François
de Kock

INDIVIDUAL DIFFERENCES IN JUDGMENT ACCURACY IN PERSONNEL SELECTION:

What Makes the 'Good Judge'

Individual Differences in Judgment Accuracy in Personnel Selection

What Makes the 'Good Judge'?

François de Kock

The research presented in this dissertation was supported in part by the Andrew W. Mellon Foundation, National Research Fund (NRF), University of Cape Town Research Office, Police Academy of the South African Police Services, Military Psychological Institute of the South African Military Health Services, and the University of Stellenbosch.

ISBN 978-94-6299-245-0

Cover designed by Linda van Zijp, StudioLIN, Rotterdam
Lay-out by François S. de Kock
Printed by Ridderprint B.V., Ridderkerk

Individual Differences in Judgment Accuracy in Personnel Selection: What Makes the 'Good Judge'?

Individuele verschillen in beoordelaarsnauwkeurigheid
bij personeelsselectie – Wat kenmerkt de 'goede beoordelaar'?

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

donderdag 17 december 2015 om 15:30 uur

door

François Servaas de Kock
geboren te Uitenhage, Zuid-Afrika

Erasmus University Rotterdam

Promotiecommissie

Promotoren:	Prof.dr. M.Ph. Born Prof.dr. F. Lievens
Overige leden:	Prof.dr. A.B. Bakker Prof.dr. R.E. de Vries Prof.dr. H.T. van der Molen

Contents

Chapter 1

Introduction: Individual differences in judgment accuracy in personnel selection – What makes the 'good judge'?

The judgment accuracy of assessors has been an enduring research topic in personnel selection studies. Assessors produce ratings that affect the quality of personnel selection decisions. Although it is well known that assessors differ in judgment accuracy, we do not yet understand why this is so. This dissertation drew on social cognition literature and judgment accuracy models (Funder, 1999) to study assessor constructs that may predict their judgment accuracy in personnel selection. In order to advance contemporary practices designed to select and train assessors, an integrative profile of the 'good judge', informed by empirical evidence, is needed. The dissertation therefore presents four studies – one systematic review and three empirical studies – that investigated individual difference constructs in judgment accuracy within a personnel selection context. First, a systematic review of empirical literature was conducted, which, in addition to determining what we know and do not know about the good judge, identified focal constructs for further empirical research. In the subsequent empirical investigations, the role of specific individual difference constructs in judgment accuracy was explored. The dissertation advances an understanding of how dispositional reasoning (the complex knowledge of traits, behaviors, and situations' potential to elicit traits into manifest behaviors) and personality trait chronic accessibility (the degree to which individuals differ in the readiness with which constructs are utilized in information processing of behavioral stimulus input) may be characteristics of the good judge in personnel selection. The general project goal was to determine the extent to which assessor individual differences are able to explain judgment accuracy in personnel selection ratings.

> *The search for the good judge of personality is the oldest pursuit in the accuracy literature and was nearly its sole concern during the early incarnation from the 1930s…the prey proved to be unexpectedly elusive. Despite the research attention it has received, the good judge is the potential moderator concerning which the accuracy literature has the sparsest data and fewest firm findings to report.*
>
> Funder (1999, p. 142)

Modern personnel selection approaches rely heavily on assessors as judges of applicants' characteristics. Assessors may include interviewers, assessment centre observers, or line managers observing work-sample performances. Assessors are typically called upon to produce subjective ratings of applicants' performances in selection procedures. Eventually, these ratings form the lifeblood of important personnel selection decisions about offers of employment or promotions. Given their centrality in human resource management, it is surprising that our understanding of rater characteristics, which may affect rating quality in personnel selection, has not developed correspondingly to other factors in personnel selection (*cf*. Guion & Gibson, 1988; Hough & Oswald, 2000; Sackett & Lievens, 2008).

Over the last few decades, consistent evidence shows that individual assessors appear to differ in their judgment accuracy (e.g., Borman, Eaton, Bryan, & Rosse, 1983; Dipboye, Gaugler, & Hayes, 1990; Heneman, Schwab, Huett, & Ford, 1975; Kinicki, Lockwood, Hom, & Griffeth, 1990; Pulakos, Schmitt, Whitney, & Smith, 1996; Ryan & Sackett, 1989; Sackett & Wilson, 1982; Schneider & Bayroff, 1953; Van Iddekinge, Sager, Burnfield, & Heffner, 2006; Zedeck, Tziner, & Middlestadt, 1983). However, the reasons for these individual differences in judgment accuracy are not well known. The present dissertation reports on assessor constructs as potential explanatory variables for the variance in judgment accuracy outcomes. In this dissertation, assessors constructs are treated as 'individual difference' variables that "are linked to differences in job [criterion] performance" (Salgado, Viswesvaran, & Ones, 2001, p. 166), referring in this case to assessors' ability to produce accurate ratings in personnel selection. Individual differences between assessors that are relevant and which find easy expression in behavior are abilities (cognitive ability and physical ability), personality (including social skills, emotional intelligence, and dark traits), interests and self-evaluations (Murphy, 2012). This research on which this dissertation reports, employed a broad framework (See Figure 1.1) of psychological individual difference constructs (adapted from Farr & Tippins, 2010) as a lens to study the assessor. More specifically, this dissertation sought to advance understanding of how assessor constructs predict judgment accuracy in personnel selection.

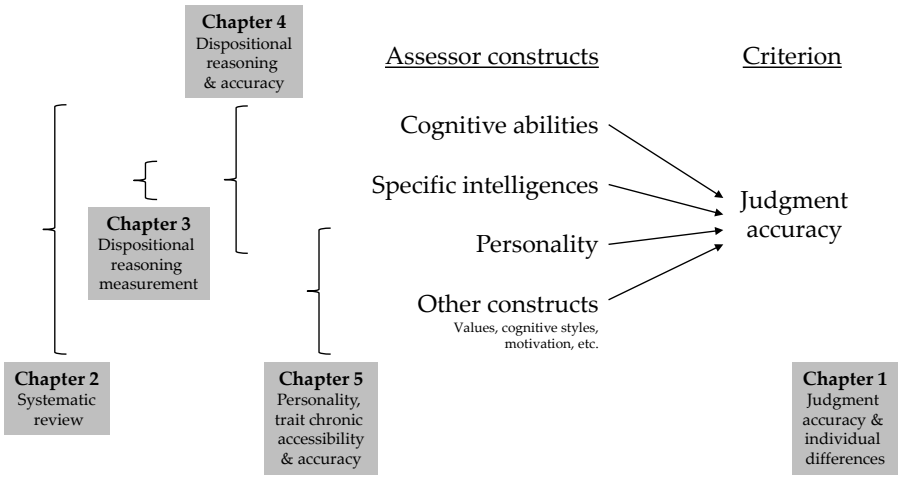

Figure 1.1 Assessor individual differences framework and visualizing the linkages between studies in the present dissertation. Adapted from Farr, J. L., & Tippins, N. T. (Eds.). (2010). *Handbook of employee selection*. New York, NY: Routledge.

Within this broad framework, the study delved into areas that held most potential for advancing understanding of judgment accuracy. The accuracy of judgments in subjective rating methods relies heavily on cognitive processes (DeNisi, Cafferty, & Meglino, 1984; Lance, Foster, Gentry, & Thoresen, 2004). As such, assessor cognitive factors are promising to help understand and improve the rating process (Jones & Born, 2008; Landy & Farr, 1980; Lievens, Tett, & Schleicher, 2009; Nathan & Alexander, 1985). In response, the present study focused on two key assessor constructs[1], namely *dispositional reasoning* (the complex knowledge of traits, behaviors, and situations' potential to elicit traits into manifest behaviors; Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005) and *personality trait chronic accessibility* (the degree to which individuals differ in the readiness with which constructs are utilized in information processing of behavioral stimulus input; Higgins, King, & Mavin, 1982).

In this chapter, first, an overview will be given of the role of assessors in personnel selection. Second, a primer on judgment accuracy will be provided. Finally, assessor characteristics will be introduced as potential individual difference predictors of judgment accuracy in personnel selection. The discussion of these factors identifies several important research questions, outlined in each section.

[1] As far as possible, these focal constructs are studied in conjunction with others (e.g. general mental ability, Big Five personality traits) in the empirical studies contained in the present dissertation.

1.1 The Role of Assessors in Personnel Selection

In personnel selection, subjective rating measures (such as interviews and assessment center evaluations) form the backbone of most personnel selection programmes. In fact, in an earlier survey of assessment practices, individual interviews were indicated as the most frequently used method across 20 countries (Ryan, McFarland, Baron, & Page, 1999). The widespread use of methods that rely on subjective ratings is also unlikely to change soon (Levashina, Hartwell, Morgeson, & Campion, 2014) and, therefore, subjective ratings have been described as "ubiquitous in the workplace" (Guion, 2011, p. 541).

Invariably, assessors' ratings in subjective assessment measures involve some degree of judgment (Gatewood, Feild, & Barrick, 2010). For example, assessors are responsible for observing and evaluating how candidates respond to an interview question or other assessment task and, consequently, they must infer the underlying characteristics (e.g., levels on interview dimensions) of these candidates. As they produce important assessment information about candidates, assessors[2] lie at the heart of subjective rating methods – assessors may be interviewers, trained assessment centre (AC) assessors, or psychologists. Assessors play a major role in the effectiveness of subjective rating methods – typically they have to elicit, evaluate and rate candidate behavior (Dipboye, Macan, & Shahani-Denning, 2012). In fact, many models of rating effectiveness (for example, Dipboye & Macan, 1988; Graves, 1993; Klimoski & Donahue, 2001; Wherry & Bartlett, 1982) place the assessor centre stage.

As such, it is ironic that of the various factors that may affect rating accuracy (for an overview, see Klimoski & Donahue, 2001) the characteristics of the assessor have been least studied. Over the last few decades, a great deal of effort has been spent on studies on rating errors or biases (e.g., Landy & Farr, 1980; Saal, Downey, & Lahey, 1980; Thorndike, 1920; Wherry & Bartlett, 1982) but these have been largely supplanted by later research (e.g., Barrick, Patton, & Haugland, 2000; Borman et al., 2001; Funder, 2012; Nathan & Alexander, 1985; Zalesny & Highhouse, 1992) which focused on judgment accuracy. *Judgment accuracy* is conceptually defined here as "a person's ability, given limited information about a target person, correctly to judge other pertinent characteristics about that person" (Jackson, 1972, p. 185). However, although the characteristics of the 'good judge'[3] (Funder, 1995) have intrigued researchers and practitioners for a long time (e.g., Adams, 1927; Cronbach, 1955; Funder, 2012; Taft, 1955), we still know very little about them. In this regard, Funder (1999) states:

---

[2] In this dissertation, the term 'assessor' is used as a general term for individuals required to observe and evaluate others in personnel selection methods such as interviews and assessment centers.

[3] Any investigation into accuracy of judgment normally requires some people to do the judging – *judges* – and those who are judged – *targets* (Cook, 1979).

> The oldest concern in the history of research on accuracy is the search for the good judge … the kind of individual who truly understands his or her fellow humans." It is not entirely clear whether such a person exists…Evidence of consistent individual differences in accuracy has always been surprisingly difficult to find (p. 138).

It is, therefore, not surprising that human resource management practices today appear to largely ignore the possibility that individual differences exist in judgment accuracy. For example, assessor training and evaluation strategies tend to focus on the assessment process, materials, key dimensions that are being rated (Krause & Thornton, 2009), and establishing a common frame of reference across assessors (Woehr & Huffcutt, 1994). Furthermore, it is also common practice in assessment center operations to combine ratings from different assessors (Gatewood, et al., 2010), reflecting an implicit belief that assessors may be interchangeable, or psychometrically speaking, 'parallel' measures (Nunnally & Bernstein, 1994) of applicants' characteristics. Judging from the available surveys on personnel selection practices used in the field (e.g., Huo, Huang, & Napier, 2002; Ryan, et al., 1999), individual difference constructs that may distinguish assessors in terms of their rating quality do not feature strongly. For example, it is uncommon for interviewers or AC assessors to be screened on any individual difference measures. Moreover, developmental methods used to enhance assessor rating quality (e.g., frame-of-reference training, FOR; Roch, Woehr, Mishra, & Kieszczynska, 2012) also tend to take a 'once-size fits all' approach and individual differences in accuracy and the constructs that may cause them are largely ignored in these training methods.

However, individual differences do exist in the ability of assessors to produce accurate judgments (e.g., Borman, et al., 1983; Dipboye, et al., 1990; Heneman, et al., 1975; Kinicki, et al., 1990; Pulakos, et al., 1996; Ryan & Sackett, 1989; Sackett & Wilson, 1982; Schneider & Bayroff, 1953; Van Iddekinge, et al., 2006; Zedeck, et al., 1983). That is, some assessors consistently outperform other assessors in terms of the quality of their ratings of candidates. As a consequence, it becomes important to understand *why* assessors differ in judgment accuracy in personnel selection methods that rely on subjective ratings. The studies contained in this dissertation explore assessor individual difference constructs as potential explanations for variability in assessor judgment accuracy.

In the remainder of this chapter, we first discuss judgment accuracy (as the main dependent variable under investigation) before, second, assessor constructs are introduced as explanatory variables of accuracy. The latter section also discusses how assessor constructs are studied by the respective chapters contained in this dissertation.

1.2 A Primer on Judgment Accuracy

The discussion of judgment accuracy in the following section comprises of an overview of conceptual and operational definitions of judgment accuracy, provided not only as broad context, but also to underline the important issue of accuracy measurement in this line of research. Next, studies are reviewed to determine whether assessors show variability in their judgment accuracy, before relevant theoretical perspectives are highlighted that may explain these differences.

Judgment Accuracy: A Universal Interest

How accurately people judge others is a problem important for its practical significance and for its theoretical implications (Cronbach, 1955). The study of judgment accuracy has come a long way (for reviews, see Funder, 1999; Jussim, 2005; Kruglanski, 1989; Zaki & Ochsner, 2011) and it has also been explored in a wide variety of disciplines where interpersonal judgments are important, including personality psychology (Funder, 1995, 2001, 2012), social psychology (Asch, 1946; Heider, 1958; Kenny, 2004a; Kenny & Albright, 1987), social-cognition (Fiske & Taylor, 2013), social neuroscience (Zaki & Ochsner, 2011), and clinical psychology (Friedman, Oltmanns, & Turkheimer, 2007). The study of individual differences in judgment accuracy has also been of interest in more specific applications, for example detecting deceit (Bond & DePaulo, 2006, 2008; Porter, Campbell, Stapleton, & Birt, 2002), witness identification (Olsson, 2000), and finally, psychometrics (Cronbach, 1955). Judgment accuracy is important in many different contexts and, therefore, it continues to attract much research interest (e.g., Borkenau, Mosch, Tandler, & Wolf, 2015; Funder, 2012; Zaki & Ochsner, 2011).

Judgment accuracy studies in these different fields use research approaches that reflect their particular focus. For example, in personality psychology studies of judgment accuracy (e.g., Blackman & Funder, 1998; Hall, Goh, Schmid Mast, & Hagedorn, 2015; Letzring, Wells, & Funder, 2006) researchers often employ small-group approaches where research participants (typically college students) rate one another's personalities after a brief interaction. In social psychological studies, the judgment tasks presented to subjects often require them to evaluate the emotions (e.g., Hall & Bernieri, 2001; Murphy & Hall, 2011) or nonverbal behavior (e.g., Ambady, Hallahan, & Rosenthal, 1995; Ambady & Rosenthal, 1992) of people depicted in brief observations of behavior (so-called 'thin-slice' stimuli) such as pictures, videos and audio segments. In some of these studies, experimental tasks occur in a 'zero-acquaintance' context, in which judges (the observers) and targets (those that are observed) are unfamiliar with one another.

In industrial organizational (I-O) psychology studies of judgment accuracy, the experimental methods and tasks used are very diverse. These appear to reflect the relevant rating context, such as performance appraisal, training evaluation or personnel selection. The bulk of rating accuracy research has traditionally been

conducted in the performance rating literature (e.g., Borman, 1977; Heneman, Moore, & Wexley, 1987; Zalesny & Highhouse, 1992). In this field, research participants are often required to rate the actual or simulated performance (in video stimuli, for example) of employees. In personnel selection studies of accuracy, researchers present judgment tasks to research participants that reflect various types of 'target constructs' to be rated, such as interview competencies (e.g., Melchers, Lienhardt, Von Aarburg, & Kleinmann, 2011), assessment centre dimensions (e.g., Lievens, 2001; Melchers, Kleinmann, & Prinz, 2010), or personality traits (Barrick, et al., 2000; Blackman, 2002; Powell & Goffin, 2009; Schmid Mast, Bangerter, Bulliard, & Aerni, 2011; Townsend, Bacigalupi, & Blackman, 2007). In summary, the study of individual differences in judgment accuracy is not only of interest in many areas of psychology, but also across different domains of I-O psychology.

Defining Accuracy

Conceptual Definitions

A number of theoretical perspectives of accuracy in social judgment exist (for reviews, see Funder, 1995; Jussim, 2005; Kruglanski, 1989). As a consequence, a rich diversity of conceptual definitions of judgment accuracy has developed over the years. These are reflected in the use of various terms to denote judgment accuracy, such as rating accuracy (e.g., Borman, 1977; Sulsky & Balzer, 1988; Zalesny & Highhouse, 1992), inferential accuracy (Jackson, 1972), rater validity (Zedeck, et al., 1983), realistic accuracy (Funder, 1995), empathic accuracy (Davis & Kraus, 1997; Hall & Schmid Mast, 2007; Taft, 1966), interpersonal sensitivity (Hall, Andrzejewski, & Yopchick, 2009; Kenny & Winquist, 2001), signal detection (Lord, 1985), to mention only a few. In the I-O psychology literature alone, a confusing array of definitions of 'rating accuracy' exists (for extensive reviews, see Murphy, 1991; Murphy & Balzer, 1989). Although these definitions will not be discussed at length in this dissertation, some brief comments are relevant.

Types of Judgments. Conceptual definitions for accuracy should be distinguished on the basis of the nature of inferences (see Binning & Barrett, 1989) required of the judge – are they about measurement or about prediction? Assessors make *diagnostic judgments* of applicants' characteristics when they infer the degree to which the target (e.g. the interviewee) holds certain characteristics. Diagnostic judgments result in statements such as, "… the applicant is probably a three (out of five) on the dimension of verbal communication". Diagnostic inferences may also differ in depth, that is, 'shallow' inferences may require a judgment of whether a behavior has occurred or not (i.e. behavioral inference). Other 'deeper' inferences – known as dispositional inferences (Trope & Higgins, 1993) – may require a judgment of the underlying trait or dimension that may have caused a behavior. Research shows behavioral and dispositional inferences are separate and complexly interrelated (Trope, Cohen, & Alfieri, 1991), for example, the ability to observe performance and the ability to evaluate performance are different things altogether (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). The second broad class of inferences that assessors

may be required to make in personnel selection are *predictive judgments*, performed when assessors infer or predict applicants' future behavior (e.g. expected future job performance) from measures (which may be diagnostic judgments themselves or scores from standardized tests). Predictive judgments are also known as 'clinical predictions' (Dawes, Faust, & Meehl, 1989; Meehl, 1954).

In light of these distinctions between types of inferences in assessor judgment, the focus of the present study was on diagnostic judgments about others' underlying characteristics, for example, applicants' levels of interview competencies, personality traits, and so forth. The accuracy of dispositional inferences is also known as inferential accuracy. *Inferential accuracy* has earlier been defined as "a person's ability, given limited information about a target person, correctly to judge other pertinent characteristics about that person" (Jackson, 1972, p. 185). When applied to judgments in personnel selection rating devices (e.g. interviews and ACs), inferential accuracy is the ability of an assessor to correctly infer the 'true' characteristics of applicants from behavioral information gathered with personnel selection measures.

Being 'wrong' vs being 'right'. Practitioners and researchers may use assessors' proneness to make errors in their judgments on the one hand and their accuracy on the other, as conceptually related. Practically speaking, it is often thought that being accurate means being free from error, and vice versa. This is not the case, however. On the basis of meta-analytic findings (Murphy & Balzer, 1989), it is clear that rating accuracy is not the opposite thing as rating error. For example, in their meta-analysis, Murphy et al. reported an average correlation of .05 between indices of rating error and rating accuracy (see also Kasten & Weintraub, 1999).

Operational Definitions of Accuracy

The measurement of judgment accuracy with meaningful indicators has been a persistent issue in accuracy research (for a review of performance rating measures of accuracy, for example, see Sulsky & Balzer, 1988). In this regard, Fiske and Taylor (2008, p. 203) state:

> The task of assessing if and when a person's inferences are accurate is more complex than one might suppose. One can determine whether a judgment corresponds to some criterion…; whether it is shared with others…, or whether it is adaptive, pragmatic or useful.

Operational measures of accuracy tend to "describe both the strength and kind of relation between one set of measures and a corresponding set of measures (e.g. true scores) considered to be an accepted standard for comparison" (Guion, 1965, quoted in Sulsky & Balzer, 1988, p. 498). As such, a rich diversity of accuracy measures have been developed in different fields, including 'consensus' and 'agreement' (Funder & Colvin, 1997; Funder & West, 1993).

Early Approaches. Reviews of the development of accuracy measures (e.g.,

Funder, 1987; Funder & Colvin, 1997) show that these early approaches to accuracy measurement often involved simple indices of correlation or agreement between the assessors' judgments of targets' characteristics and the actual profiles of the targets. For example, Cronbach and Gleser (1953) proposed the D-index to assess profile similarity, essentially the sum of the squared deviations of corresponding scores. In their review of the development of accuracy measures in performance rating, Sulsky and Balzer (1988) report that difference scores were also widely used to assess rater judgment accuracy in early research. However, they also point out that simple indices for estimating accuracy were flawed since they collapse potentially meaningful information into a single index, and, for instance, do not take into account the direction of deviations from true scores.

Cronbach Accuracy Measures. Cronbach and colleagues (Cronbach, 1955; Gage & Cronbach, 1955) further criticized simple squared deviation indices of accuracy as confounding various important aspects of accuracy. Cronbach's (1955) seminal contribution represented a watershed moment to accuracy research, however, inadvertently, his critique reached quite the opposite effect: instead of improving accuracy research by providing much-needed improved operationalisations of accuracy, his critique of squared deviation measures stalled accuracy research for the next few decades (Funder, 1995). Cronbach (1955) and colleagues' (Gage & Cronbach, 1955) solution was to decompose accuracy into four separately measurable components. In short, these accuracy measures provided an overall measure of accuracy of a judge's ability to perceive others by averaging his or her squared errors over all items and all targets. This overall score should be broken down into four components, that is, elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA). For the sake of brevity, only differential accuracy (DA) will be discussed[4], as it has particular relevance to studies in the present dissertation. Differential accuracy reflects the ability to detect differences between targets on any item, averaged over items (Cronbach, 1955). As such, it represents the ability of the judge to diagnose the individual target's trait profile (Powell, 2008).

Borman's Differential Accuracy. Borman (1977) identified Cronbach's DA as the most conceptually appropriate component score in performance rating accuracy research because it helps determine how accurately raters can discriminate among people being evaluated on a number of dimensions. Sulsky and Balzer (1988) explain that, to compute Borman's DA index, an assessor's ratings for each dimension are correlated with corresponding true scores across ratees, yielding a DA score per dimension. An overall DA score is then derived by averaging the correlations across dimensions using Fisher's (date) *r*-to-*z* transformation. However, Borman's DA provides only correlational information and the actual distances between a subject's

[4] Interested readers may consult Sulsky and Balzer (1988) for an overview of Cronbach's remaining indices of accuracy.

ratings and true scores are not considered (Becker & Cardy, 1986). Sulsky and Balzer (1988) argued that, conceptually, Borman's DA does not qualify as an index of accuracy because it is insensitive to distances between ratings and true scores. Rather, it should be seen as an index of rater validity that provides important preliminary information for accuracy.

In summary, it is clear that many different operational measures of judgment accuracy exist, each focusing on a unique aspect of judgment accuracy. Different measures do not necessarily co-vary empirically (Roch, et al., 2012; Sulsky & Balzer, 1988) and may imply different aspects of accuracy. As such, they may all be relevant to understand accurate judgment, depending on the need of the researcher. On the basis of earlier recommendations by Borman (1977) and Sulsky and Balzer (1988), the present study will rely on the differential accuracy (DA) criterion measures, although some chapters of this dissertation also report findings on other indices.

True score estimation. A key issue in conversations about accuracy measures is the 'criterion problem' – the lack of a dependable golden standard against which assessors' judgments may be compared (Colvin & Bundick, 2001). For this reason, Funder (1995, 2012) suggests that comprehensive accuracy criteria, called *realistic accuracy*, should be considered in accuracy studies, as they overcome shortcomings of idiosyncrasies associated with specific rater perspectives on the target's behavior, such as peer, self- or subject matter expert (SME) perspectives. Funder suggests that by combining these views through triangulation it is more likely that a true representation of the target's characteristics may be obtained. Realistic accuracy measures are often represented operationally in scores aggregated from various perspectives. However, it stands to reason that realistic accuracy measures would stand and fall to the degree to which there is agreement between various sources of rating of the target's characteristics. As the issue of true score estimation in accuracy research is so important as the basis for calculating accuracy criterion scores (Sulsky & Balzer, 1988) each of the empirical studies presented in the current dissertation report thoroughly on true-score estimation and the determination of accuracy scores.

Are Assessors Accurate?

Despite earlier criticism about people's abilities to make judgments about others (Einhorn & Hogarth, 1978; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954)[5] evidence to the contrary view – that judges can be quite accurate – seems more voluminous and also better substantiated with empirical evidence. Even using so-called 'thin slices' (Ambady, Bernieri, & Richeson, 2000) of expressive behavior, research experiments have shown that people can make remarkably accurate judgments. For example, investigations that utilize so-called 'zero acquaintance'

[5] This line of research appears to investigate a different issue than the issue addressed in the present dissertation, that is, the notion that judges can make 'clinical predictions' (Guion, 2011) from sets of assessment scores about candidates. In the present dissertation, the issue is whether or not assessors can infer the characteristics of targets.

settings have found that judges are able to judge personality and intelligence from limited cues (Borkenau & Liebler, 1993; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004) and also predict others' behavior at greater than chance levels (Levesque & Kenny, 1993). However, these effects increase with acquaintance (Paunonen, 1989) and may depend on the trait being judged (Gangestad, Simpson, DiGeronimo, & Biek, 1992). Even naïve judges can infer others' personality profiles from relatively brief information (Carney, Colvin, & Hall, 2007; Schmid Mast, et al., 2011). An extreme example of subjects' ability to infer others' characteristics from thin-slice stimuli was in a study using extreme rating conditions. Borkenau, Brecke, Möttig, and Paelecke (2009) showed photographs of faces to students as stimuli from which they had to infer the personalities of individuals portrayed in the photographs. Students were able to judge extraversion accurately after receiving only 50 milliseconds of exposure to the photographs. And in a recent study (Borkenau, et al., 2015), strangers were able to judge the personality of targets accurately from only text-based information provided by targets: Targets wrote essays on their hobbies, friends, family and academic studies, and judges had to infer the personality profiles of targets from cues in this text-based information.

The effect sizes for accuracy of interpersonal judgments in accuracy studies are impressive, given the complexity of the judgment task. Whereas most of the studies outlined above used 'above chance' levels as a threshold for inferring accuracy, others have quantified accuracy more clearly. Similarly, in a recent meta-analysis of accuracy studies (k = 263) where Big Five personality traits were judged, Connelly and Ones (2010) found substantial rater accuracy for three different accuracy criteria, namely rater consensus (uncorrected mean r_{rr} = .30 to .45), self–other agreement (corrected $\rho_{o=\infty}$= .71 to .82), and accuracy for predicting job performance (corrected ρ = .31 to .55). In an earlier review of 32 studies in the social and personality psychology literature, consensus correlations ranged from zero to .30 (Kenny, Albright, Malloy, & Kashy, 1994). Taken together, these results support the notion that assessors can be accurate judges of others' characteristics.

