal factors and organisational variables. *Human Resource Management Journal, 18*, 295–313. https://doi.org/10.1111/j.1748-8583.2008.00067.x.
Wyer, R. S., & Srull, T. K. (Vol. Eds.), (2014). *Handbook of social cognition: Basic processes*(2nd ed.). *Vol. 1*. New York, NY: Psychology Press.
Zalesny, M. D., & Highhouse, S. (1992). Accuracy in performance evaluations. *Organizational Behavior and Human Decision Processes, 51*, 22–50.
Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance, 18*, 269–294. https://doi.org/10.1016/0030-5073(77)90031-9.
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*, 361–378. https://doi.org/10.1080/08959285.2010.501049.