Do Assessors Differ in their Judgment Accuracy?

An important assumption of the present study was that assessors do, in fact, differ in their levels of judgment accuracy. If there are no differences, a search for predictors of these differences becomes moot. In this section, we address the question, "Are judgments of some people more valid, as measures of traits or as predictors, than those of others?" (Guion, 2011, p. 391).

Studies of judgment accuracy tend to support the notion that some assessors' judgments are more accurate than those of other assessors (e.g., Biesanz, 2010; Borman, et al., 1983; Dipboye, et al., 1990; Heneman, et al., 1975; Kinicki, et al., 1990; Pulakos, et al., 1996; Ryan & Sackett, 1989; Sackett & Wilson, 1982; Schneider & Bayroff, 1953; Van Iddekinge, et al., 2006; Zedeck, et al., 1983). In laboratory research, some judges appear to be more effective than others. For example, Ambady and

Rosenthal (1992) meta-analyzed (k = 44) the accuracy of judges' predictions of various objective outcomes from brief (< 5 min) observations of expressive behavior, where the overall effect size ($r_{\text{effect size}}$) for prediction accuracy was .39. In a later replication within a workplace setting, (Ambady & Rosenthal, 1993) found in their 'thin-slice' study that student ratings of brief (6 s, 15 s, 30 s) silent videotapes of their college lecturers predicted job performance outcome criteria – end-of-semester teacher evaluations, as well as the principal's ratings of the same teachers – strongly. Various other investigations (e.g., Ambady, Krabbenhoft, & Hogan, 2006; DeGroot & Motowidlo, 1999) report similar findings. In the recent meta-analysis of accuracy studies (k = 263) by Connelly and Ones (2010), where observer ratings of Big Five personality traits were considered, accuracy indices reported in the cited studies varied considerably.

Whereas the aforementioned investigations considered diagnostic judgments (i.e. trait or dimension inferences), individual differences in judges' predictive judgments (i.e. behavioral predictions) have also been found. For example, the results of Zedeck, et al. (1983) demonstrated how 10 interviewers in a military training organization differed in the criterion-related validity of their ratings of N = 121 candidates admitted to officers' training school. In their study, three interviewers' ratings were more predictive of candidates' training evaluation scores after a six-week period, as compared to the remaining interviewers. In a later investigation using structured interviews, Pulakos, Schmitt, Whitney, and Smith (1996) studied behavioral interview ratings by N = 62 interviewers of N = 515 federal agency employees, where each interviewer assessed between 11 and 48 interviewees. Again, the results showed that interviewers differed in their criterion-related validity for predicting subjective ratings of job performance, although the authors were concerned about the effect of sampling error on observed differences in validity between interviewers. In another study of accuracy within a structured interview, Van Iddekinge, Sager, Burnfield, and Heffner (2006) examined differences in criterion-related validity of individual interviewers' (N = 564) ratings of 944 military non-commissioned officers (NCOs) in the US Army. Their results showed considerable variation in coefficients estimating interviewer validity in relation to multiple performance criteria, although these authors were also concerned that sampling error may partially have accounted for these variations.

In summary, individual differences in assessor judgment accuracy are consistently found in most accuracy studies conducted in I-O psychology. To shed light on how individual differences in judgment accuracy may develop, we turn briefly to relevant judgment theories.

Interpersonal Judgment Theories

The process of forming judgments in interviews, for example, is essentially a person perception process (Parsons, Liden, & Bauer, 2001). Interviewers pose questions to interviewees and evaluate the verbal and non-verbal responses, before assigning a

rating that reflects their impression of the interviewee's characteristics. The accuracy of these judgments relies heavily on cognitive processes (DeNisi, et al., 1984; Lance, et al., 2004) and the individual differences that drive them (Jones & Born, 2008).

An impressive array of theories has been developed to explain how individuals form impressions of others and the factors affecting the accuracy of interpersonal judgments. These theories generally fall in the domain of social cognition – defined as the medium through which our worlds are construed and actions initiated (Fiske & Macrae, 2012). Although various judgment theories[6] (for a review, see Fiske & Taylor, 2013; Wyer & Srull, 2014) and judgment process models (for example, Biesanz, 2010; Kenny, 2004b) exist to explain judgment accuracy processes and outcomes, amongst other things, social judgment theory (SJT) (Brehmer, 1988) has laid the foundation for the study of accuracy of person perception. Social judgment theory is described (Brehmer, 1988, p. 13) as:

> … the result of a systematic application of Brunswik's probabilistic functionalism to the problem of human judgment in social situations. Brunswik's theory of perception is also called "cue theory". According to such a theory, a person does not have access to any direct information about the objects in the environment. Instead, perception is seen as an indirect process, mediated by a set of proximal cues. In accordance with this view, SJT defines judgment as a process which involves the integration of information from a set of cues into a judgment about some distal state of affairs.

Realistic Accuracy Model (RAM)

The most influential manifestation of social judgment theory in personnel selection research has been the realistic accuracy model (Funder, 1995, 1999, 2012). On the basis of earlier 'cue' models of perception (e.g., Brunswik, 1956), RAM proposes that the path to judgment accuracy involves a number of interdependent steps (see Figure 1.2). In this framework, an interviewer's judgment accuracy results from a social cognitive process that proceeds in four stages (from left to right in the diagram). The behavior displayed by a target (interviewee) serves as a cue and must be *relevant* to the trait being judged, in a manner where this information is *available* to the interviewer, who must then *detect*; and correctly *utilize* the information to form an accurate judgment. Only when all four steps have been achieved effectively – these processes are interdependent – can a perceiver correctly judge another person's characteristics (Funder, 2012).

[6] A selection of specific theories will be introduced and reviewed in greater detail within the respective chapters of this dissertation.

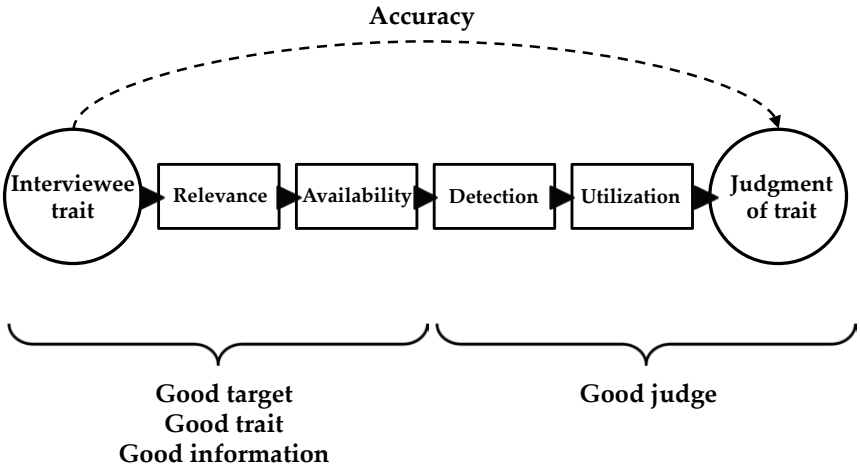

Figure 1.2 The Realistic Accuracy Model applied in interview judgments: Processes and moderators. Adapted with permission from Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, p. 659.

By implication, cue detection and cue utilization may or may not work effectively. RAM further proposes that the degree to which trait or dimension judgments are accurate is moderated by a number of key factors: good targets, good traits, good information, and finally, good judges (Funder, 2012). *Good targets* are simply highly judgable people – their behavior is relevant to their underlying personalities and they may be more transparent than poor targets. *Good traits* (e.g. expressiveness) are more visible than others (e.g. deceptiveness) and therefore, are more easily judged. *Good information* is a function of both quantity (e.g. a one-hour interview provides more trait cues than a speed-dating interview) and quality (e.g. when a person is relaxed and responds to good interview questions, higher-quality cue information results). Finally, *good judges* are better able to detect and use behavior cues to form an accurate personality trait inference.

The RAM is an important framework that highlights how 'good judges' (in addition to good targets, good traits and good information) can be moderators of accuracy. Naturally, perceivers have to bring together all the relevant information about the ratee and other factors (for example, about the situational context, ratee characteristics to be rated, rating procedure, and rater characteristics; Guion, 2011) and form an overall judgment. Although the RAM framework explains how accurate judgments are formed, and also addresses potential moderators of accuracy, it does not provide much detail on the characteristics of the good judge[7]. That is, more work

[7] Funder (1999) suggests a cursory selection of individual difference constructs that may enhance personality judgment accuracy, but these individual differences are not fully representative of the literature and their empirical support is not fully assessed. In addition, they are also not linked to judgment accuracy in the human resource management (HRM) literature.

is needed to determine the individual differences of assessors that may facilitate important judgment processes.

1.3 Individual Differences Constructs in Judgment Accuracy

There are likely to be identifiable assessor constructs that may help explain their judgment accuracy (Guion, 2011). By implication, an individual differences approach is needed to uncover assessor constructs that may affect rating accuracy in personnel selection (for example, see discussion in Jones & Born, 2008). Our main research question, therefore, is:

> To what extent do assessor constructs explain differences in judgment accuracy in subjective rating selection measures?

In addition to its theoretical relevance, a better grasp of the assessor constructs that influence accuracy could help us to select and train the most effective interviewers or assessors. For example, interviewers may be screened with measures that tap into constructs associated with rating accuracy. Or they may receive training to develop specific individual differences that may enhance accuracy. So, research that explores the individual differences in judgment accuracy may have potential practical implications for personnel selection.

The Profile of the Good Judge in HRM: A Systematic Review

As basis for this dissertation, **Chapter 2** reviews empirical literature in HRM that have addressed the characteristics of a good judge. As a consequence of its breadth, RAM does not focus on assessor individual differences that may facilitate judgment processes. Although Funder (1999) suggested a few characteristics of the good judge, a better understanding is needed of what these are. Furthermore, we do not know how assessor constructs (for example, see discussion in Jones & Born, 2008) may affect judgment processes that lead to accuracy – it is unclear from the HRM research base how various assessor individual differences facilitate judgment processes. Stated otherwise, we do not know the link between individual differences and cue detection and cue utilization in the HRM domain. This might be another reason why personality and social psychological studies mostly failed to identify characteristics associated with being a good judge.

Over more than 60 years in HRM literature, we have seen a steady flow of empirical studies in human resource management that tested individual difference constructs as predictors of accuracy (e.g., Borman, 1979; Christiansen, et al., 2005; Powell & Goffin, 2009). However, this work has not been synthesized into a profile of the good judge that is supported by empirical evidence. As such, we are not sure what we know, or do not know, about the individual difference constructs of accurate assessors in human resource management (Guion, 2011). Furthermore, a

better understanding is needed of how assessor individual differences can be linked to judgment processes outlined in RAM.

In order to provide a synthesis of empirical research studies, Chapter 2 will construct an evidence-based profile of the 'good judge' in the specific HRM context. Our overview draws from three primary HRM fields: performance appraisal, interviews, and assessment centres (ACs).

By weighing the evidence in support of individual difference predictors of accuracy in **Chapter 2**, we were able to outline what we know, and do not yet know, about the characteristics of the good judge in a HRM context. In this review, the overall aim is to crystallize the evidence pertaining to the link between various individual differences and accuracy and to identify directions for future research. A key contribution of this study will be to link individual difference constructs to processes thought to influence judgment accuracy. Funder's RAM (1999, 2012) proposes that the ability to detect and use behavior cues may influence the quality of the judges' impression. As such, our review develops a conceptual model that portrays empirical support for (and gaps between) individual difference constructs and key judgment processes.

Consequently, the following research question is formulated:

> *Research Question 1:* From the empirical HRM literature, which individual differences explain judgment accuracy, in other words, what is the profile of the good judge?

The various individual difference constructs that are investigated in the three empirical studies of the present dissertation are now discussed in more detail.

The Internal Factor Structure of Dispositional Reasoning

As a result of the conceptual review conducted in Chapter 2, **Chapter 3** reports the first of three empirical studies that explored specific individual difference predictors of judgment accuracy. Of the many individual difference constructs that may predict accuracy, the most promising factors are likely to be cognitive characteristics (Guion, 2011). Indeed, a contemporary review of general interview literature (Dipboye, et al., 2012) concludes that good interviewers have higher general cognitive ability relative to poor interviewers. In recent years, the focus of individual differences research has shifted toward exploring specific abilities related to being a good judge (for example, see Letzring, 2008; McLarney-Vesotski, Bernieri, & Rempala, 2011; Powell, 2008).

One of these specific abilities, dispositional reasoning, is the focus of both Chapters 3 and 4. *Dispositional reasoning* is defined as the complex knowledge of traits, behaviors, and situations' potential to elicit traits into manifest behaviors (Christiansen, et al., 2005). Earlier research findings suggest that dispositional

reasoning may be a key individual difference construct to explain why judges differ in accuracy. For example, Christiansen et al. (2005) asked students (N = 122) to watch videotaped segments of individuals responding to employment interview questions, after which students also judged the personality of the video interviewees and rated acquaintances who later completed self-report personality inventories. Among a set of constructs that included general cognitive ability and personality, dispositional reasoning was the best predictor of interview accuracy (r = .42). The ability of dispositional reasoning to predict interview-related accuracy was partially replicated in a follow-up study (Powell & Goffin, 2009). In summary, these findings suggest that dispositional reasoning seems to facilitate interviewer accuracy.

Dispositional reasoning was initially conceptualized as a broad set of conceptually distinguishable components, namely behavior-trait knowledge, judges' implicit personality theories, and judges' understanding of situation-trait relevance (Christiansen, et al., 2005). In the present dissertation, these components are labelled as 'trait induction', 'trait extrapolation', and 'trait contextualization', respectively. As studies up to now (e.g., Christiansen, et al., 2005; Powell & Goffin, 2009) have not been able to measure these components reliably, they collapsed scores from conceptually distinct clusters of items into a broad dispositional reasoning measure. In doing so, the underlying facets of dispositional reasoning lie obscured from measurement and further use in research and practice. **Chapter 3** takes a closer look at the internal factor structure of dispositional reasoning. In the current investigation, a revised measure of the interpersonal judgment inventory (IJI) (Christiansen, et al., 2005) was developed[8] to yield reliable and measurement valid component scores. In doing so, the study sought further evidence of construct validity of the measure in the form of internal validity.

Finally, we were also interested in the usefulness of the revised measure in two different populations of assessors, namely managers and psychology students. These populations were selected as they typically receive assessor training in personnel selection (Krause & Thornton, 2009). Users of the dispositional reasoning measure may want to administer the measure in different populations of assessors. Therefore, the issue of measurement invariance needs to be addressed in order to determine whether "an assessment instrument is measuring the same constructs in exactly the same way across groups" (Byrne & Stewart, 2006, p. 287). The analysis of measurement invariance has become popular in HRM (Schmitt & Kuljanin, 2008) as a means to establish measurement properties of a measure when it is used between two or more groups. If the issue of measurement (Millsap, 2011) equivalence between managers and students is not addressed, between-group comparisons of test scores may be misleading, because we would not be sure if observed group differences are 'real' or confounded with differences in the structure of the

[8] The development of the revised interpersonal judgment inventory (RIJI) (De Kock, Lievens, & Born, 2015) is described at length in Chapter 4. Chapter 3 presents an analysis of its measurement validity.

constructs and/or functioning of the measurement scales (Cheung, 2008). As such, the present study determines to what extent does the revised dispositional reasoning instrument measure the same constructs for managers compared to students?

Information on the measurement properties of the revised interpersonal judgment inventory can be useful for researchers and practitioners in human resource management in different ways. It may help to support the use of the measure of dispositional reasoning in judgment accuracy research studies – a reliable and valid measure that provides both overall as well as component-level scores. In practice, the revised measure may be used to establish the profile of assessors on overall- and component-level scores of dispositional reasoning. For example, interviewers and assessors can be screened on the measure, or feedback by targeting specific assessor constructs may be given to assessors to develop them in rater training (e.g., Lievens, 2001; Roch, et al., 2012).

The research question addressed in **Chapter 3**, therefore, is as follows:

Research Question 2: How can dispositional reasoning be measured reliably and with measurement validity at the component level?

a) Can the components of dispositional reasoning be measured reliably with a revised measure, the RIIJ?
b) Can the components of dispositional reasoning be measured validly with a revised measure, the RIIJ?
c) Which factor structure best represents a dispositional reasoning construct, among a choice of a global factor, a component-only (first-level) model, or a hierarchical model (where a second-order global factor influences component scores at first level)?
d) To what extent does the revised dispositional reasoning instrument measure the same constructs for managers compared to students?

To address these questions, **Chapter 3** reports on a factor analysis of dispositional reasoning scores in two distinguishable samples of respondents, namely managers and psychology students. We tested alternative hypothesized factor structures for dispositional reasoning. This allowed us to accomplish three important objectives that all focus on assessing internal measurement properties:

- First, we tried to shed light on the internal composition of the dispositional reasoning construct, that is, is it a single, broad ability, or many specific abilities that are related? Or, could it be considered both, in other words, would a hierarchical factor structure make more sense, where a broad dispositional reasoning ability influences specific components?
- Our second objective was to make practical recommendations about the best way to use the dispositional reasoning measure: For example, should it be used as an overall measure, or rather, can the individual subtests be used to assess the subcomponents reliably and validly?

- Our third objective addressed generalizability issues: we compared the factor structures underlying our dispositional reasoning measure between different assessor types (managers vs psychology students). It was important to compare the construct validity of dispositional reasoning between these groups, as they both provide pools of assessors that are often trained to provide ratings of applicants in workplace assessments (Krause & Thornton, 2009).

An In-depth Look at Dispositional Reasoning and Interviewer Accuracy

As opposed to the previous chapter that focused on internal construct validity issues, Chapter 4 focuses on external construct validity and criterion-related evidence. On the basis of theories that suggest that judges' interpretation of behaviors, traits, and situations are intertwined (e.g., Trope, 1986), Christiansen, Wolcott-Burnam, Janovics, Burns, and Quirk (2005) introduced dispositional intelligence[9] and defined it as complex knowledge of traits, behaviors, and the potential of situations to elicit traits into manifest behaviors. According to Christiansen et al., dispositional intelligence is a "declarative knowledge structure" (p. 126) that enables behavioral information processing. According to the original framework, judgment accuracy may depend on three separate components in this construct, namely behavior-trait knowledge (the ability to know how traits manifest themselves in behavior), understanding of trait co-occurrence (an understanding of how traits and their behavioral manifestations naturally co-vary), and situation-trait relevance. In order to enhance conceptual clarity and ease of use in our discussion, we use the following labels for these components, namely *trait induction*, *trait extrapolation*, and *trait contextualization*.

However, these earlier studies on dispositional reasoning did not consider its underlying components, that is, the role of trait induction, trait extrapolation and trait contextualization in judgment accuracy is unclear. As a result, we know little about the componential nature of dispositional reasoning and how these components individually facilitate interviewer accuracy. Furthermore, if it were possible to measure the components of dispositional reasoning reliably, it would enable tests of whether it may be understood as an intelligence measure. For it to be considered an intelligence measure, a specific mental ability should meet several conceptual and empirical criteria (Carroll, 1993; Flanagan, Genshaft, & Harrison, 1997; Mayer, Caruso, & Salovey, 1999). To this end, we developed a revised measure of dispositional reasoning, that is, one with reliable components. In **Chapter 4**, therefore, we report on the testing of whether the components of dispositional reasoning adhered to the criteria for classic intelligence measures.

[9] In our view, it is too early to conclude that this construct can be classified as an intelligence, and therefore we label it as dispositional *reasoning* henceforth in the present dissertation.

In summary, then, the second research question addressed by the present study was formulated as follows:

> *Research Question 3:* Does dispositional reasoning meet the classic criteria for an intelligence measure, considering the relationship between interviewers' dispositional reasoning components, general mental ability, personality, and their judgment accuracy for rating interview dimensions?
>
> a) Do the components of dispositional reasoning (trait induction, trait extrapolation, and trait contextualization) converge with one another and with general mental ability?
> b) Do the components of dispositional reasoning show discriminant validity with Big Five personality measures?
> c) Do the components of dispositional reasoning show incremental validity over general mental ability in predicting interview dimension rating accuracy?

Although the implicit question embedded in these series of tests was whether or not dispositional reasoning would adhere to the standard criteria for a classic intelligence, our research also sought to position the dispositional reasoning construct within a nomological network of other individual differences and judgment accuracy. Taken together, this evidence allowed us to further assess the construct validity (Cronbach & Meehl, 1955; Messick, 1995) of dispositional reasoning.

Does it take one to know one? Interviewer personality, chronically accessible traits, and trait judgment accuracy

Reporting on the final empirical study in this dissertation, **Chapter 5** presents the findings of an investigation into the relationship between interviewer personality traits, their chronically accessible personality traits, and trait-specific judgment accuracy. Personality judgments are increasing in importance in personnel selection as they not only underlie ratings used in selection devices such as interviews (Huffcutt, Conway, Roth, & Stone, 2001; Lievens, De Fruyt, & Van Dam, 2001; Van Dam, 2003), but also hold many potential benefits over self-report measures of personality. For example, these so-called 'observer ratings' (such as those by work colleagues or supervisors) can be more predictive of criteria (e.g. academic performance and job performance) and incremental to self-ratings (Connelly & Ones, 2010; Zimmerman, Triana, & Barrick, 2010).

For these reasons, it is important to understand the factors that affect the accuracy of observer ratings of personality, such as interviewer characteristics. Although interviewer cognitive factors (Christiansen, et al., 2005) appear to be consistent predictors of personality trait judgment accuracy, interviewer personality traits show relatively inconsistent, poor, and often counterintuitive relationships (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003). The reasons for

these findings are counterintuitive as they suggest that interviewer personality plays little to no role in person perception in organizations. However, personality is the predisposition to respond to stimuli in a certain way (John, Robins, & Pervin, 2008) and it affects most areas of functioning, including social functioning and social judgment in the workplace (e.g., Tziner, Murphy, Cleveland, Yavo, & Hayoon, 2008).

Earlier studies on the relationship between interviewer personality traits and personality trait judgment accuracy have two major drawbacks. First, they often used trait-generic accuracy criteria that measure how well an interviewer is able to infer a complete personality profile. A relatively unexplored avenue is the notion that personality trait judgment accuracy may be *trait-specific*, rather than trait-generic. As predicted by Funder (1995) that traits have been found to differ in their judgability – accuracy scores for judging Big Five traits vary predictably (Allik, Realo, Mõttus, & Kuppens, 2010; Connolly, Kavanagh, & Viswesvaran, 2007; Funder & Dobroth, 1987). Given these findings, we know little about how interviewers' traits may predict trait-specific, or *trait-level* judgment accuracy.

The second drawback of earlier studies was neglecting to offer an explanatory social-cognitive mechanism through which interviewer personality traits may affect their ability to judge others' traits. A growing line of research suggests that the self may be a basis for social cognitive schemas when forming impressions of others (Dunning & Cohen, 1992; Dunning & McElwee, 1995). Along these lines, interviewers' understanding of particular traits may be related to their own personalities. For example, is it possible that extraverts would be better at rating extroversion, in other words, 'does it take one to know one?' in personality trait judgments? A complementary explanation may be that interviewers' own traits are more salient in their perceptual schemas. Drawing on construct accessibility theory (Higgins, 2012) we also determine in **Chapter 5** whether interviewers' chronically accessible traits, defined as the degree to which individuals differ in the readiness with which each construct is utilized in information processing of behavioral stimulus input (Higgins, et al., 1982, p. 45), may possibly predict their personality trait judgment accuracy. Moreover, would a more parsimonious account, where the chronic accessibility for a trait partially mediates the effect of an interviewers' personality trait on judging the corresponding trait in others, be supported? To our knowledge, no earlier studies have tested these ideas empirically.

Research question 4: Would interviewers' personality traits and their chronically accessible traits predict personality judgment accuracy that is trait-specific?

a) To what extent are interviewers more accurate at judging traits they share with targets?
b) Is there a relationship between interviewers' personality traits and chronic accessibility for corresponding personality traits?
c) Is interviewers' trait accessibility for personality traits related to the degree of accuracy for judging corresponding traits?

d) What is the incremental validity of personality trait chronic accessibility over personality in predicting trait judgment accuracy?
e) Does trait chronic accessibility partially mediate the effect of traits on judgment accuracy for corresponding traits?

In answering these questions, **Chapter 5** presents two empirical studies. In Study 1, college students were asked to infer the personality profiles of hypothetical interview applicants from behavioral descriptors. We also measured students' own personalities and determined whether trait-level accuracy could be predicted from their personality traits. In Study 2, the first investigation was replicated in a field sample of interviewers in a large financial company and we also measured their chronically accessible personality traits and investigate their predictiveness of trait-specific judgment accuracy for corresponding traits. Together, these studies aimed to add to the current understanding of the role of interviewer personality and trait construct accessibility as individual differences in interviewer judgment accuracy.

The four research questions as described above have guided the research that is presented in the five chapters to follow. Each chapter describes a unique study, which may be read independently from the other chapters of this dissertation. The samples utilized between studies were relatively independent[10] and some studies contained multiple samples. In closing, **Chapter 6** presents answers to the research questions and presents practical implications and recommendations. Chapter 6 also carves out ideas for future research.

[10] The only exception was the inclusion of dispositional reasoning item scores of managers in the study reported in Chapter 4, within our sample of managers used for the analysis of measurement properties, reported in Chapter 3.

Chapter 2

The profile of the 'good judge' in HRM: A systematic review[11]

In this chapter, we provide an overview of individual difference characteristics that have been associated with the 'good judge' in human resource management over more than 60 years. We review empirical findings to identify what we know and do not know about the individual difference causes of rating quality. Overall, findings suggest that cognitive factors show stronger and more consistent relationships with rating accuracy than personality-related factors. The discussion concludes with some thoughts about the future of individual differences studies.

[11] This chapter is in preparation for publication review as:
De Kock, F. S., Born, M. Ph., & Lievens, F. (2015). *The profile of the 'good judge' in HRM: A systematic review*. Manuscript in preparation.
An earlier version of the study in this chapter was presented at the 29th annual conference of the Assessment Centre Study Group (ACSG), Stellenbosch, South Africa, March 2009.

2.1 Introduction

In human resource management (HRM), ratings of others are ubiquitous (Guion & Highhouse, 2011). As organizations rely on ratings to make important selection and promotion decisions (Schmitt & Chan, 1998), it is easy to understand why so much effort has gone into understanding not only how people evaluate others (see Graves & Karren, 1992; London, 2001; Parsons, et al., 2001) but also into identifying the characteristics of effective raters (e.g., Christiansen, et al., 2005; Graves, 1993; Powell & Goffin, 2009).

In social and personality psychology, a longstanding line of research has sought to determine what makes the 'good judge'. In this vein, Funder (1999, p. 142) concluded that:

> The search for the good judge … is the oldest pursuit in the accuracy literature and was nearly its sole concern during the early incarnation from the 1930s … the prey proved to be unexpectedly elusive. Despite the research attention it has received, the good judge is the potential moderator concerning which the accuracy literature has the sparsest data and fewest firm findings to report.

In pursuit of a coherent explanation for judgment accuracy, Funder developed the Realistic Accuracy Model (RAM; Funder, 1995, 1999, 2012), one of the most well-known models of a good judge in personality and social psychology. RAM suggests that judges[12] are important moderators of accuracy. In Funder's (1995) framework (see Figure 2.1), accurate judgment results from a social cognitive process that proceeds in four stages, in which the target person first emits a behavior that is *relevant* to the trait to be judged, in a manner where this information is *available* to the perceiver, who must then *detect* and correctly *utilize* the information to form an accurate judgment. In an interview, for example, assuming there is useful behavioral information available to the evaluator, the interviewer can make accurate judgments of an interviewee only if it is possible for them to correctly detect and use behavioral cues displayed by the interviewee.

[12] In this chapter, the terms 'judge' and 'rater' refer to interviewers and assessors (e.g. psychologists, managers, or other trained assessors).

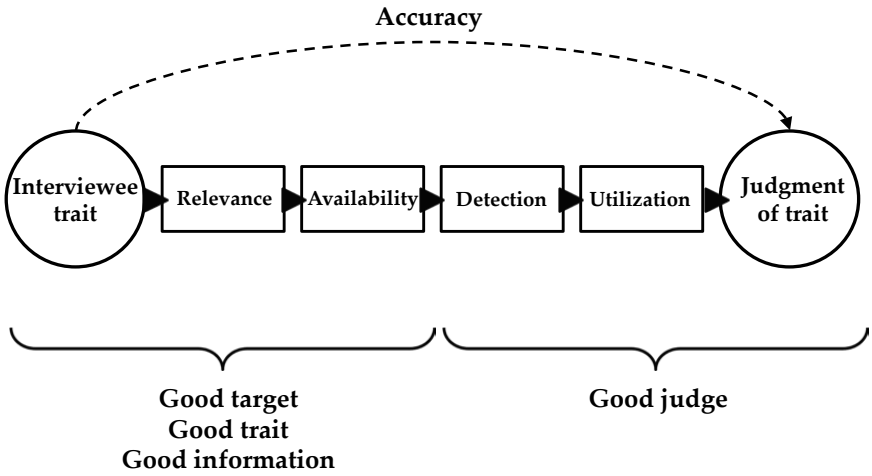

Figure 2.1. The Realistic Accuracy Model, applied in interviewer judgments: processes and moderators. Adapted with permission from Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, p. 659.

As a consequence of its breadth, RAM does not focus on rater individual differences that may facilitate judgment processes. Although Funder (1999) suggests a few characteristics of the good judge, a better understanding is needed of what these are. Furthermore, we do not know how assessor constructs (for example, see discussion in Jones & Born, 2008) may affect judgment processes that lead to accuracy. It is unclear from the HRM research base how various rater individual differences facilitate judgment processes. Stated otherwise, we do not know the link between individual differences and cue detection and cue utilization in the HRM domain. This might be another reason why personality and social psychological studies often fail to identify characteristics associated with being a good judge.

A review of more than 60 years in the HRM literature shows a steady flow of empirical studies in HRM that tested individual difference constructs as predictors of accuracy (e.g., Borman, 1979; Christiansen, et al., 2005; Powell & Goffin, 2009). However, this work has not been synthesized into a profile of the good judge that is supported by empirical evidence. As such, we are not sure what we know, or do not know, about the individual difference constructs of accurate raters in HRM (Guion, 2011). Furthermore, a better understanding is needed of how rater individual differences can be linked to judgment processes outlined in RAM.

Information on the good judge can be useful for HRM in different ways. A profile of the characteristics of accurate raters cannot only help reveal how individual differences enable better judgments; it may also have practical benefits. For example, interviewers and assessors can be screened on measures of individual differences that predict their judgment accuracy. One such individual difference

construct, dispositional reasoning, is defined as a judges' complex understanding of traits, behaviors and a situation's potential to manifest traits into behaviors) (Christiansen et al., 2005). In addition to rater screening applications, knowledge about rater constructs may inform better ways to develop raters (e.g., Lievens, 2001; Roch, et al., 2012) by targeting specific constructs in rater training. By illustration, if managers' knowledge about performance dimensions affects rating quality (Woehr, 1992) it may help to implement training programmes that focus on enhancing these schemas.

The Present Study

In the present study, we reviewed empirical research on individual differences that predict rating quality in the HRM domain.[13] As an organising framework for our review (see Figure 2.2), we relied on the Realistic Accuracy Model (RAM) (Funder, 1995, 1999). According to this framework, judges' ability to detect and use behavior cues holds the most potential to advance our understanding of individual difference constructs that may affect rating quality in HRM. As such, we also link the constructs reviewed to the specific stages of RAM and underline their relative degree of empirical support.

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 are able to outline what we know and do not yet know about the characteristics of the good judge in HRM. We summarize this empirical research base in Table 2.1. Finally, the review outlines 20 questions (see Table 2.2) that hold most potential for advancing knowledge of individual differences in judgment accuracy.

2.2 Method

Literature Search

Collecting Possible Studies

Four methods were used to locate relevant studies. First, a computer search of Web of Science and Dissertation Abstracts was conducted to retrieve research studies containing the terms *accuracy*, *rater*, *rating*, *judgment*, *interview*, *performance*, *assessment centre*. We filtered the resulting list 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 that have published accuracy studies, including *Journal of Applied Psychology*, *Personnel Psychology*, *International Journal of Selection and Assessment*, *Human Performance* and others. Third, publications were located within reference lists of major accuracy review studies,

[13] For this reason, our study does not consider investigations in adjacent disciplines, such as social psychology, nor in more specific lines of research, such as the study of judgments of non-verbal behavior.

and individual relevant studies, published both in journal article and book form. Last, we also trawled the personal research websites of active accuracy researchers.

Inclusion Criteria

These searches resulted in a large number of studies. To be included in our review, a study had to meet the following six criteria:

1. Most important, the study had to include individual differences as predictors of rating quality indices. Rating quality is broadly defined here as the degree to which ratings reflect desirable psychometric characteristics. As such, rating quality measures include rating accuracy indices (construct validity) and rating validity coefficients (criterion-related validity), but none of the various rating error measures, such as halo, leniency, etc. Even though both accuracy and error measures are often used as rating outcomes, these measures show little empirical relationships with one another. For example, in a meta-analysis by Murphy and Balzer (1989), the average correlation between rating error (various indices) and rating accuracy indices was a mere .05 (see also Kasten & Weintraub, 1999). Because of this limited overlap, rating error measures will not be considered as dependent variables in this review; rating accuracy measures will be our focus.
2. We excluded accuracy studies that used rating tasks, stimuli or target dimensions that were not immediately relevant to practical HRM applications. For example, studies that used students to judge the sexuality of other students from non-verbal behavior were not considered relevant. This criterion excluded a large number of studies from our main review, as many studies (e.g., Davis & Kraus, 1997; Murphy & Hall, 2011) focused on judging moods, emotions or affective states of others. As the bulk of this work may be peripheral to HRM applications, we do report on their findings in this review. However, a summary of these studies is available from the first author.
3. We considered only peer-reviewed publications, but irrespective of publication date. The studies retained ranged from 1953 to 2011.
4. We only retained empirical studies reporting relationships between any rater characteristic and rating quality indices such as accuracy and validity. For example, if a study reported the effect between a rater's demographic characteristic, for example, gender, and accuracy, it was retained. Quite a large number of studies reported multiple rater characteristics and, as such, the number of observations in our review are greater than the number of studies reviewed.
5. A few studies were available in both dissertation and journal article format. If we could confirm it was the same study, only the journal article information was included. Often only dissertation abstracts were available – if incomplete

or ambiguous information was available to determine the study result, a study was omitted from the review.

6. Finally, some studies included individual differences that we considered 'borderline' to our exclusion criteria, as we were not sure whether they actually addressed constructs related to the good judge. Examples are perceived similarity in attitudes with targets (Zalesny & Highhouse, 1992), demographic or cultural similarity (Letzring, 2010), and interpersonal acquaintance (Kenny, et al., 1994; Paunonen, 1989). These were deleted from the review, although they are available from the first author for future studies of possible moderators of accuracy.

Coding Procedures

Each study that fulfilled our criteria was coded by the first author on the following dimensions: (1) year, (2) sample size, (3) type of sample (students, employees, managers, country, etc.), (4) criterion measure, (5) rating task/stimuli, (6) target dimension/trait, (7) theoretical framework, (8) accuracy operationalization, and (9) effect size.

Study Characteristics

Table 2.1 shows the studies reviewed in terms of categorization variables. The majority of effects were observed in college samples (79.3%), with only a small proportion in field samples (14.9%) or mixed groups (5.8%). To employ laboratory studies that were cross-sectional was a popular choice. The mean sample size for the studies reviewed was approximately 166 participants (*SD* = 116; Min = 44; Max = 898). Apparently, the majority of studies were conducted in North America, although it is hard to quantify the proportion as many studies did not reveal the location of the research. More information about the study characteristics, such as tests used, choice of accuracy measure, and other information that was coded for each study reviewed, may be requested from the first author.

2.3 Results

We present the empirical research on individual difference characteristics in rating quality according to our organizing framework (see Figure 2.2). In this heuristic, we group judges' individual differences into constructs that are general, such as cognitive ability and personality traits, or others that are more specific, such as observation ability, cognitive styles and others. The RAM suggests that, first, target characteristics emit behavior cues that are available and relevant to the perceiver. In turn, the degree to which good judges can detect and utilize these cues will influence their rating quality (Funder, 2012). In these processes, a variety of general and specific characteristics can play an important role, reviewed next.

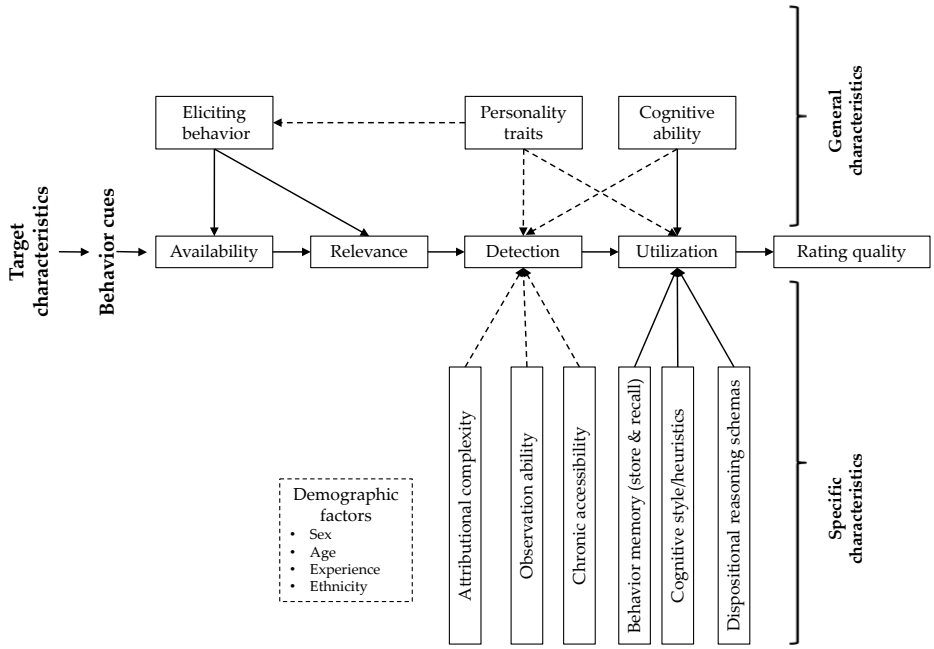

Figure 2.2. A model of individual differences in judgment accuracy, structured according to Funder's (1995) RAM. Solid arrows indicate relationships with empirical research support, whereas dotted arrows indicate relationships with limited/inconsistent research support.

General Characteristics

General Intelligence

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). We found 18 studies that reported the relationship between cognitive characteristics and accuracy as dependent variable (see Table 2.1). These include general intelligence, and other specific factors,[14] including specific intelligences (dispositional reasoning, verbal or spatial reasoning), cognitive heuristics and attributional complexity as predictors.

Intelligence may affect rating quality because intelligence enables behavior information processing: a key process in trait cue utilization (Funder, 1999). Because 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, et al., 2004), cognitive processing abilities may be an important key to producing accurate judgments (Dipboye, et al., 2012; Wyer & Srull, 2014).

[14] The specific cognitive characteristics are discussed in a later section "Specific Characteristics".

In an early review of studies on cognitive ability and accuracy, Allport (1937) concluded, "Experimental studies have found repeatedly that some relationship exists between superior intelligence and the ability to judge others" (p. 514). In later work that we reviewed, intelligence is the most consistent predictor of rating accuracy (uncorrected validity coefficients = .31; Borman, 1979; .24; Borman & Hallam, 1991; .25; Christiansen, et al., 2005; .23 - .34; Hauenstein & Alexander, 1991; .36; Lippa & Dietz, 2000; average .54; Schneider & Bayroff, 1953) of all individual differences we reviewed. That said, effect sizes are often rather modest (e.g. uncorrected $.10 < r < .30$), and some studies actually found no relationship between intelligence and accuracy (Letzring, 2008; Powell, 2008; Powell & Goffin, 2009).

Although the evidence in support of the link between cognitive ability and rating accuracy is substantial, there are still unresolved questions. First, intelligence may be complexly related to accuracy, as opposed to having a simple relationship with accuracy. For example, while most studies have assumed that these constructs are linearly related, their relationship may in fact be non-linear: Smither and Reilly (1987) for instance found that the most intelligent raters were generally less accurate than moderately intelligent raters, although moderately intelligent raters were more accurate than the least intelligent raters. To add to this complexity, studies often report significant relationships between intelligence and accuracy dependent on moderators, such as interview structure (George, 2006), motivation (Davis, 1999) or environmental complexity (Brecker, 1988). To address this complexity, more studies should explore possible moderating, mediating and non-linear effects of intelligence to unravel its role in judgment outcomes. As intelligence has been the most consistent predictor of accuracy, it is a prime candidate for these studies.

Second, in addition to the need to explore more complex effects, the field should also consider how intelligence explains accuracy in different judgment tasks and rating contexts. Could the influence of intelligence on judgment quality be affected by what is being judged, for example, interview dimensions, personality traits, or assessment centre dimensions? Logically, we expect the effect of intelligence on accuracy to increase with task complexity, as 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". In assessment centre judgments, for example, information processing loads are higher than in interviews, for example, as multiple candidates are judged, often on multiple dimensions, and also in varying situations (Melchers, et al., 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. Studies could also consider varying task complexity by manipulating aspects of the rating design, such as rating stimuli (e.g. vignettes, videos vs live people), or number of targets rated (e.g. single, typical in interviews vs multiple, typical in assessment centres). Researchers may therefore want to explore the intelligence–accuracy link by considering variations of the rating context.

Third, we speculate that intelligence may be more important to rating quality than the extant empirical literature suggests. With few exceptions, all the studies reviewed here used college students, which may cause restriction of range in ability-based measure scores. Restriction of range (see Nunnally & Bernstein, 1994) could have deflated the observed correlations we have reported, due to the relatively homogenous nature of the typical university samples used in accuracy studies. That is, intelligence may actually predict accuracy in general, non-college populations (where people may reside who may often do lay interviews) at a much higher level than usually observed in college samples.

Finally, we note that all the studies we have found have tested direct effects between intelligence and rating quality as distal outcome measure. As a consequence, the role of intelligence in the intermediate stages of cue detection and cue utilization is not yet clear. However, direct measures of cue utilization and cue detection to test this hypothesis have not been used before. The effect of intelligence on cue detection[15] has not been explored and deserves more attention. From a signal-detection theory (Lord, 1985) point-of-view, it is possible that intelligence may enable an increased sensitivity to the perception of social cues.

Personality-Related Variables

Personality Traits. Judges' personalities may regulate their social functioning in the workplace, including aspects of interpersonal judgment (e.g., Tziner, et al., 2008). Personality may affect rating quality, but this link is probably more complex, as compared to the case of cognitive predictors. A number of conceptual arguments have been proposed for the process through which personality may affect accuracy (for an overview, see Christiansen, et al., 2005; Funder, 1999). Broadly, these can be grouped into two streams. First, more proximally, personality may enhance accuracy motivation, primarily by affecting perceptual processes. That is, some people may be inclined to study and interpret the behaviors and dispositions of others because of the instrumental value of understanding others' intentions, or because they simply enjoy studying others socially. So, as a function of their traits, judges may differ in the importance they attach to social information and in their motivation to judge others accurately.

[15] See our discussion of problems with behavior observation measurement later in this chapter.

Second, a distal explanation holds that personality may influence social interaction which, in turn, gives the judge the opportunity to develop accuracy faster than judges who have less social interaction. Stated otherwise, personality may affect accuracy through the mediating role of social interaction. Instead of exerting a 'main effect' on accuracy, personality affects preferences for social interactions – and importance attached to relationship skills in general – to such an extent that accurate judgment either develops or deteriorates because of increased or decreased social interaction.

Specific traits may enhance or detract from accuracy when we apply both of these views. For example, agreeable individuals show more concern for others' feelings (Digman, 1990) and should, therefore, be more attuned to other individuals with whom they are required to interact and about whom they form impressions. Extraverts are further 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), resulting in higher accuracy. However, a counter-argument could hold that, due to their higher tendency to be focused on the self (Goldberg, 1992), extroverts may be less likely to detect behavioral cues about others' behavior. Conscientiousness, on its part, manifests in greater detail orientation (Goldberg, 1992) generally, but may also affect how we form impressions about others. For example, high conscientiousness judges are likely to be more attentive in cue detection than low-conscientious judges and show greater consistency in cue utilization. Furthermore, persons higher in openness are more inquiring and frequently enjoy working with abstract ideas or concepts (Goldberg, 1992). They are also more likely to actively develop mental representations of other's traits and behavior (Kihlstrom & Hastie, 1997), seek patterns of consistencies and inconsistencies, and form and test hypotheses about behavior (Kruglanski & Ajzen, 1983). This personality trait may also relate to either the cognitive style of the judge, such as his/her need for cognition – a construct which has explained accuracy of performance judgments in at least one study (Palmer & Feldman, 2005). Or, openness could affect accuracy through its relationship with dispositional reasoning (Christiansen, et al., 2005). Openness to experience tends to correlate with measures of cognitive ability (Ackerman & Heggestad, 1997) and social intelligence (Shafer, 1999) – both characteristics which have been linked to higher accuracy. There are therefore numerous ways in which openness to experience could potentially enhance accuracy. We urge researchers to look into these opportunities.

Despite their theoretical relevance to social interaction and judgment accuracy, these hypothesized links between personality and accuracy have received very little and generally inconsistent empirical support (e.g., Borman, 1979; Borman & Hallam, 1991; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003). 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 (r = .23, p < .05) showed a small to medium effect[16] with one of the accuracy measures, namely interview accuracy. Overall, it appears that the good judge most likely does not score higher (or lower) on certain traits, given the empirical evidence we have gathered thus far. That is, no trait seems to emerge as consistent predictor of accuracy. Our review shows that most empirical studies including personality as predictor of accuracy outcomes have shown null or inconsistent findings (for a detailed list, see Table 2.1). Generally, the observed effects of personality on accuracy, where found, are also rather small (e.g. .10 < r < .20). In fact, there are some traits that may also be detrimental to accuracy, for example aggression (Borman, 1979), being domineering and vindictive (medium effects, Letzring, 2008), and neuroticism (Gibson, 2006). Counterintuitively, judges who are less sociable may be more accurate (Ambady, et al., 1995; Sait, 2014) than sociable individuals.

Moreover, we interpret the personality–accuracy literature with caution, as some of the studies we surveyed are often plagued by high family-wise error rates. That is, by combining multiple personality traits and behaviors (often more than 30) and various operationalizations of accuracy (e.g. by relying on permutations of 'true-score' source, accuracy index, and so forth) very large correlation matrices result. Even when using these 'empirical dragnets' (often employed in the earlier stages of exploratory research of individual differences), only a few studies report relatively few notable effects (for an example of such a 'broad' study, see Borman, 1979).

A number of questions remain about possible links between personality and accuracy. A useful area for future research is to explore more complex hypotheses. Perhaps personality constructs are more complexly related to rating quality than thought earlier. For example, personality may affect accuracy indirectly as a moderator variable. On the basis of their analyses, Christiansen et al. (2005) concluded that the relationship between 'dispositional intelligence' and acquaintance accuracy was moderated by conscientiousness and agreeableness. When elevation on these two traits was high, dispositional intelligence predicted acquaintance accuracy better than when elevation on these traits were low. So, as with cognitive constructs, perhaps the field of accuracy research can benefit from moving away from simple effects studies and explore more complex questions.

Personality-related Behavior. Another avenue for study is to consider the links between judges' behaviors, as manifestations of their personality, and rating quality. It is likely that judges' behaviors affect the availability and relevance of cues. Good judges are not 'passive perceivers', but seem to participate actively in interpersonal situations when forming impressions (Graves, 1993). When judges interact with

[16] Effects reported are observed correlations and have not been corrected for unreliability, nor for restriction of range, unless stated explicitly that they have been corrected.

targets, judges' personalities may play an important role in creating the conditions for forming accurate impressions. For example, in an experimental study using unstructured interactions in triads of previously unacquainted students, Letzring (2008) found that students' judgment accuracy of their acquaintances was related to their social skill, agreeableness and adjustment, implying that judges' personalities and behaviors are important for creating situations within which targets are likely to reveal relevant personality cues (Letzring, 2008). More specifically, accurate judges emphasized others' accomplishments (uncorrected correlation = .32), engaged in constant eye contact (.28), compared themselves to others (.21), expressed warmth (.21), enjoyed the interaction (.20), displayed ambition (.20), seemed interested (.19), and expressed sympathy (.18). This emerging line of research is introducing new elements into RAM, as the interviewers' behavior does not relate to either detecting or using cues better, but actually, to elicit cues. 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* is a promising future avenue for work on the good judge (Lievens, et al., 2015).

A relatively unexplored avenue lies in the use of behavior prompts to actually test or confirm initial impressions of targets, just like an interviewer would use verbal prompts to confirm or disconfirm an initial 'impression hypothesis'. For example, in Kruglanski's lay epistemic theory of judgment (Kruglanski, 1990), it is proposed that judges go through a cyclical process of hypothesis generation and hypothesis testing (see also Sackett, 1979, 1982) of an inferred profile of the target. If so, how do interviewers employ specific behaviors to test these impressions by using verbal and non-verbal behaviors? We encourage more research along these lines.

Motivation. For the purposes of this review, rater motivation is discussed under the broad umbrella term of 'personality-related factors'. When raters are motivated to produce accurate ratings, they may produce more accurate ratings. Developed in the performance appraisal context, the cognitively oriented rater motivation model of Harris (1994) suggests that personal (mood, self-efficacy, information quantity) and situational (accountability measures, trust, forms, etc.) factors are likely to affect cognitive processes (rating context, including observation, storage, retrieval, integration and rating) thought to affect judgment accuracy.

To enhance rater motivation, perceived intrinsic and/or extrinsic rewards for producing accurate ratings may be effective, but it appears there is little research evidence on this matter. In one study, incentives to produce accurate ratings led to higher accuracy (Salvemini, Reilly, & Smither, 1993), but it is unclear whether accuracy motivation was the cause of this effect – accuracy motivation was not measured explicitly – or something else. Another plausible, but relatively

unexplored question is whether the avoidance of negative consequences (e.g. scorn by one's supervisor) may lead to accuracy motivation.

Rater motivation may be influenced by perceived accountability for one's ratings (Mero & Motowidlo, 1995; Mero, Motowidlo, & Anna, 2003). For example, across a number of studies (e.g., Strupeck, 2004; Wood & Marshall, 2008), accountability expectations were fairly effective at improving judgment accuracy, showing medium to large effects. In another investigation (Rosenbaum, 1992), when study participants were held accountable for their ratings, or perceiving high consequences from their ratings, it promoted accuracy. However, the type of accountability may also play a role: Brtek and colleague (2002) observed that procedure accountability increased interview validity (r = .26), although outcome accountability lowered it (r = -.17), but these results ran counter to findings from an earlier investigation (Craven, 1988). From empirical work, it seems that being held accountable for one's ratings may enhance accuracy, but more research is needed to explore the generalizability of these findings across HRM contexts, as most of these studies were conducted in performance ratings.

Motivation may increase attention to behavior cues, as well as lead to judges assigning greater cognitive resources to cue utilization, that is, through expending greater effort. For example, when raters believe that rating effort would result in desired rating outcomes (e.g., expectancy, valence, and instrumentality; Vroom, 1964) they may invest greater effort in the rating process. However, it does not seem that rating effort enhances accuracy (Borman, 1979).

These questions therefore remain conjecture, however, pending empirical evidence. Such studies could provide answers to how rater motivation may enhance accuracy. A fruitful avenue for future research is to test dual-processing theories of judgment developed in the social-cognition field, where it is well established that factors such as motivation or mood may act as a 'gear lever' that selects the operation of either conscious (and deliberate) judgment processes, or unconscious (and automatic processes) (Fiske & Taylor, 2013).

Specific Characteristics

Demographic characteristics

A number of demographic characteristics have been considered as direct effects on rating quality measures, although none have considered why or how the judges 'physical' characteristics may affect cue availability, relevance, detection or utilization. As such, we discuss only the available empirical evidence on 'direct' effects.

Rater Gender. Of all the demographic characteristics, rater gender has been the most often-studied predictor of judgment accuracy. Hypotheses about gender differences in accuracy have been driven by gender disparities in constructs that are thought to drive accuracy, for example, interpersonal sensitivity (Hall & Bernieri,

2001). However, research findings are not clear-cut. In some studies (Ambady, et al., 1995; Carney, et al., 2007; Letzring, 2010; Schmid Mast, et al., 2011; Vogt & Colvin, 2003), female judges were more accurate than male judges, while others showed no gender differences (Christiansen, et al., 2005) or reported mixed findings (Chan, Rogers, Parisotto, & Biesanz, 2011; Letzring, 2008). In others, the trait moderated this effect. For example, female judges are sometimes better at rating only certain single specific traits, but not all, e.g. extraversion and positive affect (Ambady, et al., 1995), neuroticism (Lippa & Dietz, 2000; Schmid Mast, et al., 2011), and vulnerability to stress (Powell, 2008). In the emotions literature, women's superiority in accuracy is also qualified by the content domain and gender-specific motivation. In the judgment of non-verbal expressions of emotions, women are generally more accurate (Hall & Schmid Mast, 2008) – an interesting avenue for future research to explore possible links between gender-related personality traits (i.e. measures of masculinity and femininity) and accuracy (Lippa & Dietz, 2000).

Rater Age. Except for gender, other demographic factors (e.g. ethnicity, culture, age and others) have received scant research attention. Age does not appear to affect accuracy in the few studies that have reported these findings (e.g., Borman, 1979), although we may expect age effects when age is confounded with rater experience (see below; rating experience).

Culture/Ethnicity. An interesting direction for research is to consider culture or ethnicity effects in judgment accuracy (e.g., Albright et al., 1997). For example, in one study of performance dimension judgments, collectivism was related to lower differential accuracy (Paquet, 2005). In an increasingly multicultural workplace, cultural or ethnic differences in judgment accuracy should be explored further.

Rating Experience. Rating experience may also enhance rating quality. When judges continuously refine schemas and heuristics, based on feedback in the form of observing judgment outcomes (Fiske & Taylor, 2013), it becomes possible to refine their judgment 'algorithms'. We expect that, when an initial impression of a target turns out to be accurate (the target does something that confirms the judges' expectations), this information may enhance not only the rater's schemas, but also bolster rater self-efficacy and confidence. Empirical studies show that accuracy may be higher for judges with more experience (Kolk, et al., 2002), but these effects are rather small (uncorrected validities = .18, Wood & Marshall, 2008) or negligible (Borman, 1979). In one study, observational accuracy was actually lower for judges with more experience (-.16, Borman & Hallam, 1991) than for judges with less experience. In an experimental investigation (Schmid Mast, et al., 2011), recruiter job experience positively predicted lie detection accuracy, but lowered accuracy for judging extraversion. The link between rater experience and accuracy is therefore not as straightforward as theory may suggest.

Attributional/Cognitive Complexity

Attributional Complexity. Attributional complexity is defined here as the tendency to engage in complex social information processing and inferential reasoning (Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986). Raters with high levels of attributional complexity spend more time processing complex social cognitive stimuli (Fletcher, et al., 1986), most likely because they are more motivated to understand other's behavior and show preference for in-depth attributional reasoning, than raters low on attributional complexity (Fletcher, Rosanowski, Rhodes, & Lange, 1992). It is clear, then, that attributional complexity may influence cue detection and utilization: high-complexity raters are more likely to spend time looking for, and thinking about, others' behavior and what it says about them.

The research base for attributional complexity as a predictor of accuracy is scant. High-complexity raters produce more accurate causal answers to difficult causal social cognitive problems (Fletcher, et al., 1986) than low-complexity raters and make more accurate judgments of traits and attitudes (Fletcher, Grigg, & Bull, 1988). This is because "both [attribution making and empathic accuracy] may be outcomes or products of a more general epistemic attempt to 'understand' another person" (Ickes, Stinson, Bissonnette, & Garcia, 1990, p. 736). However, more work must be done to link attributional complexity with rating quality outcomes in HRM contexts. As far as we know, there are no empirical studies that have tested the predictiveness of attributional complexity measures to predict interviewers' or assessors' rating quality.

Cognitive Complexity. Attributional complexity is closely related to 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). Rooted in personal construct theory (Kelly, 1955), cognitive complexity speaks to raters' ability to perceive behavior in a multidimensional fashion. As with the sparse evidential basis for attributional complexity as predictor of rating quality, later research efforts on cognitive complexity (Adair, 1987; Bernardin, Cardy, & Carlyle, 1982; Borman, 1979; Gerber, 2013) have consistently failed to support the complexity of a rater's attributional processes as a predictor of accuracy, despite early optimism.

It is possible that more conceptual clarity and better operationalization of cognitive complexity may lead to more empirical support in accuracy studies (Guion, 2011; Woehr, Miller, & Lane, 1998). We urge more research that addresses measurement issues associated with attributional and cognitive complexity. For example, these measures are typically based on self-reports. As such, we have questions about their validity as indicators of *actual* complexity, as opposed to preferences for or self-perceived complexity. For example, it is not uncommon to see that ability-based measures have higher criterion-related validity than trait-based measures (e.g., for emotional intelligence; Van Rooy & Viswesvaran, 2004).

Behavior Observation

Raters are also required to detect manifestations of traits (Funder, 1999) in order to pass this information on to cue utilization functions. Observation and categorisation of ratees' behavior is the first task in producing judgments about performance (Borman & Hallam, 1991). Surprisingly little research has studied behavior observation ability as a predictor of judgment accuracy. We say this because most studies of 'behavior accuracy' (e.g., Lewis, 2002; Middendorf & Macan, 2002; Murphy & Balzer, 1986; Murphy, Garcia, et al., 1982; Sanchez & De La Torre, 1996; Sulsky & Balzer, 1988) seem to confound behavior detection with behavior recall. That is, studies typically use measures that target both the ability to *observe* behavior, as well as the ability to *remember* what they saw in the task. For example, these studies would ask raters to use a behavior list and tick off behaviors they 'remember seeing', after viewing an interview video. Often, there is a considerable delay between showing behaviors and asking raters to indicate their observations.

More thought should therefore be given to tasks that do not confound behavior observation with behavior recall (Murphy, Martin, & Garcia, 1982). More specifically, cue detection (in RAM) (Funder, 1995) may be facilitated by both cue identification on the one hand (signal detection), and cue encoding, storage and recall (memory processes) on the other. Therefore, 'pure' tests of behavior observation ability – the ability to detect behavior cues – would require study participants to indicate cue detection as behavioral cues occur, that is, in a 'live' stream of verbal and non-verbal behavior cues. To test cue detection, video stimuli may be pre-coded by expert raters who 'flag' objective behavior cues. Against this normative base-line, raters' cue detection can be analysed through their verbal protocols (e.g. 'think-out-loud studies'). Alternatively, raters may be shown video footage of targets and asked to tag behavior cues using a clicker. After viewing the interview footage, raters may be asked to explain their behavior markers after viewing the video. These tests are not contaminated with behavior memory (or recall) and would be considered 'pure' tests of behavior observation ability.

Chronic Accessibility

Raters' ability to detect and perceive cues may also be influenced by their perceptual 'filters'. One such filter, construct chronic accessibility, can be defined as the degree to which individuals differ in the readiness with which constructs are utilized in information processing of behavioral stimulus input (Higgins, et al., 1982). Individual differences in the subjective meaning of social events may be especially evident in the personal constructs individuals employ (Mischel, 1973). A personal construct system is "a kind of scanning pattern which a person continually projects upon his world. As he sweeps back and forth across his perceptual field he picks up blips of meaning" (Kelly, 1955, p.145). In turn, person-memory (that is, of targets' traits and behaviors) may be affected by trait accessibility, because trait category accessibility affects the storage, encoding and retrieval of behavioral information

(Bargh & Thein, 1985; Srull, 1981, 1983). Observed stimulus information that is related to a rater's individual accessible constructs would therefore be more readily processed and retained than information that is related to inaccessible constructs (Srull & Wyer, 1979). In this way, chronically accessible constructs may affect the degree to which performance-related dimensions are accessible for use (Woehr, 1992).

As far as we know, no studies have used chronic accessibility measures to explain rating quality outcomes directly, nor explored their link with the stages of RAM. Nevertheless, there is a compelling argument: accessibility may affect accuracy because it influences perceptual selection (Higgins, et al., 1982). Individuals with accessible constructs are more sensitive (than individuals with inaccessible constructs) to stimuli associated with those constructs (Bargh & Pratto, 1986). As such, we do not know whether judges would have higher accuracy for traits that are accessible to them. For example, would a judge with extroversion accessibility also be more adept at detecting and using extroversion-related cues? If so, it may have implications for practice. For example, assessors with chronically accessible traits may be employed as trait experts to rate specific traits in an interview. Overall, chronic accessibility is a relatively unexplored predictor of rating quality and deserves more research attention.

Behavior Memory

Social cognitive theories of judgment that are based on person-memory (e.g., Srull & Wyer, 1989) assume that memory of others' behavior and traits plays an important role in producing accurate judgments. The ability to store and recall information about targets is an important link in the judgment process. Personality trait terms can be considered as summary labels for broad conceptual categories that are used to encode information about others' behavior into memory (Srull & Wyer, 1979).

Empirical studies on memory-accuracy links are mixed, as some (DeNisi & Peters, 1996; Murphy, Garcia, et al., 1982; Rush, Phillips, & Lord, 1981) find positive effects with accuracy, and others (Lewis, 2002; Middendorf & Macan, 2002; Murphy & Balzer, 1986; Sanchez & De La Torre, 1996; Sulsky & Balzer, 1988) find trivial or no effects.

To address this apparent stalemate, researchers may benefit from rethinking the way we understand person-memory. Conceptually, person-memory models have evolved to a view that splits person-memory into two general types of information: one about specific behaviors demonstrated by a person, and the other referring to abstract personality traits or dispositions (Srull & Wyer, 1989). For example, an interviewer would recall both applicants' actions during the interview, but also the trait impressions the interviewer formed about each applicant. These memories are distinct. As we do not yet understand the comparative role of impression-memory versus behavior-memory in judgment processes, we would like

to see the comparative predictive validity for accuracy criteria. Moreover, we reiterate our view earlier that behavior memory and behavior detection measures are typically confounded in accuracy studies – these need to be teased apart in future accuracy studies.

Cognitive Style and Heuristics

The way judges think about others' behavior and form resulting impressions may affect accuracy (Brehmer, 1994). Perhaps good judges have special ways they assemble other-related information when forming a mental picture of the target person. While we know that individual differences exist in a number of 'heuristics', including implicit theories of performance (Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Hauenstein & Alexander, 1991), 'personal constructs' (Borman, 1987), decision processes (Arvey & Campion, 1982; Graves & Karren, 1992; Ostroff & Ilgen, 1992), and weighting of applicant information (Dougherty, Ebert, & Callender, 1986; Kinicki, et al., 1990; Sackett & Hakel, 1979; Zedeck & Kafry, 1977), we need more empirical studies to show whether and how these heuristics predict rating quality. For example, Hauenstein and Alexander (1991) found that students possessing a normative implicit theory of performance were more accurate in their ratings of teachers' performance than students possessing an idiosyncratic implicit theory. However, like many other largely 'once-off' investigations of individual constructs (see Table 2.1) in this category, the study has not been replicated. Therefore, the field needs larger studies and a more solid research base before firm conclusions may be drawn about the usefulness of cognitive styles and heuristics.

Dispositional Reasoning

In addition to general cognitive ability constructs, more specific abilities may also be involved in judgment accuracy. 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, et al., 2015). The 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). The same authors experimentally tested the notion that dispositional reasoning may allow good judges to process behavioral information towards accurate trait inferences. Using a lab study where students (N = 122) watched videotaped segments of individuals responding to employment interview questions, judged the personality of the video interviewees and rated acquaintances who later completed self-report personality inventories, dispositional reasoning was the best predictor of various accuracy indices (with r ranging from .41 to .52), in fact, better than general mental ability and personality. More recent work has partially replicated the predictive validity of dispositional reasoning for judging personality (Powell & Goffin, 2009) and interview dimensions (De Kock, et al., 2015), but trivial effects have been observed with judges' self-rating accuracy (that is, self-perceptions of their own strengths and weaknesses) (Janovics, 2003).

Although the research base about this construct is in its adolescence, findings are promising. More work is needed to consider its constituent components and their role in producing quality ratings. For example, De Kock, et al. (2015) developed reliable component[17] measures of dispositional reasoning (for *trait induction*, *trait extrapolation*, and *trait contextualization*) that each predicted accuracy and jointly incremented general cognitive ability to predict accuracy. Also, judged against a number of criteria for classic intelligences (Mayer, et al., 1999), the results showed that dispositional reasoning was characteristic of an intelligence measure. Future studies could assess the nomological placement (in a network) of dispositional reasoning components in relation to 'adjacent' constructs that are conceptually similar, such as emotional intelligence and social intelligence. For example, Christiansen, et al. (2005) report that dispositional reasoning co-varied with general intelligence (r = .43) and also mediated its effect on accuracy outcomes, suggesting a more complex interrelationship with accuracy as a mediator of other variables.

Other Characteristics

In this category, a number of 'other' individual difference constructs are discussed, none of which have received much attention, nor have they suggested potential usefulness as predictors of accuracy. Attitudes have been inconsistent (*cf*. Gibson, 2006; Hartog, 1991) and vocational interests (e.g., Holland's [1973] six interest types were studied in Borman, 1979) were poor predictors of judgment accuracy. In this sparse research base, findings are often counterintuitive (for example, 'social interest' correlated negatively, -.17, in Borman, 1979). A number of other diverse constructs, not discussed, are listed in Table 2.1. On the whole, there is not yet a compelling case to take note of 'other' characteristics as predictors of rating quality. For the time being, it would be better for the field to invest energy elsewhere where more promising research findings have emerged.

Pseudo-Constructs

We reserve this category for individual difference constructs that may require more attention to operationalization. That is, a number of effects that we found relied on self-report measures, for example self-rated non-verbal communication skills and rating effectiveness (Schmid Mast, et al., 2011), and self-perceived attributional complexity (Davis, 1999), which raises questions about successful operationalisation of these constructs. Future studies should consider using ability-based measures of individual differences were abilities are implied, for example, to allow more conclusive evidence.

[17] These components were introduced by Christiansen et al. (2005) and they were relabelled for ease of use in De Kock et al. (2015) – the latter study is contained in Chapter 4 of this dissertation.

2.4 Discussion

Main findings

Our review of the literature on individual differences in accuracy suggests that more is known about the 'good judge' than earlier thought. In our review of research from 1953 to present, we found at least 126 individual effects reported in 48 works (published articles and unpublished dissertations and theses). Individual differences research therefore remains an active interest in HRM and I-O psychology. Together, these studies have explored many rater characteristics that span across various functional psychological domains.

Overall, empirical evidence suggests that cognitive factors play a dominant role in judgment accuracy. For example, the good judge is not only more intelligent than less accurate judges, but also has better understanding of others' behaviors, traits and situations. By virtue of better social information processing and memory, good raters can process, store and recall targets' behavior better than poor raters can. Good judges also seem to possess better developed schemas that relate to social information (for example, about the traits that underlie behaviors, implicit personality theories, and understanding situational contexts).

Our review of the individual differences literature shows that effect sizes for cognitive factors are normally moderate to large, and these appear to be relatively consistent in laboratory studies. It is interesting to note that effect sizes tend to be larger when ability-based measures (as opposed to self-report measures) of cognitive factors are used. This trend is similar to findings that reveal how ability-based measures of emotional intelligence show stronger effects with related outcomes than self-reported measures of emotional intelligence (O'Boyle, Humphrey, Pollack, Hawver, & Story, 2011).

It is not only cognitive *abilities* that may help to explain why judges are accurate or not, but also the way judges think about others' behavior or process behavior information (for example, implicit theories, judgment policies and analytic orientation). Although replication studies about the predictive validity of non-ability cognitive factors are relatively sparse, the emerging results are promising because they open up new questions for research. First, can these rating algorithms be developed by means of systematic training, similar to frame-of-reference (FOR) training (Roch, et al., 2012; Woehr & Huffcutt, 1994), where evaluative schemas for judging behaviors into dimensions are imposed on raters in order to enhance accuracy? Or are these judgment 'algorithms' better acquired by long-term experience? Second, are these judgment processes deliberate or do they operate automatically? Dual-processing theories of judgment in person perception (Fiske & Macrae, 2012) have become the norm in social cognition literature, and these suggest that judges use both conscious and unconscious processes to evaluate others. We

urge more research that takes a broader view of individuals' judgment processes (in other words, including both deliberate and automatic aspects) and rating quality in HRM applications.

The good judge most likely does not score significantly higher or lower than others on a particular personality profile or personality trait. Our review showed that none of the broad Big Five traits are consistent predictors of accuracy. Narrow traits that are socially oriented may be useful as predictors of accuracy, but little research has considered this angle. In studies that do report statistically significant effects for personality traits on accuracy, effects tend to be relatively small or trivial (Cohen, 1988) and most likely, also practically negligible. This conclusion does not preclude the possibility that personality may be more complexly related to accuracy, for example, as a moderator of the influence of cognitive variables on judgment accuracy (e.g. Christiansen et al., 2005).

Theoretical implications

In sum, our study provides support for the notion that 'good judges' are moderators of accuracy, as proposed by Funder's RAM (Funder, 1995, 1999). Our review of the empirical evidence shows that a vast array of rater constructs has been considered as predictors of accuracy in HRM settings such as interviews, assessment centres and performance rating. Overall, we found evidence that some individual differences may indeed explain the 'good judge', but that much remains to be learned. We have synthesized this literature into a comprehensive 'good judge' model (see Figure 2.2) that answers calls (see Jones & Born, 2008) to explain how assessor constructs may facilitate specific judgment processes.

Moreover, our model links individual differences to rater judgment processes thought to cause accuracy (RAM) (Funder, 1999), namely cue detection and cue utilization, and we weigh the evidence in support of each hypothesized link. The bulk of work has implicitly argued that constructs that were studied may enhance cue utilization. As such, assessor constructs that may enhance cue detection is an area that is ripe for study. In fact, not many studies have taken a broad perspective by integrating individual differences that are implied in both cue detection and cue utilization – both processes are needed in order to produce an accurate impression of the target.

Our review has also identified potential extensions of Funder's RAM. Built on the basis of physical perception models (e.g., Brunswik's Lens model; Brunswik, 1956), RAM implies a view of the judge as a passive observer: waiting to pick up on behavior signals and using these validly in forming a mental picture of the target. In contrast, research findings (e.g., Letzring, 2008) suggest that when they interact with targets, good interviewers actively *elicit* good behavior cue information. That is, they actively encourage the interviewee to express useful trait-relevant information: They use interviewing and other skills (for example, listening or non-verbal

communication) to put the interviewee at ease, draw out more information, test initial impressions, and so forth. These findings suggest the need to consider adding a third judgment process to RAM, namely *cue elicitation* (see Lievens, et al., 2015).

Future research directions

The field of individual differences research on the good judge in HRM needs to grow in several directions. First, researchers in this field stand to benefit from drawing upon social-cognition frameworks and methods to illuminate judgment accuracy issues in the workplace. Although social psychology theories may sometimes not generalize easily to other settings (Ilgen & Favero, 1985), our review shows many successful applications of social cognition theories and principles to workplace rating accuracy problems. In this work, it should be kept in mind that "individual differences ought to be considered central in theory construction, not peripheral" (Underwood, 1975, p. 129, cited in Revelle, Wilt, & Condon, 2013). From our review of the field, we suggest that two primary candidate constructs, dispositional reasoning and personality trait chronic accessibility, in conjunction with general intelligence, may feature in future theory development, as their conceptual linkage with the judgment processes in RAM is compelling and emerging evidence is suggestive.

Second, we foresee opportunities to determine how interviewers and assessors manage the interpersonal interaction to elicit useful behavioral data for their judgments. Only recently, have constructs related to cue elicitation (such as behavior; Letzring, 2008) entered the mainstream of accuracy research. As studies up to now have also not disentangled judgment processes (cue detection vs cue utilization), experimental research that considers the main and interactive effects of interviewers' cue elicitation, cue detection, and cue utilization (see Funder, 1999) would be useful. Along the same lines, we note that with few exceptions (such as Borman, 1979) extant work on individual differences in accuracy have considered judges' characteristics in isolation. But we also know that behavior is complexly determined (Stanovich, 1992), that is, they are a function of various constructs that operate in different functional psychological domains. Future studies should, for example, consider how judges' characteristics interact in producing accuracy. For example, Davis (1999) found that assessors' cognitive ability and motivation interacted to produce accuracy. Are these linearly combined to produce accurate judgments, or can they compensate for one another? Or, if we consider the various constructs that 'cause' accuracy, which would provide incremental validity in predicting accuracy? We also do not know whether two judges can produce judgments with similar accuracy, although they rely on different judgment strategies and/or abilities. That is, can they achieve the same objective (same ends) but use different paths (different means) to accuracy?

Third, there is an interesting and growing divergence in accuracy research between the domains of social-cognition and I-O psychology. In particular, we note that the dominant perspective in social psychological literature on impression formation and interpersonal judgment relies on dual-processing theories of judgment (Fiske & Macrae, 2012). In contrast, judgment studies in HRM settings are hesitant to explore the role of unconscious processes (e.g., Highhouse, 2008). We encourage more work that incorporates both conscious and unconscious judgment processes, especially those that determine the possible interplay between these processes in producing accurate judgments. Dual-process theories suggest that judges employ both deliberate (conscious) and automatic (unconscious) processes to interpret others' behavior, and the degree to which either is used depends on aspects of the rating task and other individual differences of the judge (e.g. mood) (Fiske & Taylor, 2013). More research can be devoted to this issue in HRM research.

Finally, we reiterate earlier calls for continued attention to measures of accuracy (Colvin & Bundick, 2001). Accuracy studies often find low congruence between multiple accuracy operationalisations (e.g., Sulsky & Balzer, 1988), and our review showed this incongruence sometimes extends to substantive conclusions that depend on which measure is used. Accuracy research can benefit from consensus-seeking on the meaning and measurement of accuracy by students of accuracy in different domains (e.g. HRM contexts, personality, social psychology, non-verbal behavior). Similar to other domains of personnel selection, the criterion problem (Austin & Villanova, 1992) in individual differences in accuracy research must be addressed more satisfactorily before the field can flourish.

Recommendations for practice

In our view, there are two promising ways to advance practices used for rater training and selection, given the findings of our review. The first is the identification of judges on the basis of constructs that predict accuracy. In contrast to earlier scepticism that empirical evidence would be found of individual differences that predict accuracy, it appears there are constructs that do consistently predict accuracy. These appear to lie mostly in the cognitive domain, especially when ability-based measures are used. Our results suggest that organizations could consider using cognitive ability measures to select raters, as these measures predict judgment accuracy and are therefore 'job-relevant'. The validities in some studies approach those for predicting job performance (Schmidt & Hunter, 1998). In addition to general cognitive ability, specific intelligences related to behavioral information processing, such as dispositional reasoning, show promising results as single and incremental predictors of accuracy (Christiansen, et al., 2005; De Kock, et al., 2015) in HRM. Organizations should consider using, for example, measures of dispositional reasoning to select interviewers, assessors and raters of performance.

A second way to advance rating practices is to develop raters' constructs that predict accuracy. But first, we need to know whether these constructs can be

developed. That is, are they malleable at all? The deeper underlying question is whether or not the good judge is born, or made? A good example, again, is dispositional reasoning. It has been shown that dispositional reasoning may broadly adhere to the criteria for an intelligence (De Kock, et al., 2015). Also, early attempts to enhance one of the components of dispositional reasoning behavior-trait knowledge (also known as 'induction'), have been unsuccessful (Powell & Goffin, 2009). So, before we can recommend dispositional reasoning training, we need empirical evidence to support it and we call for more research along these lines.

2.5 Conclusion

Through the ebbs and flows over the last century, the question of 'what makes the good judge?' has endured. Individual differences research in judgment accuracy has come a long way. Our review of the empirical research shows that good judges do seem to share a number of characteristics (outlined in Table 1). Among these are cognitive factors, including dispositional reasoning and general mental ability. Although assessors' personality traits show an inconsistent relationship with rating accuracy, specific behaviors may be effective to elicit trait-related behavioral cues from targets. Not only are we coming closer to understanding the rater constructs that shape judgment accuracy, but we are also gaining insight into how these constructs facilitate processes that lead to accurate impressions (portrayed in Figure 2.2). At the same time, there are aspects of rater individual differences that have hardly been touched on, such as emotional intelligence, behavior observation acuity, and chronically accessible personality traits. Our study summarizes these in 20 researchable questions (see Table 2.2). Answers to these questions about what makes the good judge hold much potential to enhance the quality of ratings in HRM settings.

Table 2.1

Research Evidence[a] on Individual Differences Constructs Predicting Judgment Accuracy in HRM

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Hartog (1991)	250	Attitude	Attitudes	Not avail. (dissertation abstract)
Gibson (2006)	*nr*[c]	Attitude	Life satisfaction	No/negligible
Letzring (2008)	142	Behavior	Social behavior	Various (up to medium)
Murphy et al. (1982)	44	Behavior detection/memory	Behavior observation	Small to large
Borman (1979)	146	Biographical	Age	No/negligible
Paquet (2005)	181	Biographical	Culture (Individualism/collectivism)	Not avail. (dissertation abstract)
Carney et al. (2007)	334	Biographical	Gender	Small
Ambady et al. (1995)	90	Biographical	Gender	Small (negative) to medium
Lippa et al. (2000)	109	Biographical	Gender	Various (No/negligible to small)
Vogt et al. (2003)	102	Biographical	Gender	Medium
Christiansen et al. (2005)	122	Biographical	Gender	No/negligible
Powell (2008)	164	Biographical	Gender	Small to medium
Letzring (2008)	138	Biographical	Gender	Small to medium
Letzring (2010)	80	Biographical	Gender	Small to medium
Mast et al. (2011)	131	Biographical	Gender	Not avail. (not reported)
Chan et al. (2011) [study 1]	898	Biographical	Gender	Not avail. (not reported)
Letzring (2010)	80	Biographical	Similarity (demographic: various)	Small
Schneider et al. (1953)	400	Cognitive	Academic performance	Large
Bayroff et al. (1954)	*nr*	Cognitive	Aptitude	Not avail. (dissertation abstract)
Schneider et al. (1953)	400	Cognitive	Aptitude	Large
Borman (1979)	146	Cognitive	Attention span	No/negligible
Brtek et al. (2002)	338	Cognitive	Attentiveness	Medium
Adair (1987) [study 1]	147	Cognitive	Cognitive complexity	Not avail. (dissertation abstract)
Adair (1987) [study 2]	*nr*[c]	Cognitive	Cognitive complexity	No/negligible
Bernardin et al. (1982)	72	Cognitive	Cognitive complexity	Small

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Brecker (1988)	122	Cognitive	Cognitive complexity	Not avail. (dissertation abstract)
Borman (1979)	146	Cognitive	Cognitive complexity	No/negligible
Janovics (2003)	410	Cognitive	Dispositional reasoning	Not avail. (dissertation abstract)
Christiansen et al. (2005)	122	Cognitive	Dispositional reasoning	Large
Powell (2008)	164	Cognitive	Dispositional reasoning	Various (up to medium)
Janovics (2003)	410	Cognitive	General mental ability	Not avail. (dissertation abstract)
Borman et al. (1991)	79	Cognitive	General mental ability	Medium
Lippa et al. (2000)	109	Cognitive	General mental ability	Medium
Christiansen et al. (2005)	122	Cognitive	General mental ability	Small to medium
Letzring (2008)	138	Cognitive	General mental ability	No/negligible
Davis (1999)	82	Cognitive	General mental ability	Not avail. (dissertation abstract)
George (2006)	301	Cognitive	General mental ability	Small (but contingent)
Smither et al (1987)	90	Cognitive	General mental ability	Various (up to medium)
Brecker (1988)	120	Cognitive	General mental ability	Not avail. (dissertation abstract)
Hauenstein et al. (1991)	100	Cognitive	General mental ability	Small to medium
Powell (2008)	164	Cognitive	General mental ability (verbal/quant)	Small (both negative and positive)
Borman (1979)	146	Cognitive	General mental ability (verbal)	Medium
Borman et al. (1991)	79	Cognitive	Spatial ability	Medium
Adair (1987) [study 1]	147	Cognitive style/Heuristics	Attribution	Not avail. (dissertation abstract)
Johnson (1987)	73	Cognitive style/Heuristics	Cognitive modelling	Not avail. (dissertation abstract)
Lee (1988)	95	Cognitive style/Heuristics	Cognitive style	Medium
Willis (1985)	264	Cognitive style/Heuristics	Cognitive style	No/negligible
Clevenger (1991)	*nr*[c]	Cognitive style/Heuristics	Field independence	Not avail. (dissertation abstract)
Cardy et al. (1984)	359	Cognitive style/Heuristics	Field independence	Small-to-medium
Hauenstein et al. (1991)	100	Cognitive style/Heuristics	Implicit rating theory	Medium
Uggerslev et al. (2008)	236	Cognitive style/Heuristics	Implicit rating theory	Various (Small)
Borman (1979)	146	Cognitive style/Heuristics	Problem-solving style	No/negligible

Author(s)	*N*	Cluster	Predictor	Effect size[b]
George (2006)	301	Cognitive style/Heuristics	Prototypes	Not avail. (dissertation abstract)
Kolk et al (2002)	121	Demographic	Experience	Small to medium
Mast et al. (2011)	131	Experience	Job experience	Small to medium (negative)
Borman et al. (1991)	79	Experience	Rating experience	Small effect (negative)
Borman (1979)	146	Experience	Rating experience	No/negligible
Wood et al. (2008)	194	Experience	Rating experience	Small
Borman (1979)	146	Interests	Interests	No/negligible
Borman (1979)	146	Judgment task: base rate estimation	Base rate estimation	No/negligible
Mast et al. (2011)	131	Judgment task: deception	Deception detection task	Small
Ambady et al. (1995)	90	Judgment task: decoding	Decoding skills	Various (small)
Ambady et al. (1995)	90	Judgment task: non-verbal	Non-verbal sensitivity	Various (small to medium)
Sanchez et al. (1996)	262	Memory	Behavior memory	Various (small)
Lewis (2002)	149	Memory	Behavior memory	Not avail. (dissertation abstract)
Wood et al. (2008)	194	Motivation	Accountability	Medium
Strupeck (2004)	*nr*[c]	Motivation	Accountability	Not avail. (dissertation abstract)
Brtek et al. (2002)	338	Motivation	Accountability	Various (small to medium)
Rosenbaum (1992)	579	Motivation	Accountability	Not avail. (dissertation abstract)
Craven (1988)	*nr*[c]	Motivation	Accuracy motivation	Not avail. (dissertation abstract)
Salvemini et al. (1993)	108	Motivation	Accuracy motivation	Medium
Borman (1979)	146	Motivation	Effort	No/negligible
Zalesny et al. (1992)	83	Perceptions of others	Teaching attitudes	Various (medium neg to large pos)
Lewis (2002)	149	Perceptions of process	Expectations	Not avail. (dissertation abstract)
Davis (1999)	82	Perceptions of self	Attributional complexity	No/negligible
Letzring (2008)	138	Perceptions of self	Attributional complexity	Small
Borman (1979)	146	Perceptions of self	Empathy	Small to medium
Borman et al. (1991)	79	Perceptions of self	Evaluative tendency	No/negligible
Powell (2008)	164	Perceptions of self	Interpersonal orientation	Small to medium

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Mast et al. (2011)	131	Perceptions of self	Rating efficacy	No/negligible
Mast et al. (2011)	131	Perceptions of self	Rating efficacy	Medium
Wood et al. (2008)	194	Perceptions of self	Rating efficacy	Medium to large
Borman (1979)	146	Perceptions of self	Self-competence	No/negligible
Borman (1979)	146	Perceptions of self	Social interest	Small to medium (negative)
Borman (1979)	146	Perceptions of self/others	Assumed similarity and affect	No/negligible
Letzring et al. (2006)	180	Perceptions of self/others	Acquaintance	Medium to large
Borman (1979)	146	Personality	Aggression	Small
Letzring (2008)	142	Personality	Big 5	Small to medium
Letzring (2008)	138	Personality	Big 5	Small to medium
Janovics (2003)	410	Personality	Big 5	Not avail. (dissertation abstract)
Gibson (2006)	*nr*[c]	Personality	Big 5	Not avail. (dissertation abstract)
Christiansen et al. (2005)	122	Personality	Big 5	Small to medium
Powell (2008)	164	Personality	Big 5	Various (small to medium)
Davis (1999)	82	Personality	Conscientiousness	Not avail. (dissertation abstract)
Borman (1979)	146	Personality	Detail orientation	Small to medium
Borman et al. (1991)	79	Personality	Detail orientation	Small
Letzring (2008)	138	Personality	Interpersonal problems	Small to medium
Lippa et al. (2000)	109	Personality	Masculinity/femininity	Small
Letzring (2008)	138	Personality	Narcissism	Small to medium
Davis (1999)	82	Personality	Need to evaluate	Not avail. (dissertation abstract)
Gibson (2006)	*nr*	Personality	Need to evaluate	Not avail. (dissertation abstract)
Borman et al. (1991)	79	Personality	Personal adjustment	Small
Lippa et al. (2000)	109	Personality	Personality (Big 5)	Various (up to small to medium)
Vogt et al. (2003)	102	Personality	Psychological communion	Medium
Letzring (2008)	138	Personality	Psychological well-being	Various (small to medium)
Human et al. (2011)	380	Personality	Psychological well-being and	Various

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Borman (1979)	146	Personality	Self-control	Small to medium
Borman et al. (1991)	79	Personality	Self-control	No/negligible
Borman (1979)	146	Personality	Self-monitoring	No/negligible
Davis (1999)	82	Personality	Self-monitoring	Not avail. (dissertation abstract)
Borman (1979)	146	Personality	Sociability	No/negligible
Borman (1979)	146	Personality	Tolerance	Small
Borman (1979)	146	Personality	Various traits	Various (up to small to medium)
Letzring (2008)	138	Personality	Various traits	Small-to-medium
Hjelle (1969)	72	Personality	Various traits	Various (small to large)
Ambady et al. (1995)	90	Personality	Various traits	Various (small to medium)
Adams (1927)	80	Personality	Various traits	Questionable method used
Borman (1979)	146	Personality	Various traits	Medium
Borman (1979)	146	Personality	Various traits	Various (negligible to medium)

Note. At least $N = 126$ distinguishable individual effects in $k = 48$ reported studies. The actual number of effects are much 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. [b]We used Cohen's (1988) guidelines to interpret effect sizes (*r*), i.e. no/trivial (.00), small (.10), medium (.30) and large (.50) effects. An effect-size interval of .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 is not reported for some studies because it may not have been available (for instance, when results were drawn from dissertation abstracts and the original dissertation could not be sourced). More information on these studies may be requested from the first author.

Table 2.2

20 Questions about Individual Differences of Judgment Accuracy in HRM

Category	Characteristic	Research Question (RQ)
Cognitive	General intelligence	RQ1. Is the relationship between intelligence and rating accuracy non-linear (as opposed to linear) such that accuracy increases with intelligence, but decreases at high levels of intelligence? RQ2. Does interview structure and situational complexity moderate the effect of intelligence on accuracy? RQ3. Is intelligence more important for accurately judging certain dimensions rather than other dimensions (e.g. personality, interview competencies and assessment centre dimensions)? RQ4. Is intelligence more important for judging complex stimuli (e.g. live people) in complex situations (assessment center tasks where different situation are activated) than simpler stimuli (e.g. videos or 'paper people') in simpler situations (one-on-one interviews)?
	Dispositional reasoning	RQ5. What is the nomological placement of dispositional reasoning relative to emotional and social intelligence, and what is their relative importance in predicting judgment accuracy in personnel selection? RQ6. Can dispositional reasoning be developed with training and, if so, what is the best training method to develop it? 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) and for judging different target constructs (e.g. personality, dimensions)?

Category	Characteristic	Research Question (RQ)
	Behavior memory	RQ8. What is the comparative predictive validity of impression-memory (i.e. memory of a dispositional or trait inference) versus behavior memory (i.e. memory of an observed behavior) to predict judgment accuracy?
	Cognitive style/heuristics	RQ9. Do cognitive style and heuristics predict judgment accuracy in personnel selection ratings (as they have in performance rating studies)?
	Cognitive complexity	RQ10. How would ability-based measures of cognitive complexity predict rating accuracy as compared to self-report measures of cognitive complexity?
	Attributional complexity	RQ11. Would 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 rater's ability to elicit cues predict their judgment accuracy? RQ14. Is the increased use of interviewers' behavior prompts in interviews related to higher cue availability and eventual judgment accuracy?
	Motivation	RQ15. How do assessors' levels of accuracy motivation affect their judgment accuracy? Does it occur through enhanced attentiveness to cues, through better cue utilization, or to both judgment processes?

Category	Characteristic	Research Question (RQ)
Specific characteristics	Behavior observation	RQ16. Would innovative measures of behavior observation ability (that are not contaminated with behavior memory) predict rating accuracy in conjunction with measures of behavior memory?
	Personality trait chronic accessibility	RQ17. Would assessors' personality trait chronic accessibility for various Big Five traits predict their trait judgment accuracy?
	Rater gender	RQ18. Does rater gender affect judgment accuracy in personnel selection judgments (e.g. interviews and ACs)? If so, is the effect due to the mediating role of gender-related constructs, such as interpersonal sensitivity or masculinity/femininity?
		RQ19. Are gender differences in judgment accuracy moderated by target dimensions or traits, that is, are men and women different in their abilities to judge certain constructs?
	Culture and ethnicity	RQ20. Can assessors' cultural background characteristics (e.g. individualism) influence their judgment accuracy?

Chapter 3

The internal factor structure of dispositional reasoning: A comparison between managers and psychology students [18]

In this chapter, we assess the internal construct validity of a revised measure of dispositional reasoning. Although earlier studies used dispositional reasoning as a global construct, we test whether it consists of empirically distinct components, namely trait induction, trait extrapolation and trait contextualization. Using confirmatory factor analysis and structural equation modelling, competing models are tested to determine whether the latent structure of dispositional reasoning is better represented as (a) a single general construct, (b) three correlated components, or last, (c) a higher-order model, combining a broad factor (at the second-order) that causes variance in measures of three components at the first level. Results from a mixed validation sample (N = 321) of managers and psychology students showed that dispositional reasoning is well represented as componential in nature, with a higher-order construct underlying three lower-order components. A comparison of managers (n = 160) and psychology students (n = 161) through measurement invariance analysis showed relatively similar factor structures for dispositional reasoning between these groups, but metric invariance was not achieved in the samples we compared. Together, findings support the construct validity and limited measurement invariance of the revised dispositional reasoning measure to measure assessor' dispositional reasoning components.

[18] This chapter is in preparation for publication review as:
De Kock, F. S., Lievens, F., & Born, M. Ph. (2015). *The internal factor structure of dispositional reasoning: A comparison between managers and psychology students*. Manuscript in preparation.

3.1 Introduction

The characteristics of the good judge have intrigued researchers and practitioners for a long time (e.g., see Adams, 1927; Funder, 2012; Hall, et al., 2015; Taft, 1955; Vernon, 1933). Recent efforts to explain individual differences in judgment accuracy have shifted their focus to specific abilities as predictors. By focusing on specific abilities (as opposed to more generic constructs, such as general mental ability) we may achieve a better understanding of why and how accurate judgments are shaped. It is thought that specific schemas, abilities or knowledge structures may enable distinct processes involved in behavioral cue utilization (Funder, 1995) and broader dispositional inference (e.g., Gilbert, 1989).

In a recent study, Christiansen, et al. (2005) showed that dispositional reasoning, defined as complex knowledge of traits, behaviors and situations' potential to elicit traits into manifest behaviors, can be an important factor in producing accurate interviewer ratings. In their investigation, it was the strongest predictor of accuracy among a set of assessor individual differences that included demographics, personality and general cognitive ability. As such, it deserves more research attention.

A drawback of these earlier approaches (e.g., Christiansen, et al., 2005; Powell & Goffin, 2009) was that they tested dispositional reasoning as a broad ability, even though it was initially conceptualized as a broad set of conceptually distinguishable components[19]. The facets of dispositional reasoning are behavior-trait knowledge, judges' implicit personality theories, and judges' understanding of situation-trait relevance (Christiansen et al., 2005). However, studies up to now have not been able to measure these components reliably. Instead, they collapsed scores from the three conceptually distinct components into a broad measure. Consequently, the underling facets lie obscured from measurement and this may inhibit their further use in research and practice.

It may be useful to researchers and practitioners to expand our knowledge of dispositional reasoning by providing details of its internal composition and internal construct validity. For instance, the different components may play a meaningful role in advancing understanding of what makes the 'good judge' in personnel selection. Further, a componential view may enable stronger tests of judgment theories (e.g., Gilbert, 1989; Trope, 1986) that involve distinguishable judgment processes which, in turn, suggest the operation of different assessor constructs. Practically, understanding its internal configuration could also affect how we use the measure. For example, distinguishable components dictate that subtest scores may be utilized to select or train assessors. Also, assessor training interventions may be

[19] In Chapter 4 of this dissertation, we define these components as trait induction, trait extrapolation, and trait contextualisation.

tailored to target specific components. However, if they are indistinguishable from one another, the use of subtest scores is less justifiable.

The Present Study

In the present study, we test alternative hypothesized factor structures for dispositional reasoning. In so doing, it allows us to accomplish three important objectives. First, we provide answers to the questions about the internal composition of the dispositional reasoning construct. Is it a single, broad ability, or many specific abilities that are related? Or, could it be considered both? In other words, would a hierarchical factor structure make more sense (i.e. one where a broad dispositional reasoning ability influences specific components?). Second, we make practical recommendations about the best way to use the dispositional reasoning measure: Should it be used as an overall measure, or rather, can the individual subtests be used to reliably and validly assess the subcomponents of induction, extrapolation and contextualization. Finally, we compare the internal construct validity of dispositional reasoning between managers and psychology students – both groups provide pools of assessor that are often trained and used in workplace assessments (Krause & Thornton, 2009). In addition to the internal composition of dispositional reasoning, we also need to determine whether the revised instrument can measure the constructs in the same way between different assessor groups, that is, would it show measurement invariance? If the issue of measurement (Millsap, 2011) invariance between managers and students is not addressed, between-group comparisons of test scores may be misleading, because we would not be sure if observed group differences are 'real' or confounded with differences in the structure of the constructs and/or functioning of the measurement scales (Cheung, 2008). As such, the present study determines to what extent the revised dispositional reasoning instrument measures the same constructs for managers compared to students?

Competing Models of Dispositional Reasoning

Model 1: General-Factor Model

One way to think of dispositional reasoning is the general-factor approach. The simplest way to represent dispositional reasoning is that it is a single, monolithic entity, where judges' knowledge, abilities and understanding of information that relate to behaviors, traits and situations, join together as a whole. An example of general factor models in other domains is in Spearman's (1904) earlier view that individuals' performance at one type of cognitive task tends to be comparable to her or his performance at other cognitive tasks, or '*g*-theory'. In other literatures, there are also examples of general factors, such as general affectivity (Cropanzano, Weiss, Hale, & Reb, 2003).

The conceptual basis for a general-factor model for dispositional reasoning lies in the possibility that procedural and declarative knowledge structures that relate to multiple domains, in this case, behaviors, traits, and situations, are encapsulated in a

single broad factor. Applied in the present context, an example of a general factor model would be where items that measure knowledge and understanding of traits overlap with knowledge and understanding of trait expression in situations, for example. If dispositional reasoning is a broad factor that combines multiple, interrelated domains, then we should also observe good fit for a model with dispositional reasoning items that all load onto a single, underlying broad dimension.

In a general factor model of dispositional reasoning (see Model M1 in Figure 3.1), a broad dispositional reasoning latent construct is proposed that causes variance in all observed variables irrespective of the component that a specific indicator variable was initially designed to measure. Conceptually, this model assumes no distinction between separate dispositional reasoning components. If this model showed the best fit, a single score on overall dispositional reasoning would be the most appropriate operationalisation of the construct. This general-factor model will be treated as the baseline model for further model comparisons.

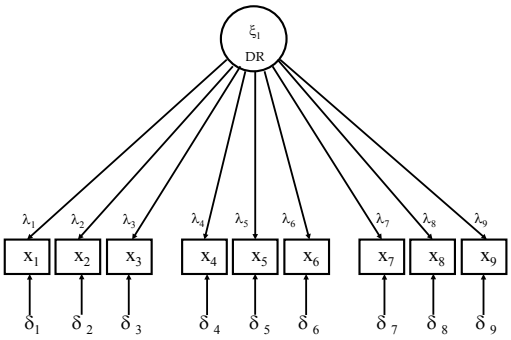

Figure 3.1. A confirmatory factor analysis of the structure of dispositional reasoning: A general factor model (Model M1). Only nine indicator variables are used in this example.

Model 2: Three Components (1st Order)

In contrast to a broad general factor, we might think of dispositional reasoning as a more complex configuration of separate narrow components. In such a componential model, specific abilities related to dealing with traits, behaviors and situations cluster tidily into three facets (Christiansen, et al., 2005). Conceptually, dispositional reasoning has been defined to consist of three distinct, yet related components, *trait induction* (the ability to know how traits manifest themselves in behavior), *trait extrapolation* (an understanding of how traits and their behavioral manifestations naturally co-vary), and *trait contextualization* (the ability to identify situations that are relevant to different traits) (see Figure 3.2) (De Kock, et al., 2015). In a componential view, judges utilize separate constructs to understand others' behavior cues, not

only broad dispositional reasoning. These may operate relatively independently in the dispositional inference process.

Componential views of constructs are widely encountered in the IO-psychology literature. For example, many intelligence theories make provision for the existence of specific forms of intelligence (e.g., Sternberg, 2000) that operate within, and sometimes beyond, the general factors. Another example is emotional intelligence (Mayer, et al., 1999) as one of multiple specific intelligences (Gardner, 1993).

A componential interpretation of the dispositional reasoning construct space would be where test items measure a distinct component only – some tap into induction, whereas others tap into extrapolation. Items do not, however, measure facets other than those to which they are most closely conceptually related. If dispositional reasoning is componential in nature, then we should also observe good fit for a model with items that load onto three separate dimensions, with no cross-loadings allowed. In a three-component model (see Model M2 in Figure 3.2), dispositional reasoning is represented as consisting of three distinct facets. From this perspective, the model specification for the latent structure of dispositional reasoning suggests that trait induction, trait extrapolation and trait contextualization represent related but distinct facets.

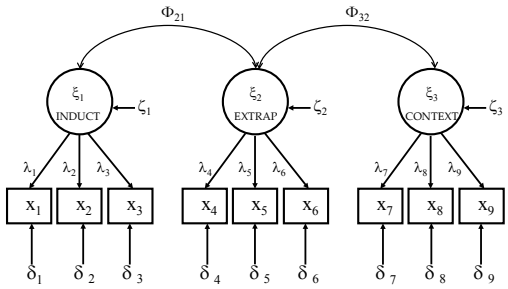

Figure 3.2. A confirmatory factor analysis of the structure of dispositional reasoning: A three-component (first-order) model (Model M2). Only nine indicator variables are used in this example.

Model 3: Hierarchical Factors (2nd Order)

We have argued that dispositional reasoning could exhibit a latent structure that supports either a general (*g*-approach) or specific (s_n approach) ability approach. It is important to consider the possibility where these views are not mutually exclusive, that is, its latent structure may be both general *and* specific.

Neither general, nor specific models of dispositional reasoning provide parsimonious conceptualisations of it as a construct. On the one hand, a general-

factor model is crude and inelegant, as the components are conceptually meaningful. On the other hand, a specific model lacks broader meaning – it does not recognize that the components cluster around a central theme as a constellation of constructs that enable dispositional inference. In other words, a component-only view sees induction, extrapolation and contextualization as isolated but not part of a whole.

The conceptual basis for the hierarchical model of dispositional reasoning lies in theories and empirical evidence that suggest the operation of higher-order constructs, for example the Big Five personality traits (McCrae & Costa, 1997) that subsume narrow facets. Also, in the intelligence literature, the early general (*g*) versus specific (s_n) intelligence debate has largely given way to a consensus view of the hierarchical nature of abilities (see Carroll, 2003, for a review) where broad factors at a higher stratum affect narrow factors at lower strata.

A hierarchical structure for dispositional reasoning would suggest that a broader dispositional reasoning latent construct causes variance in the more specific components of induction, extrapolation, and contextualization. A conceptual model for a hierarchical factor structure proposes that the higher-order factor of dispositional reasoning has direct effects on the three lower-order factors (see Model M3 in Figure 3.3). Here, the latent structure of dispositional reasoning is characterized by three first-order factors that represent distinctive facets, but all are 'caused' by a single, second-order latent construct.

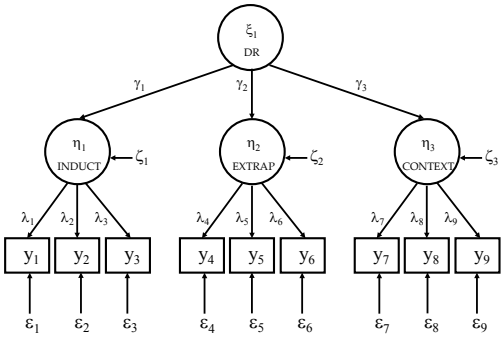

Figure 3.3. A confirmatory factor analysis of the structure of dispositional reasoning: Hierarchical (2nd order) model (Model M3). Only nine indicator variables are used in this example.

Measurement Invariance

Users of the dispositional reasoning measure may want to administer the measure in different populations of assessors. Therefore, the issue of measurement invariance needs to be addressed in order to determine whether "an assessment instrument is measuring the same constructs in exactly the same way across groups" (Byrne & Stewart, 2006, p. 287). The analysis of measurement invariance has become popular

in HRM (Schmitt & Kuljanin, 2008) as a means to establish measurement properties of a measure when it is used between two or more groups.

3.2 Method

Participants and Procedure

Table 1 shows the distributions of participants that formed part of the study sample. For our study, it was important to limit the sample to participants that may realistically form part of a broader population of assessors. A combined sample of managers and psychology students were selected as these are the people who are most likely to be trained as assessors (in interviews, ACs, or performance rating)(Krause & Thornton, 2009).

Combined Sample

There were 321 respondents in the combined sample (54.4% females and 45.6% males). In terms of race, the sample comprised of 46.3% Black African, 35.8% White, 11.1% Mixed Race and 5.9% Asian/Indian participants. Their mean age was 32.72 (*SD* = 11.13) years. Some (59.6%) had tertiary qualifications, although the student sample skewed this statistic – 59.5% of psychology students were postgraduates, and the rest were Bachelor's students. A relatively small proportion (16%) of managers had Bachelor's degrees or higher, and 27.8% held a three-year vocational training diploma. The prevalent first languages amongst these respondents were English (40.8%) and Afrikaans (19%), with the remainder speaking indigenous African languages or others. English was the official workplace language of all participants.

Group 1: Managers

We recruited 160 personnel that consisted of managers of various line and staff functions from two different organizations[20]. All of these respondents were undergoing training when they were assessed.

Group 2: Psychology Students

Our second group consisted of 161 psychology students at various levels of academic seniority. Most of the students (59.5%) were postgraduates studying Industrial-Organizational Psychology, while the rest were Bachelor's students in the same field.

A further comparison of the two samples revealed that managers were generally older (*M* = 42.3 yrs, *SD* = 6.7 yrs) than psychology students (*M* = 22.8 yrs, *SD* = 3.5 yrs), $t(221.02) = 31.142$, $p < .001$. The samples differed in terms of ethnic composition, as managers were predominantly African (71.4%), as compared to students whom were mostly White (55.6%).

[20] Some (*n* = 146) of the manager respondents were included in another study contained within another chapter in this dissertation (see Chapter 4).

Procedure

The data collection was completed in multiple sessions in two organizations and two universities in South Africa. After introducing the research as part of assessor training, we explained participants' rights and requested their informed consent. Next, we explained the research materials before participants independently completed the research questionnaire. Last, participants were debriefed and thanked for their participation. Data collection across the respective research sites was completed between 2011 and 2015.

Measures

Dispositional Reasoning

To measure the dispositional reasoning components, we used the Revised Interpersonal Judgment Inventory (IJI)(De Kock, et al., 2015). The revision of the original IJI (Christiansen et al., 2005) is thoroughly described in Chapter 4 of this dissertation. The Revised IJI consisted of 64 items that measure three distinct components. An example item for each scale can be found in the Appendix.

Induction. The contextualization component of dispositional reasoning was measured by 20 items that tapped candidates' ability to make correct behavior-trait inferences. After describing the Big Five personality traits, a list of adjectives from Goldberg's (1992) factor markers were presented. The task was to identify the traits (e.g. conscientiousness) that best matched the marker adjectives (e.g. thorough).

Extrapolation. The extrapolation component of dispositional reasoning was measured by 23 items assessing a respondent's understanding of how traits and behaviors co-occur. Items described a fictional person in terms of traits and behaviors and required respondents to select which of four descriptions was most (or least) likely also true of the person.

Contextualization. The contextualization component of dispositional reasoning was measured by 21 items that test understanding of trait–situation relevance. On the basis of empirical results (Tett & Guterman, 2000) one response option for each item was keyed as being the most consistent with empirical evidence, theoretical relationships and expert judgment. One subset of items presented a trait description, e.g. 'empathy' by listing examples of behaviors associated with high and low scorers on the trait. Next, respondents had to choose which of five situations would most likely elicit the relevant behavior. The second subset of items reversed the direction of inference.

Measurement properties for the instrument in the present study are reported in the results section of the present chapter. In a recent study (De Kock, et al., 2015) conducted with a sample of managerial staff the measure showed acceptable CFA-derived construct reliabilities for induction (.77), extrapolation (.81) and trait contextualization (.76).

Biographical Characteristics

To enable normative comparisons, we also drew respondents' biographical detail from the database of respondents that had completed the dispositional reasoning measure.

Statistical Analysis

In order to evaluate the latent structure of the revised dispositional reasoning measure, we conducted both lower-order and higher-order confirmatory factor analysis (HCFA). First-order CFA was used to assess the measurement model fit of both the global factor (M1) and three-component (M2) models.

Consequently, HCFA was used to evaluate the higher-order model (M3). Hierarchical factor analysis is often used for theory testing (Kline, 2011), for example, when it is believed that specialized facets of intelligence (e.g. verbal reasoning, memory) are influenced by a broader dimension of intelligence (*g*). In higher-order factor analysis, the factor correlations at a lower level (e.g. between specialized facets of a broader construct) become the input matrix for the higher-order factor analysis. As such, the HCFA attempts to provide a more parsimonious account for the inter-correlations among lower-order factors (Brown, 2015).

Robust maximum likelihood (RML) estimation was employed to estimate all models, unless stated otherwise. We used a number of fit indices to evaluate[21] model fit, including SBχ^2 (Satorra & Bentler, 1988), CFI, RMSEA (and its 90% confidence intervals), and SRMR. Our analyses were conducted with Lisrel 9.2 (Jöreskog & Sörbom, 2015).

Data Preparation for HCFA

Before we conducted the higher-order confirmatory factor analysis, we addressed a number of statistical issues.

Item-to-Sample Size Ratio. Our complete measure had 64 individual items. We decided not to conduct HCFA of the measurement model on *item-level* data in this study because the number of parameters to be estimated in a model with 64 observed variables – one for each item – would have led to inadequate statistical power (MacCallum, Browne, & Sugawara, 1996; Wolf, Harrington, Clark, & Miller, 2013). The issue would also have applied for separate analysis within each of our two subsamples. As such, we shortened the scales to allow for sufficient power and ensure appropriate model identification – issues that were important for the subsequent hierarchical model analyses. To abbreviate the scales, we considered the issues associated with shortening scales (e.g., Chongming Yang, Nay, & Hoyle,

[21] As recommended by Berne and Stewart (2006), the following minimum cut-offs were applied to infer acceptable model fit: SBχ^2 (Satorra & Bentler, 1988) with $p > .05$; CFI > .95; RMSEA < .08; and SRMR < .08.

2010). We decided to create four indicator variables for each first-order latent variable by using parcels of items within each scale as manifest variables. We used the procedures outlined by Little, Cunningham, Shahar, and Widaman (2002) when creating parcels and explain our parcelling strategy in detail in Appendix A. Using parcels in CFA has distinct advantages[22]. Not only do they allow retaining measurement information from many items, but in most conditions, less biased parameter estimates are encountered when parcels are used (Hair, Black, Babin, & Anderson, 2010).

Model Specification. The hierarchical CFA model (see Fig 3.3) hypothesizes[23] for both managers and psychology students the following: (a) a dispositional reasoning structure is best represented by a three-factor second-order model comprising of a single higher-order factor of dispositional reasoning and three lower order factors of induction, extrapolation, and contextualization; (b) each observed variable (i.e., parcel) would have a non-zero loading on the lower order factor it was intended to measure and zero loadings on other factors (i.e., zero cross-loadings); (c) covariation among the three lower-order factors would be explained by the higher-order factor of dispositional reasoning; (d) measurement error terms would be uncorrelated; and (e) factor disturbances would be uncorrelated.

Model Identification. A further statistical issue, latent variable scaling, was addressed before the Higher-order Confirmatory Factor Analysis was conducted. To identify a hierarchical CFA model, it must have at least three first-order factors, and the latter should have at least two indicators each (Kline, 2011). The hierarchical model (M3) (see Figure 3.3) that we hypothesized, satisfies both these requirements: Our model has three first-order factors and five indicator variables for each first-order factor. However, the second-order portion of the model has additional identification requirements as it must be identified in itself. For instance, a solution that specifies a single second-order factor over three first-order factors is just-identified (Brown, 2015). As such, the residuals of induction and extrapolation were constrained to be equal in order to achieve identification at the higher-order level of the model. To accomplish this, we used a procedure outlined by Byrne (2011), explained more fully in the results section.

Latent Variable Scaling. In addition to adequate model identification, it was necessary to scale the second-order factor of dispositional reasoning in the model. The second-order factor has no observed measures and must be provided a metric (Brown, 2015). This is accomplished either through fixing any one of the direct effects of dispositional reasoning on the first-order factors, or to fix the variance of dispositional reasoning to 1.0 (i.e. standardize it). In a single-sample

[22] However, it should be cautioned that combining items into parcels is a controversial issue as it may artificially enhance the reliability estimates of scores from the measure (Hair et al., 2012).
[23] In line with Byrne and Stewart (2006).

analysis, either method of scaling is appropriate (Kline, 2011). As the latter approach leaves all three direct effects of dispositional reasoning on the first-order factors as free parameters – these factor intercorrelations each have substantive importance in the present study – we scaled the dispositional reasoning factor by fixing its variance to 1.0. Because the indicator variables are now endogenous variables, their residual variances (disturbances) were estimated using the psi matrix (as opposed to estimation of factor variances, using the phi matrix). As such, our analysis specified that psi is a diagonal matrix; that is, freely estimate residual variances and fix all off-diagonal relationships-residual covariances to zero (Brown, 2015).

Higher-order CFA Procedure

After completing the data preparation, we followed the general sequence of higher-order confirmatory factor analysis (HCFA) proposed by (Brown, 2015): (1) develop a 'well-behaved' first-order CFA solution, in other words, one that fits well and is conceptually valid; (2) examine the magnitude and pattern of correlations among factors in the first-order model; and (3) fit the second-order model, based on conceptual and empirical grounds. These steps are described, along with the results, in the results section.

Measurement Invariance (MI)

Finally, we also conducted measurement invariance (Millsap, 2011) analysis of the best fitting factor model between managers and psychology student samples. If the issue of measurement equivalence is not addressed, between-group comparisons of test scores may be misleading, because we would not be sure if observed group differences are 'real' or confounded with differences in the structure of the constructs and/or functioning of the measurement scales (Cheung, 2008). In order to establish the measurement invariance of the first-order models of the factor structure underlying our measure of dispositional reasoning between managers and psychology students, we followed general guidelines for testing measurement invariance (MI) proposed by a number of authors (e.g., Brown, 2015; Millsap, 2011; Raykov, Marcoulides, & Li, 2012; Vandenberg & Lance, 2000). Our strategy involved an assessment of measurement invariance of the first-order model (M2) of dispositional reasoning hypothesized as three components (induction, extrapolation, and contextualization).

To assess invariance of the hierarchical model, we relied on guidelines proposed by Byrne and Stewart (2006), as well as others (Chen, Sousa, & West, 2005; Cheung, 2008). Our testing strategy involved a number of hierarchical steps. To test MI, we compared the fit of a series of more constrained models unconstrained model with the fit of the equivalence constraints with the Likelihood Ratio (LR) test (Tabachnick & Fidell, 2013). The LR test involves a comparison of the χ^2-values of the unconstrained and constrained models. A statistically significant increase in χ^2 as a result of constraining a specific set of parameters was used as a criterion for rejecting measurement invariance.

3.3 Results

Descriptive Statistics

Figure 3.4 and Table 3.1 portrays the mean dispositional reasoning scores (overall, and by component) for managers and psychology students. Results from an independent samples *t* test indicated that psychology students (M = .76, SD = .10, N = 161) scored relatively higher on overall dispositional reasoning than managers (M = .45, SD = .14, N = 161), $t(287.8) = -22.2$, $p < .001$, two-tailed. The difference of .31 scale points was large (scale range: 0 to 100%; d = 2.55, large effect size r = .79), and the 95% confidence interval around the difference between the group means was relatively precise (33.7 to 28.2). For the sake of brevity, the mean comparisons for component scores are not reported, however. They too were all statistically significant, $p < .001$. Table 3.2 reports the intercorrelation (uncorrected for unreliability) between the dispositional reasoning component scores between the two subsamples. As background to the CFA model evaluation, the item parcel means are also reported in Table 3.3 for each group.

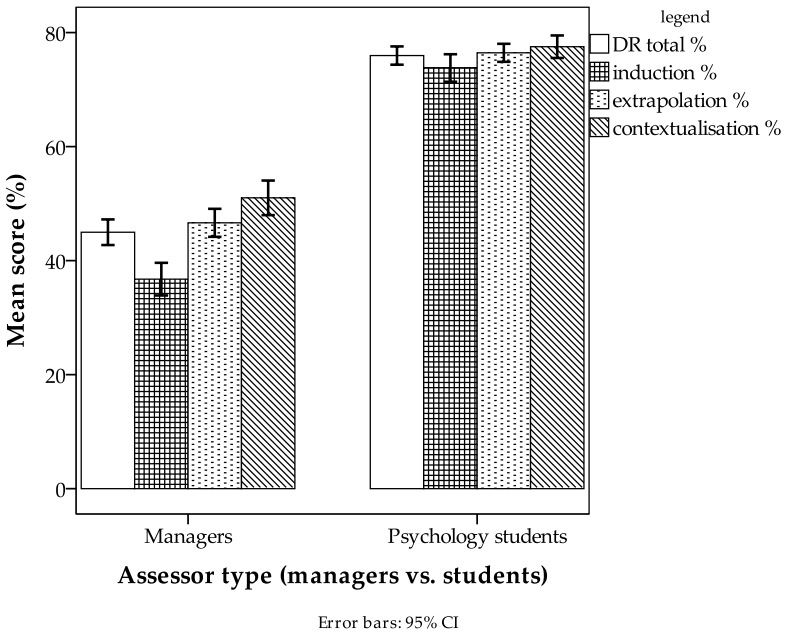

Figure 3.4. Comparison of mean scores (%) for dispositional reasoning and its components (induction, extrapolation, and contextualization) between managers and psychology students. The y-axis is interpreted as follows: 0% = no correct answers and 100% = all items correct.

Table 3.2
Descriptive Statistics and Intercorrelations for the Managers and Psychology-students Samples

Variables	Descriptives	1	2	3	4
	M (*SD*)	Correlations for managers (*n* = 160)			
1. Induction	.37 (.18)	-			
2. Extrapolation	.47 (.16)	.56**	-		
3. Contextualization	.51 (.19)	.47**	.44**	-	
4. Total DR	.44 (.15)	.83**	.80**	.80**	-
		Correlations for psychology-students (*n* = 161)			
1. Induction	.74 (.16)	-			
2. Extrapolation	.77 (.10)	.53**	-		
3. Contextualization	.77 (.13)	.53**	.36**	-	
4. Total DR	.76 (.10)	.87**	.74**	.79**	-

Note. Total *N* = 321. DR = Dispositional reasoning total scores.
* $p < .05$; ** $p < .01$ (two-tailed).

Table 3.3
Descriptive Statistics for Item Parcels for the Managers and Psychology-students Samples

Item Parcels	Managers (*N* = 160)	Students (*N* = 161)
	M (*SD*)	*M* (*SD*)
1. Induction 1	.54 (.24)	.76 (.18)
2. Induction 2	.31 (.26)	.77 (.24)
3. Induction 3	.33 (.27)	.78 (.22)
4. Induction 4	.31 (.30)	.75 (.24)
5. Induction 5	.35 (.26)	.63 (.27)
6. Extrapolation 1	.44 (.22)	.67 (.19)
7. Extrapolation 2	.38 (.21)	.73 (.17)
8. Extrapolation 3	.53 (.25)	.82 (.17)
9. Extrapolation 4	.56 (.25)	.82 (.18)
10. Extrapolation 5	.44 (.27)	.80 (.22)
11. Contextualization 1	.51 (.27)	.80 (.19)
12. Contextualization 2	.53 (.25)	.71 (.22)
13. Contextualization 3	.38 (.25)	.71 (.21)
14. Contextualization 4	.50 (.29)	.79 (.21)
15. Contextualization 5	.63 (.31)	.86 (.18)

Note. Total *N* = 321. Scores are average difficulty (p) values across items in each item parcel and, therefore, range from 0 (the mean score on items contained in the parcel was 0) to 1.00 (the mean score of participants to items contained in the parcel was 1).

Measurement Properties

Reliability

We analysed the internal consistency reliability of scores from the dispositional reasoning measure for the combined sample and separately for each subsample. Table 3.4 reports the internal consistency coefficients (Alpha coefficient) and item-analysis statistics for the overall dispositional reasoning measure, as well as for the separate subcomponent measures (induction, extrapolation, and contextualization). In the combined sample, the reliability of scores for the overall measure was excellent (α = .93) and also good for the component measures (.82 < α < .86). Table 3.4 further shows that, between the two samples that we studied, internal consistency was generally higher for managers as compared to the psychology student sample. The lower reliabilities for students may be due to substantial restriction of range in their dispositional reasoning scores. For example, on the extrapolation measure, psychology students had lower mean item variance (.14 vs .21) and scale variance (5.46 vs 13.09) as compared to the manager sample. Restriction of range in scores, or a so-called 'floor-and-ceiling' effect, is well-known to suppress internal consistency reliability, especially when item difficulties vary considerably (Nunnally & Bernstein, 1994). Furthermore, Cronbach's alpha may be a misestimator of the scale reliability if a multiple item measure when items are not essentially parallel (Raykov, 2012). As such, we believe that the internal consistency reliability of the revised dispositional reasoning measure may be underestimated in our existing data sample.

Unidimensionality

Prior to the HCFA, we fit three measurement models for dispositional reasoning on item-level data (for 64 individual items) first in order to establish overall dimensionality in the combined sample, as unidimensionality is an assumption underlying CFA (Kline, 2012). Robust diagonally weighted least squares (RDWLS) estimation was employed to estimate all models at item level as the data were not multivariate normal. Results revealed satisfactory evidence of unidimensionality within each subscale.

Table 3.4

Internal Consistency Reliability Estimates of Scores for Dispositional Reasoning Measure[a]

Group	*N*	Full measure (64 items)	Component		
			Induction (20 items)	Extrapolation (23 items)	Contextualization (21 items)
Combined	321	.93	.86	.82	.82
Managers	160	.85	.74	.66	.75
Psychology students	161	.79	.69	.44[b]	.59[b]

Notes: *N* = 321, [a]As a conservative approach that avoids capitalization on chance, no items were removed on the basis of item analysis prior to the CFA analyses. [b]Lower reliabilities are likely due to substantial restriction of range for the psychology student sample. For extrapolation, as example, psychology students had lower mean item variance (.14 vs .21) and scale variance (5.46 vs 13.09) as compared to the manager sample.

Assessment of Models

General Factor Model (M1)

Model assessment was conducted by testing a series of confirmatory factor analytic models. The results of these tests are reported in Table 3.5 for the combined sample. Table 3.6 reports the results separately for managers and psychology students. The general factor model (M1) of dispositional was assessed by a first-order confirmatory factor analysis based on data from the combined sample. The analysis used fifteen item parcels created to represent sets of items of the revised interpersonal judgment inventory (De Kock, et al., 2015). The hypothesized model is presented in Figure 3.1 where circles represent latent variables, and rectangles represent measured variables. In this figure, a one-factor model of dispositional reasoning is portrayed. The item parcels serve as indicators of the general dispositional reasoning factor. The general factor model (M1) was tested and the fit was acceptable, $\chi^2(90, N = 321) = 191.50$, $p < .001$, Satorra-Bentler χ^2 (90, $N = 321$) = 180.99, $p < .001$, Robust CFI = .96, TLI = .95, RMSEA = .06, 90% CI: [0.05; 0.07], although the relative large chi-square statistic suggested the need for further model improvement.

Three-Component Model (M2)

Next, we proposed a three-component factor model of dispositional reasoning, with trait induction, trait extrapolation and trait induction as separate components (see Figure 3.2). The three factors were hypothesized to co-vary with one another and the respective item parcels created from each of the subscale items serve as indicators of the respective factors. A three-component model showed relatively good fit, χ^2 (87, $N = 321$) = 117.60, $p = .016$, Satorra-Bentler χ^2 (87, $N = 321$) = 113.29, $p < .05$, Robust CFI = .99, TLI = .98, RMSEA = .03, 90% CI: [0.015; 0.048]. All fifteen item parcels (three first-order latent variables with five item parcels each) were significant indicators of their respective latent factors. We inspected the results of the phi matrix providing the correlations among the latent variables, or factors. Consistent with our expectation, all factors were significantly interrelated (range of *z*s = 6.76 – 10.48). Factor intercorrelations (amongst the various subdimensions of the dispositional reasoning components, M2) were generally large ($.84 < \varphi < .95$) when disattennuated for measurement error. As such, the pattern of correlations speaks to the feasibility of the suggested second-order model, which posited that trait induction, trait extrapolation and trait contextualization are more specific dimensions of broad underlying dispositional reasoning.

Hierarchical Factor Model (M3)

Finally, a hierarchical (2^{nd}-order) factor model of dispositional reasoning – this model proposes a general component, influencing the three specific components of induction, extrapolation, and contextualization – was proposed. The hierarchical model was tested and support was found as the model showed good fit, χ^2 (87, $N = 321$) = 117.60, $p = .016$, Satorra-Bentler χ^2 (87, $N = 321$) = 113.29, $p < .05$, Robust CFI = .99, TLI = .98, RMSEA = .03, 90% CI: [0.015; 0.048]. As is evident, the goodness of fit

of the hierarchical model is the same as the three-component first-order model (M2) in which factors are allowed to co-vary freely. This is so because a solution that specifies a single second-order factor over three first-order factors is just-identified (Brown, 2015). Brown recommends that it is not appropriate to statistically compare M3 with M2: only when the higher-order model is over-identified, can the nested χ^2 be used to determine whether the specification in M3 produces a significant degradation in fit relative to the first-order solution. However, as the higher-order solution does not result in a decrease in model fit, it may be concluded that the model provides a good account for the correlations among the first-order factors. Despite being just-identified, the magnitude and statistical significance of the factor loadings in the higher-order part of the model may be meaningfully interpreted (Brown, 2015). In the completely standardized estimates from the solution, each of the first-order factors loads strongly on the second-order dispositional reasoning factor: induction (γ = .98) and extrapolation (γ = .96) loaded more strongly than contextualization (γ = .88). As such, dispositional reasoning as a higher-order factor accounted for substantial proportions of variance[24] in the individual components: induction 96% (1 - .04), extrapolation 91.5% (1- .085), and contextualization 77.1% (1 - .229).

HCFA Model Modification. To provide estimates for a model that would be over-identified, we modified the second-order model (M3) using a procedure outlined by (Byrne, 2011). In this procedure, value constraints were placed on both the variance of the second-order factors and the disturbance terms associated with the Induction and Extrapolation factors. In reviewing the goodness-of-fit statistics it was evident that imposing of constraints to this parameter resulted in substantial degradation of fit with respect to SBχ^2 (without equality constraint, SBχ^2 = 113.183, p = .031; with equality constraint, SBχ^2 = 157.898, p = .000), the noncentrality parameter (30.603 vs 77.295), as well as the ECVI (.572 vs .711). The modified model reflecting the equality constraint (imposed to achieve identification of the higher-order portion of the model) was re-estimated and the results are reported in Table 3.5 (as M4).

[24] The estimates in the psi matrix indicate the proportion of variance in the lower-order factors that is not explained by the second-order factor (Brown, 2015). As such, they are calculated as the difference between unity (1) and the completely standardized disturbances.

Table 3.5

Fit Indices for Factor Structure Models of Dispositional Reasoning Measure in Combined Sample[a]

Model	χ^2	*S-B*χ^2	*df*	*S-B*χ^2/ *df*	*NNFI/TLI*	*CFI*	*SRMR*	$p_{close\ fit}$	*RMSEA (CI)*
M1	191.50**	180.99**	90	2.01	0.95	0.96	0.041	.09	0.059 (0.048; 0.071)
M2	117.60*	113.29*	87	1.30	0.98	0.99	0.031	.98	0.033 (0.015; 0.048)
M3	117.60*	113.18*	87	1.30	0.98	0.99	0.031	.98	0.033 (0.015; 0.048)
M4	166.30**	98.45**	89	1.11	0.96	0.97	0.039	.38	0.052 (0.040; 0.064)

Notes. N = 321, [a]Models tested here use item parcels as indicator variables and not individual items. M1 = single-factor structure; M2 = three-factor structure (Christiansen et al., 2005); M3 = hierarchical 2nd-order factor structure (De Kock et al., 2015); M4 = hierarchical 2nd-order factor structure (De Kock et al., 2015) with equality constraint on two factor variances for identification; χ^2 = Normal Theory Weighted Least Square Chi-Square; S-Bχ^2, Satorra-Bentler Scaled Chi-square; *df* = Degrees of Freedom; NNFI, Non-Normed Fit Index, a.k.a. Tucker-Lewis index; CFI, Comparative Fit Index; SMSR, Standardized Root Mean Residual; $p_{close\ fit}$ = p value for close fit (RMSEA < .05); RMSEA, Root Mean Square Error of Approximation with 90% confidence interval.
*p < .05. **p < .01.

Table 3.6

Sample Comparison of Fit Indices for Alternative Factor Structure Models of Dispositional Reasoning

Model	Group	χ^2	S-Bχ^2	*df*	S-Bχ^2/ *df*	NNFI/TLI	CFI	SRMR	$p_{close\ fit}$	RMSEA (CI)
M1	Managers	167.228**	168.758**	90	1.88	0.83	0.85	0.071	.02	0.073 (0.056; 0.090)
	Students	111.205	106.555	90	1.18	0.92	0.94	0.060	.79	0.038 (0.000; 0.060)
M2	Managers	99.200	101.210	87	1.16	0.97	0.98	0.052	.91	0.030 (0.000; 0.054)
	Students	103.510	98.275	87	1.13	0.94	0.95	0.058	.85	0.034 (0.000; 0.057)
M3	Managers	99.200	101.210	87	1.16	0.97	0.98	0.052	.91	0.030 (0.000; 0.054)
	Students	103.510	98.275	87	1.13	0.94	0.95	0.058	.85	0.034 (0.000; 0.057)
M4	Managers	137.495**	139.892**	89	1.57	0.98	0.91	0.080	.23	0.058 (0.038; 0.077)
	Students	101.061	104.795	89	1.18	0.92	0.94	0.057	.79	0.038 (0.000; 0.060)

Notes. $N_{managers}$ = 160, $N_{students}$ = 161; M1 = single-factor structure; M2 = three-factor structure (Christiansen et al., 2005); M3 = hierarchical 2nd-order factor structure (De Kock et al., 2015); M4 = hierarchical 2nd-order factor structure (De Kock et al., 2015) with equality constraint on two factor variances for identification; χ^2 = Normal Theory Weighted Least Square Chi-Square; S-Bχ^2, Satorra-Bentler Scaled Chi-square; *df* = Degrees of Freedom; NNFI, Non-Normed Fit Index, a.k.a. Tucker-Lewis index; CFI, Comparative Fit Index; SMSR, Standardized Root Mean Residual; $p_{close\ fit}$ = *p* value for close fit (RMSEA < .05); RMSEA, Root Mean Square Error of Approximation with 90% confidence interval.

*$p < .05$. **$p < .01$.

Model Comparison

We compared the baseline model (general factor model, M1) with the comparison models. A chi-square difference test indicated that the nested model (M2) showed significantly poorer fit[25] compared to the baseline (M1) model, Satorra-Bentler $\chi^2_{diff}(3, N = 321) = 45.033$, $p < .001$. Therefore, the three-component model of dispositional reasoning fits significantly better than a general-factor model. As the fit for the hierarchical model (M3) are the same as those for M2, the same conclusion may be reached, that is, a higher-order model with dispositional reasoning as a general factor influencing induction, extrapolation, and contextualization, explains variance in test scores better than a general factor model that disregards a componential structure. The model fit strategy outlined above (for testing M1, M2, and M3) was repeated in each separate subsample and the results are reported in Table 3.6.

Measurement Invariance

To compare the factor structure and latent means of dispositional reasoning between managers and psychology students, we conducted measurement invariance analysis. The results of testing various forms of measurement invariance for the first-order constructs are reported in Table 3.7.[26] First, a baseline model was established in each group, followed by tests of equivalence across groups at each of several increasingly stringent levels of invariance.

First-order (M2) Invariance

Preliminary analyses. It is preferable to conduct multiple-groups CFA with relatively balanced sample sizes, as in the present study (managers: $N = 160$; students: $N = 161$). Prior to the CFAs, the data were screened to ensure their suitability for the ML estimator (i.e., normality, absence of multivariate outliers). The test for multivariate normality held in the manager sample ($\chi^2 = .702$, $p = .70$), but not in the student group ($\chi^2 = 50.546$, $p < .001$), nor when the two samples were combined ($\chi^2 = 36.00$, $p < .001$). As a result, the Robust ML estimator was used in estimation of all models and, therefore, all analyses are based on the Satorra-Bentler scaled statistic (SBχ^2; Satorra & Bentler, 1988). In line with the recommendations of Byrne and Stewart (2006) we rely on SBχ^2, as well as on CFI, the root mean square error of approximation (RMSEA), and SRMR to evaluate[27] all models. The first item parcel within each subscale was used as a marker indicator to define the metric of the latent variable.

---

[25] This figure was calculated using a macro (available from Bryant & Satorra, 2013) discussed in Bryant and Satorra (2012).

[26] The analytic strategy and reporting standard generally follows Brown (2015).

[27] The evaluation criteria we apply for each fit index are outlined in Byrne and Stewart (2006). Values that adhere to the following cut-offs indicate significant reduction in fit when comparing two nested models: (1) if corrected $\Delta SB\chi^2/\Delta df$ shows statistical significance; (2) $\Delta CFI > .01$; and (3) the root mean square error of approximation (RMSEA) > .08.

Testing for Baseline Models. As the estimation of baseline models involves no between-group constraints, the data were analysed separately for each group. Prior to conducting the multiple-groups CFA, we ensured that the suggested three-factor model is acceptable in both groups. As shown in Table 3.7, in both managers and psychology students in this data set, overall fit statistics for the three-factor solution are consistent with good model fit. On both groups, all freely estimated factor loadings are statistically significant (all *ps* < .01*).*

Testing for Configural Invariance. Configural invariance represents the observance of the same number of factors and factor loading pattern across groups – no parameter equality constraints are imposed[28]. As the baseline models are now fitted simultaneously in a multigroup evaluation, the criterion for configural invariance is that goodness-of-fit (GOF) should indicate a well-fitting model. The fit of the configural model provides the baseline against which all subsequent more constrained invariance models are compared. As such, we conducted the simultaneous analysis of equal form. As shown in Table 3.7, this solution provides an acceptable fit to the data. This solution serves as the baseline model for subsequent tests of measurement invariance and population heterogeneity.

Testing for Factor Loading Invariance. In this step, equality constraints are imposed for all freely estimated first-order factor loadings, except for three items fixed to 1.00 for the purposes of latent variable scaling. Invariance for this step holds if GOF is adequate and if there is minimal degradation in fit from the configural model. The analysis evaluates whether factor loadings (unstandardized) of the DR component indicators are equivalent in managers and psychology students. In our data, the equal factor loadings models had an overall good fit to the data, although it significantly degraded fit relative to the equal form solution, $\chi^2_{diff}(12) = 39.60$, $p < .001$. As this value is statistically significant, it suggests that the constraints of equal factor loadings in the restricted model do not hold, that is, the two models are not equivalent across the manager and psychology student groups (Byrne & Stewart, 2006). Because the constraint of equal factor loadings significantly degrades the fit of the solution, it can be concluded that the indicators do not evidence comparable relationships to the latent constructs of dispositional reasoning components in managers and psychology-students (Brown, 2015). As such, a unit change in the underlying latent variable is not associated with statistically equivalent change in the observed measures (item parcels) in both groups. A failure to demonstrate metric invariance (i.e., factor loadings are not equivalent across the two groups) was sufficient evidence to terminate the evaluation of further constraints. The results of further tests are reported in Table 3.7, however.

[28] For this model, as with subsequent tests in our invariance analysis where equality constraints are imposed on particular parameters, data for the two groups are analysed simultaneously in a file combining data for both groups to obtain estimates.

Table 3.7
Tests of Invariance of Dispositional Reasoning in Managers and Psychology-students

Model	χ^2	*df*	χ^2_{diff}	Δ*df*	RMSEA (90% CI)	Cfit	CFI	TLI	NFI	PNFI	NFI_{diff}
Single group solutions											
Managers (*n* = 160)	99.200	87			.030 (.000 - .054)	.91	.98	.97	.84	.70	
Psychology students (*n* = 161)	103.510	87			.034 (.000 - .057)	.85	.95	.94	.76	.63	
Measurement invariance											
Equal form (configural)	224.495	179			.040 (.020 - .055)	.85	.98	.97	.89	.76	
Equal factor loadings (weak)	264.102**	191	39.60**	12	.049 (.034 - .063)	.54	.96	.96	.87	.79	0.02
Equal indicator intercepts (scalar)	301.070**	203	76.58**	24	.055 (.041 - .068)	.26	.95	.95	.85	.83	0.04
Equal indicator error variances	344.114**	213	119.62**	34	.062 (.050 - .074)	.05	.93	.93	.83	.85	0.06
Equal factor variances	390.790**	216	166.29**	37	.071 (.060 - .082)	.00	.91	.91	.81	.83	0.08
Equal factor covariances	409.503**	219	185.01**	40	.074 (.063 - .085)	.00	.90	.90	.80	.84	0.09
Equal factor means	728.912**	222	504.42**	43	.119 (.110 - .129)	.00	.72	.74	.65	.68	0.24

Note. $N = 321$. χ^2_{diff}, nested χ^2 difference; RMSEA = root mean error of approximation; 90% CI = 90% confidence interval for RMSEA; CFit = test of close fit (probability of RMSEA ≤ .05); SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker-Lewis index; NFI = normed fit index; PFI = parsimonious fit index.
* $p < .05$. ** $p < .01$. *** $p < .001$.

3.4 Discussion

Main findings

In recent empirical studies (e.g., Christiansen, et al., 2005; Powell & Goffin, 2009) dispositional reasoning emerged as an important predictor of judgment accuracy, but its internal composition is still unclear. Does it have components, or is it better understood as a single global construct? Or would a hierarchical model be a better representation of the dispositional reasoning factor structure? In response to a need to shed light on the possible componential nature of dispositional reasoning, the present study investigated its components using confirmatory factor analysis. Specifically, competing factor structure models were compared to identify the best fitting model from a general factor model (M1), three-component model (M2) and a hierarchical model (M3). The stability of the factor structure was also assessed in two different samples of assessors (managers and psychology students).

Although previous studies of dispositional reasoning as a predictor of rating outcomes have treated it as a single broad ability, our analyses show that it is better understood as having three specific ability components – induction, extrapolation, and contextualization. Component-level analyses were not possible when using earlier versions of the Interpersonal Judgment Inventory (Christiansen et al., 2005) because the subcomponents could not be reliably measured. This drawback has hampered further research on the role that specific facets of dispositional reasoning may play in rating outcomes such as accuracy.

The results of the present study showed that the internal composition of dispositional reasoning follows in the footsteps of intelligence constructs (Carroll, 1993, 2003) because the study findings supported a general dispositional reasoning factor, at a higher stratum, that appears to drive three specific facets at a lower stratum. As it is common to observe 'positive manifold' (Horn & Cattell, 1966) among measures of ability, especially between those that fall within the same conceptual domain, the hierarchical structure we observed for dispositional reasoning mimics findings from the intelligence literature where a *g*-factor underlies more specific abilities. However, it is too early to categorize dispositional reasoning as an ability from our results alone – other tests are required strict conceptual and correlational criteria for a classic intelligence (for example, see the classical criteria for an intelligence measure suggested by Mayer, et al., 1999).

The superiority of both a three-component and hierarchical factor model of dispositional reasoning relative to a single global-factor approach seems to extend to different assessor-types. In both of the samples tested here, results showed better fit for componential- and stratum-models than for undifferentiated models. Dispositional reasoning scores demonstrated a hierarchical structure, ordered in two

strata (one general, and the other specific) in both managers and psychology students. A relatively common factor structure for dispositional reasoning in both samples is interesting because earlier studies have shown accuracy differences between these groups. For example, Lievens (2001) showed that managers provided significantly more accurate ratings than students, although managers distinguished less between the dimensions being rated. Their relatively lower ability to distinguish between dimensions may have been because of lower levels of dispositional reasoning overall, and not because their dispositional reasoning is constituted in a different way than psychology students, for arguments' sake.

However, our invariance tests showed that measurement invariance was not evident for the first-order model of dispositional reasoning between managers and students. If measurement invariance cannot be established, then the finding of a between-group mean difference in dispositional reasoning scores cannot be unambiguously interpreted as it is unclear whether score differences are due to true dispositional reasoning differences, or to different psychometric responses to the scale items (Cheung & Rensvold, 2002). Results showed a lack of metric invariance, that is, the factor loadings were not equivalent between managers and students. As the relationship between the observed variables (item responses) and the latent traits specified in the measurement model were not equally strong across the groups (Kline, 2011), it implies that differences may exist between managers and psychology students in how the dispositional reasoning components are manifested. Nevertheless, metric equivalence is required in order to make meaningful between-group comparisons of scores on the dispositional reasoning measure and, therefore, we urge for more research into the origins of item response differences between managers and psychology students.

Limitations

The present study had a few limitations that must be acknowledged. First, by grouping assessors into two relatively course categories (managers vs students) it may obscure other individual differences within these groups, such as gender and ethnicity. As our respondents were drawn from only a few organizations in a single country, more research is needed to see how stable the reported factor solutions for dispositional reasoning are to other settings and nationalities. While acknowledging the need for replications and investigations of generalizability, there is yet no compelling reason to suspect that factor structures may differ.

Second, respondents were tested in various settings and not at the same time. This was necessary because, by virtue of their vocation, managers are not easy to access in large numbers. As such, data collection was accomplished by accumulation to achieve sufficient sizes that would allow the CFA analyses. While our 'snowball' sampling strategy has limitations, we limited participation to individuals that typically receive training to conduct ratings in organizations.

Finally, the modest sample sizes used in the present study prohibited fitting some using item-level data and, instead, item parcels were used as indicator variables. We acknowledge the limitations with parcelling as a strategy (Little, et al., 2002) and tried to counter this limitation by fitting the measurement models first at the item level, albeit only in the larger combined sample. The effect of different parcelling strategies on the study's final results were tested and found negligible.

Implications for Theory

In addition to theoretical implications already outlined, our study may unlock better tests of judgment theories. For example, Gilbert's (1989) model of social inference holds that, after observed behavior is categorized and characterized, an important correction stage follows, where the initial dispositional inference is adjusted, given the situational constraints on behavior expression. This theory implies that different components of dispositional reasoning may be utilized at different judgment stages. Our evidence in support of components, and the possibility to measure them well, may encourage researchers to seek to understand which of these, and how, relate to judgment outcomes.

The present investigation has also shed some light on assessor-type effects on dispositional reasoning. Although results have shown that the general factor structure underlying dispositional reasoning may be relatively similar between assessor types (managers and students in the present case), other results pointed to potential differences. The descriptive statistics in our study showed that psychology students outperformed the managers that we tested by a substantial margin on the measure of dispositional reasoning. We speculate that, as compared to managers, psychology students may have better developed schemas that relate to understanding traits, behaviors, and situations, by virtue of their specialization area (i.e., the behavioral sciences). However, these results require further replication before a conclusion is warranted. More work is needed not only to determine if, and how, various assessor types differ in terms of dispositional reasoning and its components, but also, how these differences may ultimately affect rating quality.

Implications for Practice

We found general support for the internal construct validity of the revised dispositional reasoning measure (De Kock et al., 2015) in samples drawn from different populations of raters. As results showed that the components can be reliably and validly measured, practitioners are encouraged to consider using the brief subscales in their assessor training and selection programmes. However, more work in the form of criterion-related validity evidence is needed before we can recommend the measures for rater selection and training. Future studies could also explore the effect of shortening the component measures to make them more convenient to use.

By extension, the study findings open up new possibilities to conduct predictive studies (of judgment accuracy as criterion) using dispositional reasoning at a componential level. Support for both componential and hierarchical models of constructs suggest that they may be used as predictors of criteria at either lower- or higher levels of abstraction (Hair, et al., 2010). As such, human resource practitioners may use either overall dispositional reasoning scores, or component-level scores, depending on pragmatic considerations. If only a brief and general measure is required, items may be selected at random as long as sufficient measurement properties can be maintained (for example, using Spearman's prophecy formula, Nunnally & Bernstein, 1994). Alternatively, component measures may be used individually (i.e. in a 'stand-alone' fashion) in their current form, or shortened where they exhibit high reliability of scores in particular sample types. We urge caution, however, that users should not be surprized to find lower internal consistencies in highly homogenous samples that are range-restricted (Nunnally & Bernstein, 1994).

3.5 Conclusion

The present study explored the theoretical factor structure of dispositional reasoning – a promising predictor of judgment outcomes such as accuracy. Whereas earlier studies treated it as broad ability construct, we showed that dispositional reasoning is better represented by a componential model, one which contains three facets, namely trait induction, trait extrapolation and trait contextualization. Furthermore, broad dispositional reasoning may act as a higher-stratum ability that affects performance on its facets at a lower stratum, in line with stratum theories of intelligence. Finally, our study findings suggest that different assessor types (for example, managers and psychology students) may be relatively similar in how the overall factor structure of dispositional reasoning is configured. However, dispositional reasoning and its components may not manifest in the same way between managers and psychology students, as we failed to find evidence of metric invariance. In our study, the similarity between managers and students did not extend to overall levels of dispositional reasoning, however – psychology students appear to have much better developed knowledge structures pertaining to behaviors, traits, and situations, as compared to the managers that we tested. We hope to see more work on the additional research avenues opened up by our study findings.

Appendix A: Parcelling Strategy

Dimensionality Considerations

An appropriate parcelling strategy should be identified given the dimensionality of the factor structure underlying a set of item scores. Exploratory factor analysis[29] of our item-level data (using Principal Axis Factoring, with Oblimin rotation, considered appropriate for our data, as suggested by Tabachnick & Fidell, 2013) indicated possible multidimensionality within all three first-order factors, namely for induction, extrapolation, and contextualization. However, we also had to consider the possibility that multidimensionality within each component of dispositional reasoning may be due to statistical artefacts. For example, multiple dimensions may also be artificially created when items vary in terms of their difficulty levels. Even if various items measure the same construct, the resulting correlation coefficients between these items may be low if the response thresholds vary much (Lord & Novick, 1968). As a result, techniques that are based on correlations, such as factor analysis, may cause artefacts in the form of spurious 'difficulty factors' with little if any psychological meaning (Bernstein & Teng, 1989; Reise, Waller, & Comrey, 2000). Stated otherwise, it is possible that items with similar distributions may tend to form factors irrespective of their item content. The *p*-values of the 64 items in our combined dispositional reasoning measure varied (M_p = .61; SD_p = 17; Min_p = .20; Max_p = .93).

Although some authors (e.g., Bandalos & Finney, 2001) argue that parcelling should be reserved for conditions of uni-dimensionality, Little and colleagues (2002) suggest two specific strategies for parcelling items when item scores indicate a multidimensional factor structure. First, an *internal consistency* approach creates parcels that use the facets observed as grouping criteria. In this approach, items contained within a facet are clustered to form a combined item parcel, yielding internally consistent facets as manifest indicators of the higher stratum construct and keeping the multidimensional nature of the construct explicit. Second, the *domain-representative* approach is a method that creates parcels by joining items from different facets into combined item clusters. For example, a parcel would contain items from each facets identified through dimensionality analysis. So, each parcel reflects all of the facets present within a set of items – this solution accounts for the multidimensionality inherent in a set of items. The domain representation approach has shown to be superior in some studies (e.g., Kishton & Widaman, 1994). Finally, a *random item assignment* strategy may be used. We decided to utilize random item assignment as a parcelling strategy, as it recognizes the possibility that difficulty factors may cause spurious dimensions within each component of dispositional reasoning. We also ran the analyses using the two other parcelling strategies – the choice of parcelling strategy had no substantive effect on the final results.

[29] Due to space constraints, we do not report these results, although they are available from the first author upon request.

Chapter 4

An in-depth look at dispositional reasoning and interviewer judgment accuracy[30]

Dispositional reasoning is defined as general reasoning about traits, behaviors and situations. Although earlier accuracy studies found that it predicted interview judgment accuracy, they did not distinguish between its underlying components (i.e. trait induction, trait extrapolation, and trait contextualization). This drawback has hampered insight into the nature of this construct. Therefore, we use a componential approach to test whether it adheres to classic criteria for an intelligence. Results from 146 managerial interviewers who observed videotaped interviewees showed the dispositional reasoning components had positive manifold and predicted interview accuracy. Moreover, they demonstrated discriminant validity with personality and incremental validity over cognitive ability in predicting interview accuracy. Together, findings suggest that dispositional reasoning broadly adheres to the classic criteria for an intelligence.

[30] This chapter has been published as:
De Kock, F. S., Lievens, F., & Born, M. Ph. (2015). An in-depth look at dispositional reasoning and interviewer accuracy. *Human Performance*, *28*, 1-23. doi: 10.1080/08959285.2015.1021046
An earlier version of the study in this chapter was also presented at the 28th International Congress of Applied Psychology, Paris, France, July 2014.

4.1 Introduction

The characteristics of the 'good judge' have intrigued researchers and practitioners for a long time (Adams, 1927; Cronbach, 1955; Funder, 2012). An understanding of the constructs that influence accuracy could help us to select and train the most effective interviewers, assessors, and raters. Although a contemporary review of interview literature (Dipboye, et al., 2012) concludes that good judges[1] have, for example, higher general cognitive ability, the perennial search continues for other characteristics that might help identify accurate judges.

In recent years, the focus of individual differences research has shifted toward exploring specific abilities related to being a good judge in personnel selection (for example, see Letzring, 2008; McLarney-Vesotski, et al., 2011; Powell, 2008). On the basis of theories that suggest that judges' interpretation of behaviors, traits, and situations are intertwined (e.g., Trope, 1986), Christiansen, Wolcott-Burnam, Janovics, Burns, and Quirk (2005) introduced dispositional reasoning and defined it as complex knowledge of traits, behaviors, and situations' potential to elicit traits into manifest behaviors. According to the dispositional reasoning framework, judgment accuracy may depend on three components, labelled here as *trait induction* (the ability to know how traits manifest themselves in behavior), *trait extrapolation* (an understanding of how traits and their behavioral manifestations naturally co-vary), and *trait contextualization* (the ability to identify situations that are relevant to different traits) (see Figure 4.1).

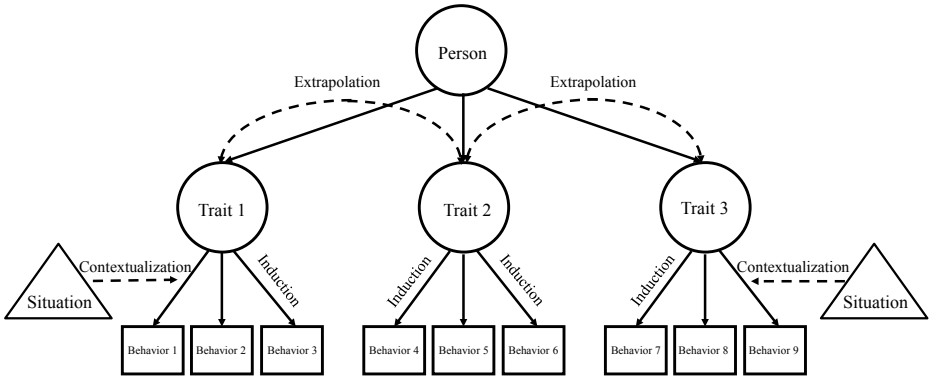

Figure 4.1. Understanding the components of dispositional reasoning: trait induction, trait extrapolation and trait contextualization.

According to Christiansen et al. (2005), the dispositional reasoning components are "declarative knowledge structures" (p. 126) that enable behavioral information processing. To test this notion, they asked students (N = 122) to watch videotaped segments of individuals responding to employment interview questions. Students judged the personality of the video interviewees and rated acquaintances who later completed self-report personality inventories. Overall, dispositional reasoning was the best predictor of interview accuracy (r = .42), among a set of rater individual differences. A follow-up study (Powell & Goffin, 2009) partially replicated these findings in a training context. In sum, these findings suggest that dispositional reasoning seems to facilitate interviewer accuracy.

However, the role of its subcomponents in accuracy is unclear. Prior studies did not consider its underlying components, as they could not measure them reliably. Instead, they collapsed the components into a single, broad measure. As a result, we know little about the componential nature of dispositional reasoning and how these components individually may facilitate interviewer accuracy.

In the present study, we test fine-grained hypotheses about the relation between the subcomponents of dispositional reasoning, other individual differences, and judgment accuracy. To this end, we develop a revised measure of dispositional reasoning – one with reliable components – and test whether the components to the criteria for classic intelligence measures.

It is important to determine whether dispositional reasoning represents an intelligence for both practical and theoretical reasons. In addition to expanding our insight into the specific rater constructs (see Jones & Born, 2008) that affect accuracy, it may also hold practical benefits. For example, the success of rater selection and

rater training may depend upon a valid model of the specific interviewer constructs that drive accuracy. In addition, if dispositional reasoning were further distinguishable into specific components, rather than a single, broad construct, it implies that we should target the specific components in interviewer training and selection.

4.2 Study Background

Criteria for an Intelligence

For it to be considered an intelligence, a specific mental ability must meet several criteria (Carroll, 1993; Flanagan, et al., 1997; Mayer, et al., 1999). First, a conceptual[2] criterion applies: The proposed intelligence should be operationalized as a set of mental abilities (that focuses on a specific concept, e.g. general memory and learning, Carroll, 1993), rather than preferences, interests or inclinations. Second, a correlational criterion holds that the abilities should form a related set, and be related to other intelligences. Although the intelligence measures must be related, they should still show unique variance, i.e. be empirically distinct. To this correlational criterion, we add the requirement that it should also predict related external criteria (for example, interviewer accuracy, in our study).

Positive Manifold

It is common to observe positive manifold (Horn & Cattell, 1966) among measures of ability, especially between those that fall within the same conceptual domain. There is considerable conceptual overlap among the dispositional reasoning components, because we need all three to utilize behavior cues effectively when constructing a mental picture of the interviewee (Christiansen et al., 2005). Information about traits, behaviors, and situational contexts are linked because they all represent trait relevant information (Kihlstrom & Hastie, 1997).

This conceptual overlap between the components should also manifest itself in empirical overlap. In the intellectual ability literature, for example, special intelligences (or narrow abilities) tend to co-vary with conceptually related abilities (Carroll, 1993). In a widely published study by McGrew, Werder, and Woodcock (1991) (as presented in Carroll, 2003) the mean correlation between sixteen narrow abilities was .37, indicating medium to large[3] positive manifold effects among narrow abilities.

In addition to showing internal positive manifold, the narrow dispositional reasoning components should also load moderately on external measures of general mental ability, which is typical of the relationship between narrow and general abilities (Neisser et al., 1996). We also expect similar effects between dispositional reasoning and general mental ability at the component level. Thus,

Hypothesis 1: The components of dispositional reasoning will show positive manifold internally, i.e. they will moderately ($r > .30$) relate to each other, as well as externally, that is, they will moderately ($r > .30$) relate to a measure of general mental ability.

Predicting Accuracy from Dispositional Reasoning Components

Trait induction and interviewer accuracy. Trait induction refers to the ability to know how traits manifest themselves in behaviors. An example of a person who has high trait induction would be someone who knows that an acquaintance who is talkative (i.e. observed behavior) is also most likely to be an extrovert (i.e. underlying trait) (Goldberg, 1992). In contrast, managers who have low trait induction are unlikely to infer that a person who continuously checks up on other colleagues is being *neurotic***.** A judge with high trait induction ability can therefore infer accurately which trait underlies (or drives) others' manifest behaviors (see Figure 4.1).

The theoretical origins of trait induction lie in trait theory. Trait theorists (e.g., Allport, 1937) view traits as habitual patterns of behavior, thought, and emotion which are relatively stable over time and influence behavior. On the basis of factor-analytic evidence, we know that particular clusters of behaviors reliably co-vary to form underlying traits (Cattell, 1965; Eysenck, 1970; McCrae & Costa, 1997). High levels of trait induction imply more accurate schemas of how behaviors actually cluster around traits.

In the past, various instruments were used to measure trait induction (see Cantor & Mischel, 1979; Hampson, John, & Goldberg, 1986; Klein, Loftus, Trafton, & Fuhrman, 1992; Maass, Colombo, Colombo, & Sherman, 2001). One popular measure, the behavior-trait knowledge subtest of the Interpersonal Judgment Inventory (Christiansen et al., 2005), consists of items that require the respondent to match a list of behaviors to corresponding Big Five general personality traits. The correct responses for each item were generated from large-scale empirical evidence of actual behavior-trait links (e.g., Goldberg, 1992).

There are many reasons to expect that trait induction would predict judgment accuracy. People generally try to form mental representations of others (Kihlstrom & Hastie, 1997) and, in doing so, two processes are used: (1) behavioral identification, where behavior is evaluated in terms of relevant categories; and (2) dispositional inference, where behavior information is integrated with situational information (Trope, 1986). In the initial identification stage, behaviors act as cues for inferring underlying or latent characteristics of the target. Judges encode these behaviors in terms of trait concepts when reading others' actions (Wyer & Srull, 2014). Ultimately, interviewers who make correct behavior-trait inferences would therefore form a more accurate overall impression of the interviewee. Considering these arguments, we posit:

Hypothesis 2a: Trait induction will moderately ($r > .30$) relate to interviewer accuracy.

Trait extrapolation and interviewer accuracy. Trait extrapolation can be defined as the understanding of how traits and their behavioral manifestations naturally co-vary. For example, someone who can extrapolate from one trait to another would know that people who are honest are generally also reliable (Goldberg, 1992). When an interviewer infers that an applicant exhibits one trait (e.g. honest), she extrapolates the existence of the other (e.g. reliable) by drawing on a personal understanding of how traits tend to co-vary. So, by understanding the general covariation between traits, he or she is able to fill-in missing information to form a more coherent person impression of the applicant.

The theoretical frameworks that inform trait extrapolation have a long history in personality literature (e.g., see Asch, 1946; Jackson, 1972). Generally, the notion of individual differences in understanding true (objectively determined) trait covariation is a fundamental premise of implicit personality theory (IPT, Jackson, Chan, & Stricker, 1979; Schneider, 1973). This theory posits that people use naïve, common sense IPTs to form impressions (Bruner & Tagiuri, 1954). As such, an IPT is a set of perceived or expected relations among personality traits, which may or may not be true or accurate (Van der Kloot & Kroonenberg, 1982; Wiggins & Blackburn, 1976).

Trait extrapolation has been measured by various methods and instruments (for an early review, see Schneider, 1973). Christiansen et al. (2005) presented subjects with items describing hypothetical persons and asked them to select, from additional trait or behavior descriptors, the most appropriate option. The correct answer was derived from empirical studies of the covariation of traits and behaviors.

In interviews, trait extrapolation is likely to affect judgment accuracy. Interviewers are often constrained by limited behavioral information (e.g. 30-minute interactions with interviewees) from which to extract trait and person impressions. When forming impressions of others from limited information, judges most likely rely on heuristic mechanisms such as trait extrapolation to fill-in missing aspects of the person impression. Accurate interviewers are likely to have more accurate IPTs – they can correctly extrapolate between traits. Like completing a puzzle, they correctly infer missing pieces of the target's profile, resulting in a more coherent, accurate person impression.

Empirical evidence supports this view. Not only are IPTs active in judgments of personality (Ebbesen & Allen, 1979; Wiggins & Blackburn, 1976) and performance (Krzystofiak, Cardy, & Newman, 1988), they also seem to affect judgment accuracy

(e.g., Hauenstein & Alexander, 1991; Kishor, 1995). Considering these arguments, we posit:

> *Hypothesis 2b*: Trait extrapolation will be moderately (r > .30) related to interviewer accuracy.

Trait contextualization and interviewer accuracy. Besides trait induction and trait extrapolation, dispositional reasoning also involves judgments about situations. Research shows that it may be more likely for a specific trait to be expressed (or elicited) in certain situational contexts (Tett & Guterman, 2000). As such, people consider relevant situational information when trying to understand others' behavior and the dispositions they imply (Funder & Ozer, 1983; Shoda, Mischel, & Wright, 1989). Trait contextualization refers to the ability to identify situations that are relevant to different traits. Judges with high levels of trait contextualization have insight into which situations are likely to see a trait expressed. For example, extroversion is more likely to manifest itself in a situation where a target is surrounded by other people, as opposed to one where she is alone.

Trait contextualization has its theoretical origins in the interactionist perspectives on personality (Kihlstrom, 2013; Mischel & Shoda, 1995) where behavior is considered a function of the interaction between the person and the environment. Trait activation theory (Tett & Guterman, 2000) assumes that traits are expressed (or 'activated') only in certain situations. Situations, therefore, either inhibit or entice trait expression – a notion receiving increasing research support (e.g., Robinson, 2009).

Few measures of trait contextualization exist. Tett and Guterman (2000) developed ten trait relevant scenarios written to be relevant to each of five traits measured by the revised Jackson Personality Inventory (Jackson, 1994), namely risk-taking, complexity, empathy, sociability, and organization. Two scenarios were assembled to provide opportunities to express a targeted trait in each of five life domains of college students, namely school, shopping, home, travel, and work. The researchers determined actual trait relevance for each situation-trait pair by using the mean trait-relevance ratings of 26 judges (identified from a pool of 123 judges, using scores on an independent adjective sorting task).

There are compelling reasons to expect good judges to have high trait contextualization ability. Gilbert's (1989) model of social inference holds that, after observed behavior is categorized and characterized, an important correction stage follows, where the initial dispositional inference is adjusted, given the situational constraints on behavior expression. If an interviewer fails to incorporate situational information correctly into a final dispositional inference, it may lead to an inaccurate judgment. For example, an interviewer who concludes that an applicant is a highly anxious person, without accounting for the stressful context of a panel

interview, would make inaccurate inferences about the candidate's neuroticism. Thus,

> *Hypothesis 2c*: Trait contextualization will moderately (r > .30) relate to interviewer accuracy.

Discriminant Validity with Personality

In order to show discriminant validity (Campbell & Fiske, 1959) the components of dispositional reasoning should be empirically unrelated to constructs with which they share little or no conceptual relationship. A key outstanding issue is how judges' dispositional reasoning corresponds to their personality traits.

In light of our hypothesis that dispositional reasoning components are a set of mental abilities, they are positioned in a different conceptual domain than the personality domain. Personality refers to predispositions to respond to stimuli in a certain way (John, Robins, et al., 2008). As such, personality involves a strong behavioral tendency focus (Mayer, et al., 1999). Conversely, dispositional reasoning has a strong cognitive focus, squarely rooted in information processing *about* traits, behavior and situations. Hence, dispositional reasoning should also be relatively independent from personality trait measures

To our knowledge, only one study has investigated the relationship between judges' dispositional reasoning and personality: Christiansen et al. (2005) found that only openness to experience showed moderate effects with dispositional reasoning (r = .34, p < .05). Other traits showed trivial or no effects, implying that dispositional reasoning and personality constructs were relatively distinct from one another. Therefore, we expect:

> *Hypothesis 3*: The three components of dispositional reasoning will be unrelated (i.e. there will be trivial to no effects) to Big Five personality measures.

Incremental Validity of Dispositional Reasoning

A final criterion for an intelligence measure is that it must show empirical distinctness from general cognitive ability and provide incremental validity (in predicting relevant outcomes). Failing this, it would imply that such a measure is most likely to form part of general intelligence (Carroll, 2003). By illustration, if the components of dispositional reasoning show very high (e.g. > .80) correlations with measures of general mental ability and, at the same time, fail to explain unique variance in predicting accuracy, one may conclude they essentially are 'just *g*' measures (Mayer et al., 1999).

We expect dispositional reasoning to increment general intelligence in predicting accuracy, for two reasons. First, dispositional reasoning and general

mental ability may be empirically distinct (e.g. r = .43; Christiansen et al., 2005) from one another. So, dispositional reasoning represents a related, but different set of abilities than general mental ability. This partial overlap between the constructs may increase the likelihood of finding incremental validity evidence in predicting accuracy. Second, intelligence measures explain only a part of the variance in judgment accuracy measures. In prior studies, intelligence measures showed only small-to-medium effects with interviewer accuracy (r = .25, p < .01; Christiansen et al., 2005). Thus,

> *Hypothesis 4*: The set of dispositional reasoning components will explain additional variance in judgment accuracy not already explained by raters' general cognitive ability (g). We expect this effect to be more than small (>.10).

4.3 Method

Participants

We recruited participants (police managers) undergoing a seven-week managerial training course required for promotion purposes. To increase the external validity of our study results, we ensured that all study participants were employed and had prior (at least five years) comparable work experience as managers. There were 146 managers in the sample (24.6% females and 75.4% males). In terms of race, the sample comprised of 71.2% Black African, 17.3% White, 9.4% Mixed Race and 2.2% Asian participants. The mean age of managers was 43.7 (*SD* = 5.36) years. The majority of the officers' rank was Captain (57.6%), while the rest were Warrant Officers (35.3%) and Lieutenants (7.2%). These officers represent the most likely interviewers in typical organizations, as they fall between lower-level first-line supervisors and senior management. Some (37.9%) had post-secondary school qualifications (vs 62.1% with only a senior secondary school certificate). The prevalent first languages amongst these officers were Afrikaans (26.4%), IsiZulu (13.8%), Sesotho (12.6%) and English (11.5%), although English was the official workplace language of the participating organization in South Africa. As such, no one reported difficulty understanding spoken or written English.

The data collection was completed in a single session at the end of 2011. After introducing the research as part of interviewer training, we explained the rating procedure and materials. Next, we showed five video-recorded interview segments to the group of participants, using a large video projector screen and audio equipment. The first video candidate was employed as a practice run, followed by a discussion of the ratings and final clarification of questions that remained about the rating procedure. Following each of the remaining four video segments, managers independently completed the interview dimension rating sheets. Finally, they filled in the individual difference measures before being debriefed and thanked for their participation.

Materials

Development of interview videos and materials

We decided to use videotaped segments of interviewee performance as stimuli as it allowed for the presentation of similar stimuli to all participants. We video-recorded semi-structured interviews of five graduate students recruited to take part in "an interview that would help them prepare for the job application process". The interview format was a competency-based, situational interview (Latham, Saari, Pursell, & Campion, 1980). The interview questions were designed to tap two specific dimensions, *communication* and *people management*. These dimensions were selected given their widespread use in interviews (Huffcutt, et al., 2001) and they were considered applicable (i.e. derived from job-analysis) to the fictional position ('junior management position') for which they were applying. We used eight

questions that took the form of "What would you do if..." questions, typical in situational interviews. For example, to measure communication, we used items such as "How would you handle a situation where your work colleagues ignore your ideas and input?" A brief rating guide was provided for each question to anchor possible responses to scale points. For example, for the item above, a score of '1' would be assigned to responses such as "I stop giving ideas and input". Interviewee's responses to each question were rated on a 7-point Likert scale (1 = *poor response*, 7 = *excellent response*). In the actual interviews that were recorded, an expert interviewer asked applicants to respond to the same questions, presented in the same order to all interviewees. The final video segments were shortened to a viewing time of 5 min each.

Video Interviewee True Scores

A 'true score' represents the mean of an infinite number of scores across parallel measures of a test. In line with earlier recommendations for true score estimation (for a review, see Sulsky & Balzer, 1988) our accuracy criterion combined ratings from multiple subject matter experts (SMEs). We asked a panel of seven expert judges (SMEs) – comprising of qualified industrial psychologists (with at least a Master's degree in IO Psychology) and professors in IO Psychology – to rate the video-taped applicants on the two interview dimensions. To minimize possible demographic effects, we balanced the targets and expert raters in terms of gender and ethnicity. Using the Borman (1977) procedure, we gave all expert raters the opportunity to view the video-recorded applicants as many times as they wanted before completing the structured interview rating sheet. Mean interjudge agreement between SMEs for judging both dimensions across targets was strong (LeBreton & Senter, 2008) overall, $ICC_{tot}(2, k)$ = .86, and also for separate dimensions, *communication* = .91, *people management* = .81. To obtain overall true score estimates for each interviewee, we averaged the ratings made by the respective SMEs.

Criterion Measure

Accuracy scores served as dependent variable. Consistent with the two earlier dispositional reasoning studies (Christiansen et al., 2005; Powell & Goffin, 2009) we computed an accuracy score for each participant by calculating within-person profile correlations (i.e. between the profile inferred by the rater and the accuracy criterion profile of the target) (see Borman, 1977) at the dimension level, with an *r*-to-Fisher's-*z* transformation. This method assesses the congruence (see Funder & Colvin, 1997) between the complete set of judgments made by a judge and the target.

Using the procedures of Sulsky and Balzer (1988) we also calculated all four Cronbach accuracy measures (Cronbach, 1955): (a) *elevation*, the overall tendency to rate dimensions too high or low; (b) *differential elevation*, the accuracy with which a rater can differentiate among targets, when averaging all traits; (c) *stereotype*

accuracy, the accuracy of relative distinctions produced among average trait levels, when averaged across targets; and (d) *differential accuracy*, the interviewer's sensitivity to target differences in patterns of traits.

Predictor measures

General cognitive ability. All participants completed the Wonderlic Personnel Test – Revised (WPT-R) at the beginning of the testing session. The Wonderlic Personnel Test is widely used to measure general cognitive ability (Wonderlic Personnel Test, 2002). It is a 50-item, timed test (12 minutes), with items that include word comparisons, disarranged sentences, number comparisons, analysis of geometric figures and problems requiring mathematical and logical solutions. It assesses mathematical, verbal, logical reasoning, and spatial ability, from which a measure of general mental ability is created. The Wonderlic has good predictive validities for a wide range of criteria (Wonderlic, 1998) and reliability estimates[4] generally vary between .82 and .95 (Wonderlic Personnel Test, 2002).

Dispositional reasoning. To measure the dispositional reasoning components, we revised the Interpersonal Judgment Inventory (IJI) of Christiansen et al. (2005). The original version consisted of 45 multiple-choice items that assessed a person's knowledge about personality and how it is related to behavior and situations. In their sample, the overall measure showed an internal consistency reliability estimate of .82.

We revised the IJI for three reasons. First, earlier studies (e.g., Christiansen, et al., 2005; Powell & Goffin, 2009) used only an overall score, for example, by combining scores from all items. However, a component-level measure with longer and reliable subtests was necessary to test our hypotheses. The second reason we revised the IJI stemmed from the fact that the original items contained content written for a student sample. As a result, some items were unsuitable for a sample of working adults. For example, in one item, participants had to choose a trait which was most relevant to situations like "After a morning exam, you overhear some classmates you've met only briefly talking about going to lunch at a nearby restaurant". To address this limitation, we deleted some items, rewrote others, and constructed new items with a work context in the item description.

The last reason why revision was needed was because we anticipated that non-university respondents (i.e. with lower educational levels) may have difficulty to comprehend some item stems, response options, and questionnaire instructions. Potentially problematic items (e.g. "exhibit condescending behavior") were identified in a pilot study and replaced with terms that were easier to understand (e.g. "be arrogant and 'high-and-mighty' in their behavior").

We used the same procedure described in Christiansen et al. (2005) to draft a final set of 86 pilot items. That is, an expert panel (consisting of six university

professors, with backgrounds ranging from industrial, organizational, social, and personality psychology, as well as linguistics) wrote additional items for each subscale. These items were revised for clarity and piloted in a sample of assessors (N = 19) undergoing rater training for a large scale assessment project. Some items were deleted or reworded based on item analysis. Here, we deleted items not adhering to cut-offs[5] for item variability (e.g. SD < .10), item difficulty (p < .15 or p >.85), and discrimination indices (d < .10). We also tagged poor items on the basis of distractor analysis (no or few distractor endorsements) and pilot sample feedback. The final set of three measures consisted of 66 items.

Trait induction. The first subset of 20 items measured behavior-trait inferences. After describing the Big Five personality traits, the measure presented a list of adjectives from Goldberg's (1992) factor markers. The task was to identify the traits (e.g. conscientiousness) that best matched the marker adjectives (e.g. thorough) that were provided. An example item can be found in the Appendix.

Trait extrapolation. The trait extrapolation measure (23 items) assessed a respondent's understanding of how traits co-occur. Items described a fictional person and required respondents to select which of four descriptions was most (or least) likely also true of the person. An example item can be found in the Appendix.

Trait contextualization. The last set of 23 items measured understanding of trait–situation relevance. This measure was originally based on empirical results from Tett and Guterman (2000). Originally, Christiansen et al. (2005) keyed per item one response option as being the most consistent with empirical evidence, theoretical relationships, and expert judgment. One subset of items presented a trait description, e.g. 'risk-taking' by listing examples of behaviors associated to high and low scorers on the measure. Next, respondents had to choose which of five situations listed would most likely elicit the relevant behavior. An example item can be found in the Appendix. The second subset of items reversed the direction of inference, i.e. they described a situation and respondents had to identify the trait most likely to be observed in trait-relevant behavior.

In our sample, we computed the CFA-derived construct reliabilities (Brown, 2015; Raykov, 2012) of the final measures and these were acceptable (induction = .77; extrapolation = .81; contextualization = .76).

Personality

A 20-item short form of the International Personality Item Pool (IPIP) Big-Five factor markers (Goldberg, 1992) – developed and validated across five studies by Donnellan, Baird, Lucas, and Oswald (2006) and later factor analytic studies (e.g., Cooper, Smillie, & Corr, 2010) – was used to measure interviewers' personality. Participants rated how well each of the items described themselves on a 5-point scale (*very inaccurate* to *very accurate*) (Goldberg, 2005). In our sample, the mean

inter-item correlations for each scale were comparable to those in earlier published studies (Donnellan, et al., 2006).

4.4 Results

Descriptive Statistics

Table 4.1 presents the means, standard deviations, and correlations of the study variables. The mean score of the revised dispositional reasoning measure, scored as a percentage score, was M = 46.37 (SD = 13.38). Our results indicate that participants experienced moderate difficulty to reason about traits, behaviors, and situations. Also, the standard deviation of the dispositional reasoning measure scores supports the notion of individual differences. Descriptive statistics for other measures were relatively comparable to those in earlier studies, although the general mental ability scores of managers in the present sample (M = 12.46; SD = 4.4) were lower than those published in earlier studies using university student samples.

Table 4.1

Descriptive Statistics and Intercorrelations of Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Gender[a]	-	-	–															
2. General mental ability	12.46	4.40	.34**	–														
3. Dispositional reasoning	46.37	13.38	.30**	.68**	–													
4. Trait induction	36.21	22.38	.32**	.52**	.79**	–												
5. Trait extrapolation	51.22	19.24	.18*	.47**	.72**	.48**	–											
6. Trait contextualization	56.57	23.05	.19*	.61**	.79**	.51**	.41**	–										
7. Extraversion	11.96	4.58	-.20*	-.26**	-.31**	-.25**	-.18	-.27**	–									
8. Agreeableness	13.03	4.38	.16	.28**	.25**	.16	.23*	.18	-.11	–								
9. Conscientiousness	14.13	4.03	.04	.16	.20*	.11	.16	.10	-.19*	.40**	–							
10. Neuroticism	12.77	4.20	-.04	-.07	-.09	-.15	-.14	-.11	-.21*	.12	.31**	–						
11. Intellect/Imagination	11.93	3.83	-.01	.11	.07	.05	.06	.05	.03	.40**	.54**	.20*	–					
12. Interviewer accuracy[b]	.99	.65	.16	.20*	.34**	.14	.33**	.26**	-.12	.15	.13	.07	.04	–				
13. Elevation accuracy[c]	.64	.46	-.10	-.16	-.22*	-.21*	-.26**	-.20*	.17	-.09	-.16	-.12	-.03	-.28**	–			
14. Differential elevation[c]	.66	.29	-.13	-.12	-.08	.02	.00	-.11	.00	.02	-.01	.01	.12	-.23**	-.10	–		
15. Stereotype accuracy[c]	.24	.21	-.12	-.03	-.14	-.08	-.15	-.09	.05	-.02	.03	.09	.10	-.21*	.07	-.05	–	
16. Differential accuracy[c]	.46	.22	-.04	-.12	-.18*	-.11	-.04	-.19*	.02	-.05	-.06	.02	.08	-.11	-.03	-.03	.10	–

Note. $N = 142$. Trait induction = judges' ability to infer traits from behavior cues; Trait extrapolation = understanding of how traits naturally co-vary; Trait contextualization = to know how situations affect trait expression. The dispositional reasoning scores are all expressed as percentages.

[a]Gender was coded such that men were 1 and women were 2. [b]Interviewer accuracy scores are Fisher transformed (r to z) profile correlations between participants' ratings at dimension level and subject matter experts' true score estimates. [c]Cronbach (1955) accuracy component scores were calculated using the procedures of Sulsky and Balzer (1988). Lower Cronbach scores imply higher accuracy.

* $p < .05$; ** $p < .01$ (two-tailed).

Tests of Hypotheses

Hypothesis 1 proposed that the dispositional reasoning components would show positive manifold internally, i.e. they would moderately ($r > .30$) relate to each other, as well as externally, that is, they would moderately ($r > .30$) relate to a measure of general mental ability. The correlations (see Table 4.1) among the components internally revealed medium to large ($.41 < r < .51$) (all $p < .01$) effects (Cohen, 1988). Externally, they also demonstrated large effects on general mental ability (.61, trait contextualization; .52, trait induction; .47, trait extrapolation; all $p < .01$). Confidence intervals generally contained the hypothesized effect ranges, min $r = .41$, 95% CI: [.27, .54], max $r = .61$ [.50, .70]. Therefore, Hypothesis 1 was supported. The effect sizes were even larger than anticipated.

Although the components are conceptually related (positive manifold) they should also be empirically distinct from one another. Therefore, as suggested by an anonymous reviewer, we compared (a) a baseline model (model M1) in which the correlations between dispositional reasoning are freely estimated; and (b) a nested comparison model (model M0) in which the correlations are constrained to be unity. Cognitive ability was included in both models. We used the correlations as input for the analysis and found poor fit of the nested model, $\chi^2(3, N = 142) = 86.078$, $p < .01$, RMSEA = .44 (90% CI: .36; .52). A chi-square difference test indicated that the nested model (M0, specifying the relationship between dispositional reasoning facets as perfectly correlated) showed significantly poorer fit, compared to the baseline (M1) model, $\chi^2_{diff}(3, N = 142) = 86.078$, $p < .001$. Therefore, the evidence suggests that the components are empirically distinct from one another.

Whereas Hypothesis 1 focused on the positive manifold ('related-set') criterion for a classic intelligence measure, Hypothesis 2 proposed that trait induction (Hypothesis 2a), trait extrapolation (Hypothesis 2b), and trait contextualization (Hypothesis 2c) would predict interview judgment accuracy, all with moderate ($r > .30$) effects. As shown in Table 4.1, trait extrapolation (.33, $p < .001$, 95% CI: [.18, .47]) and trait contextualization (.26, $p = .002$, [.10, .41]) showed moderate effects on accuracy, but trait induction (.14, $p = .11$, [-.02, .30]) had only a trivial effect. Therefore, Hypothesis 2 was partially supported as one effect size was weaker than hypothesized.

Hypothesis 3 stated that the three components of dispositional reasoning would show discriminant validity with Big Five personality measures. Table 4.1 shows that only three (out of fifteen) bivariate relationships between these facets showed small to medium effects, while the others showed trivial to no effects. Judges who were more extraverted had lower induction ($r = -.25$, $p = .007$, 95% CI: [-.40, -.09]) and contextualization scores ($r = -.27$, $p = .005$, [-.41, -.11]), while agreeable judges showed better extrapolation ($r = .23$, $p = .013$, [.07, .38]). In sum, there is evidence regarding discriminant validity of the components with personality measures. Therefore, our results generally supported Hypothesis 3.

Hypothesis 4 posited that individual differences in dispositional reasoning components would increment the validity of general mental ability to predict accuracy. Table 4.2 summarizes the results of the hierarchical regression analyses. In Step 1, the general mental ability score was entered. In Step 2, we entered trait induction, trait extrapolation, and trait contextualization as a set. Consistent with our hypothesis, results revealed a significant increment in the ability to explain accuracy (ΔR^2 = .09, p = .004) when trait induction, trait extrapolation, and trait contextualization were added in Step 2. So, the addition of dispositional reasoning components to the equation with general mental ability resulted in a statistically significant increment in R^2. In addition, the small to medium effect size[3] (Cohen's f^2=.11) that we observed, supports Hypothesis 4.

In addition to the analyses that tested our hypotheses, we also conducted relative weights analysis (Tonidandel & LeBreton, 2011; Woehr & Arthur, 2003) to examine which of the three components of dispositional reasoning was most important in determining interview accuracy. As shown in Table 4.2, the relative weights analysis showed that extrapolation (58.54%) and contextualization (26.76%) exerted the strongest influence in predicting accuracy, followed by induction (4.74%).

Table 4.2

Results of Hierarchical Regression Analyses of Interviewer Rating Accuracy[a] (Borman's Differential Accuracy) on General Mental Ability and Dispositional Reasoning Components

Predictor	Step 1	Step 2		
	β	β	RW_{raw} [95%CI]	RW%[b]
Step 1				
General mental ability	.22*	.00	.01 [.003, .046]	9.96%
Step 2				
Induction		-.09	.01 [.000, .011]	4.74%
Extrapolation		.31**	.08* [.025, .164]	58.54%
Contextualization		.18	.04 [.004, .113]	26.76%
Total R^2	.05*	.14**		100.00%
ΔR^2	.05*	.09**		

Note. N = 142. Induction = judges' ability to infer traits from behavior cues; Extrapolation = understanding of how traits naturally co-vary; Contextualization = to know how situations affect trait expression; RW = relative weight.

[a]Accuracy scores are Fisher transformed (r to z) profile correlations between participants' ratings at dimension level and subject matter expert true score estimates.

[b]Relative weights are not raw weights, but rescaled to express the % contribution of each predictor to overall R^2. Confidence intervals around the raw weights were calculated using the bias corrected accelerated method for generating the bootstrapped confidence intervals.
*p < .05. **p < .01.

Additional Analyses

In addition to our test of the hypotheses using the correlational accuracy measure (Borman, 1977), an anonymous reviewer suggested that we also report the results using the Cronbach (1955) accuracy measures as dependent variables. Higher Cronbach scores denote lower accuracy and, hence, negative correlations with our correlational accuracy measure would reflect agreement among the indices. There was only a trivial positive relationship between our correlational measure of accuracy and Cronbach's Differential Accuracy, r = - .11, p = .185. The Borman correlational accuracy index showed small to medium positive effects with the remaining Cronbach components, including elevation accuracy (-.28), differential elevation (-.23), and stereotype accuracy (-.21). While these effects are not negligible, weak correlations among operational definitions of accuracy are not uncommon (Becker & Cardy, 1986).

Differential accuracy is the most closely related counterpart to the Borman measure (Sulsky & Balzer, 1988). In our study, trait extrapolation (-.04) was a weak predictor of differential accuracy (whereas it was one of the best predictors of the correlational index, .33). In combination, the predictors (general mental ability and

the dispositional reasoning components) explained less of the variance in the differential accuracy criterion (R^2 = .04; see Table 4.3) than in the correlational accuracy measure (R^2 = .14; see Table 4.2). Also, the dispositional reasoning components did not show incremental validity (ΔR^2 = .02) over general mental ability when the differential accuracy measure was used.

Table 4.3

Results of Hierarchical Regression Analyses of Interviewer Rating Accuracy[a] (Cronbach Differential Accuracy) on General Mental Ability and Dispositional Reasoning Components

Predictor	Step 1	Step 2		
	β	β	RW_{raw} [95%CI]	RW%[b]
Step 1				
General mental ability	-.12	-.01	.01 [.000, .026]	16.68%
Step 2				
Induction		-.06	.01 [.000, .024]	14.02%
Extrapolation		.06	.00 [.000, .003]	3.88%
Contextualization		-.18	.02*[.001, .095]	65.43%
Total R^2	.01	.04		100.00%
ΔR^2	.01	.02		

Note. N = 142. Induction = judges' ability to infer traits from behavior cues; Extrapolation = understanding of how traits naturally co-vary; Contextualization = to know how situations affect trait expression; RW = relative weight.

[a]Accuracy scores are Cronbach Differential Accuracy scores using participants' ratings at dimension level and subject matter expert true score estimates.

[b]Relative weights are not raw weights, but rescaled to express the % contribution of each predictor to overall R^2. Confidence intervals around the raw weights were calculated using the bias corrected accelerated method for generating the bootstrapped confidence intervals.
*p < .05. **p < .01.

4.5 Discussion

Main conclusions

In personnel selection, there has been a longstanding interest in what makes a good judge. Even though empirical research suggests that judges differ in rating accuracy, the reasons why are not yet clear. The present study posited that rating accuracy is partly dependent on specific facets of judges' dispositional reasoning. By taking an in-depth look at the dispositional reasoning construct, we were able to determine whether or not its components - induction, extrapolation, and contextualization - would act as classic intelligence measures in predicting interview judgment accuracy. We tested the components against strict conceptual and correlational criteria for a classic intelligence (Mayer, et al., 1999). Although the important role of multiple components in accuracy has been suggested by earlier judgment theories

(e.g., Gilbert, 1989), to the best of our knowledge, this study is the first to test the componential view of dispositional reasoning. In doing so, we added to the current understanding of what makes the 'good judge' (Funder, 2012).

The first main conclusion is that the dispositional reasoning components generally correspond to intelligence measures. That is, they broadly adhered to the conceptual criteria and the correlational criteria we tested. For example, the dispositional reasoning components converged with one another and with general mental ability (i.e. positive manifold). As in previous studies (e.g., Christiansen, et al., 2005) broad dispositional reasoning also predicted accuracy, but we extend this research to show that managers who were better at extrapolation and contextualization, specifically, were more accurate. In addition, the components of dispositional reasoning generally showed discriminant validity with personality constructs. Our finding that the dispositional reasoning components also showed incremental validity in explaining accuracy, beyond cognitive ability, is novel. In short, these results provide evidence for a nomological network with dispositional reasoning positioned as an intelligence, namely as a specific mental ability that good judges employ to process behavioral, social, and situational information.

Our second main conclusion relates to the predictive validity of the components of dispositional reasoning. Earlier, it was found that broad dispositional reasoning can explain variance in judgment accuracy (e.g., Christiansen et al., 2005). Actually, our in-depth study of the construct suggests that dispositional reasoning is better understood as a cluster of abilities that each may be linked to accuracy. Apparently, two of the facets are required to achieve accurate judgments, namely extrapolation and contextualization. It seems that induction is less important in interview judgment accuracy, perhaps because the high-structure interviews we used actually facilitated the linking of dimension cues to specific interview dimensions. Taken together, we show that not only is general mental ability and broad dispositional reasoning related to accurate judgment of interviewees (Christiansen, et al., 2005), but the ability to deal with trait- and situation information specifically is also important.

Our results point out that judgment accuracy may depend more upon understanding how traits co-vary (extrapolation), and how situations affect trait expression (contextualization), than knowing which traits are signalled by behavioral cues (induction). The relative weights analysis confirmed that, amongst the components, extrapolation and contextualization exerted the strongest influence in predicting accuracy, in that order. To put it another way, when faced with the task of interviewing another person, interviewers apparently differ in their ability to understand trait constellations and situation-relevant information. In short, those individual differences in people's dispositional reasoning matter as they help to explain why some interviewers produce better ratings than others.

By including multiple operationalisations of accuracy (e.g. Borman's correlational accuracy measure, as well as the Cronbach components), we could also determine the stability of the effects we observed across different across measures. Overall, we found the 'same story' emerging from the results, albeit 'with different tales'. Predictor effect patterns were relatively similar, but smaller, when using the Cronbach elevation accuracy and differential accuracy measures, than when relying on the Borman index. Our results also showed that general mental ability and dispositional reasoning do not seem to adequately explain interviewers' ability to differentiate among targets (averaged across traits, i.e. differential elevation), nor interviewers' ability to make relative distinctions among average trait levels, when averaged across targets (i.e. stereotype accuracy). We encourage further research along these lines.

Limitations

First, some generalizability issues should be noted. We sampled a group of participants who worked for the same employer. Care should be taken about overgeneralizing the findings from a single organization. Future research could also investigate the extent to which our results generalize to other industries. Our study also relied on a specific type of judge, that is, managerial interviewers. We decided on this approach to extend the generalizability of dispositional reasoning studies to non-student samples. Yet, it would be interesting to know whether the results would generalize to other types of judges, for example, psychologists. Despite this limitation, it may be argued that using actual managers as interviewers, who routinely conduct interviews, bolstered the realism of the present research.

Second, our research design did not use actual life interviewees, but rather videotaped interviewees for reasons of experimental control. By using standardized video-taped stimuli in a controlled testing venue, we were able to tease apart the role of dispositional reasoning in judgment accuracy – an objective that remains difficult in field studies (where high control of extraneous factors is not possible). By using the same stimulus materials, we could hold performance constant and minimize error variance related to using real interviewees. Using standard stimuli also allowed us to determine true scores for these performances, which would have been impossible in a true field setting. We tried to mitigate the loss of fidelity by creating realistic conditions for the interviewees. All the interviewees were currently looking for jobs and used the interview exercise to prepare for real selection situations. Additionally, most participants stated that they perceived the interviews as fairly realistic. However, future studies should explore the degree to which judges' dispositional reasoning components are able to explain individual differences in the accuracy of judges interviewing real interviewees.

Implications for Theory and Future Research

Our study has implications for issues that may be important for theory building and research on individual differences in accuracy. At the broadest level, findings lend

support to mainstream theories of judgment accuracy that suggest that judges can be important moderators of accuracy (Funder, 1995, 2012). More specifically, accuracy stems in part from individual differences in their ability to utilize behavioral cues (as predicted by Funder, 1999). While Christiansen and colleagues (2005) illustrated that broad dispositional reasoning explains differences in accuracy – a finding we replicate here – we were able to offer a more fine-grained representation of how the specific facets of dispositional reasoning may play a role in producing high-fidelity impressions of the interviewee. According to our relative weights analysis, especially trait extrapolation and contextualization of behavioral information are important in producing accuracy.

In addition to showing support for existing judgment theories, the evidence for an intelligence-view of dispositional reasoning opens up interesting avenues for more theory building. For one, what is the nomological network surrounding dispositional reasoning? We see conceptual overlaps in the construct domains of dispositional reasoning and others that people use to understand people that surround them. Consider the conceptually related constructs of emotional intelligence and social intelligence. Although these individual differences may vary in focus (understanding emotions and social characteristics, respectively) from dispositional reasoning (understanding behavior and traits), they may share a common process of inductive reasoning, where judges infer underlying general characteristics from observable behavior. In light of the need to knit together a tapestry of individual differences in interpersonal judgment (for a discussion, see Lievens & Chan, 2010), we call for more studies that test the relationships between dispositional reasoning and social and emotional intelligence. These studies could also try to disentangle the constructs' relative importance in determining accuracy, especially in different judgment tasks.

Second, our finding that dispositional reasoning may be an intelligence raises questions about the malleability of the dispositional reasoning components. If it is an intelligence how does it develop, if at all? Classic intelligences are known to develop with age (Mayer et al., 1999), but we could not test this age-criterion in our investigation. Studies that show how dispositional reasoning develops with age would provide further insight into its intelligence-basis. Related to this idea, one might wonder whether it is possible to improve judges' dispositional reasoning with short-term training interventions. The answer would most likely hinge on the underlying nature of the construct. Is it closer to an innate ability (e.g. fluid intelligence) or more likely to respond to environmental exposure (e.g. crystallized intelligence)? Intelligences are likely to develop with age (Mayer, et al., 1999), especially if they are of the crystallized nature (Horn & Cattell, 1966) rather than fluid. This change probably occurs through experience and exposure to environmental stimuli (Bickley, Keith, & Wolfle, 1995). However, a recent study (Powell & Goffin, 2009) failed to observe changes in behavior-trait knowledge (i.e.

the induction component) when brief personality-knowledge based instruction was given to undergraduate students. Perhaps dispositional reasoning consists of fluid and crystallized components, similar to social intelligence (Lee, Wong, Day, Maxwell, & Thorpe, 2000). To be more effective in developing dispositional reasoning, we should explore other strategies that consider how accuracy develops (see Fiske & Macrae, 2012). In short, future work could help position the dispositional reasoning construct in a domain of individual differences by exploring how it develops.

Third, we have questions about the role of the rating context in our results. For example, would the components relate differentially to accuracy criteria, given differences in the judgment context? In our study, we used a high structure rating condition, where situations were held constant. When standardized rating materials, instructions, and criteria are provided (typical in high-structure interviews) judges are encouraged to use normative theories to interpret behavior, rather than personal, idiosyncratic theories (Uggerslev & Sulsky, 2008). As such, our judges were not required to draw on their implicit personality theories to form impressions of interviewees. When rating materials specify which interview dimensions are implied by certain behaviors, the role of the induction component could be diminished. Also, standardisation of the context limits situational variability, which is also likely to lower the need for a judge to draw on their contextualization ability. In sum, aspects of the rating context may be important boundary conditions for the relative importance of dispositional reasoning components in judgment accuracy. We call for more research that explores how situations may moderate which components of dispositional reasoning influence judgment accuracy.

Implications for Practice

Given that the effects we observed for the predictive and incremental validity of the dispositional reasoning components were small to moderate, we expect these to translate to practically significant outcomes for the workplace application. In fact, a number of practical implications for interviewer screening and training follow from our results. In terms of interviewer screening, companies might consider using candidates' scores on various dimensions of dispositional reasoning as part of a selection battery for potential interviewers or assessors. Our results showed that dispositional reasoning components add substantive variance in the prediction of judgment accuracy. So, dispositional reasoning scores are job-related and can be used in making selection decisions about interviewers. In addition, with a relatively short administration time, companies seem then better able to screen interviewers for likely judgment accuracy than from using no tests at all, or relying on measures of *g*.

However, prior to implementing our suggestion to use this assessment instrument as a simple and efficient to use tool, further research is needed. For

instance, in our study, we asked managers to rate interviewees. We do not know whether the dispositional reasoning-accuracy link exists with non-managerial judges, such as psychologists, trained assessment centre assessors, and others. We also do not know how these judges differ from one another on dispositional reasoning. Literature suggests that accuracy differences exist between different types of judges (Sagie & Magnezy, 1997), but we do not have a clear picture as to why they differ. Hence, a fruitful avenue for future research would be to examine individual differences in dispositional reasoning components and accuracy between different types of judges.

Another potentially useful training approach lies in interventions that focus on each component. Currently, most rater training is based on frame-of-reference (FOR) training – the empirical meta-analytic evidence also supports the effectiveness of FOR (Roch, et al., 2012; Woehr & Huffcutt, 1994). However, FOR does not typically entail the contextualization and extrapolation components, but generally only the induction component.

4.6 Conclusion

This study endorsed a componential view on dispositional reasoning and examined its role in interviewer judgment accuracy. No earlier studies have evaluated the role of dispositional reasoning at this level of granularity. Evidence suggests that, given our tests against strict criteria, dispositional reasoning exhibits some of the characteristics of an intelligence measure. We conclude from our findings that dispositional reasoning components are unlikely to be 'just *g*' measures.

Rather, they may represent specific intelligences in the social-cognitive domain that allow better use of behavior information in interviews. Moreover, it may have distinguishable subcomponents that could constitute pieces of the puzzle in understanding what makes the good judge in personnel selection interviews. In our study we found that, compared to poor judges of interview dimensions, accurate judges had better developed abilities to extrapolate and contextualize trait-related information. While the jury is still out on the role of dispositional reasoning components in other judgment contexts, for example, assessment centres and performance rating contexts, we hope this research will trigger further research on related issues.

Appendix: Example Items from the Revised Interpersonal Judgment Inventory (RIJI)

Trait induction

Circle the letter that corresponds most to the trait you think is represented by the word:

	Trait				
Behavior	Emotional stability	Extraversion	Openness	Agreeableness	Conscientious-ness
Sloppy					X
Irritable	X				

Trait extrapolation

For example, one item depicted 'John' as "John's co-workers all describe him as efficient, thorough, and persistent. MOST likely John also:". Next, respondents had to choose the best answer from the following four options:

A. feels the need to be around lot of people
B. has a great deal of sympathy for those less fortunate
C. **doesn't often give in to his impulses**
D. enjoys fantasising and daydreaming

Clearly, only option (C), 'doesn't often give in to his impulses' relates to the focal trait (conscientiousness) in the original person description.

Trait contextualization

For example, one item stated "Which of the following situations is most relevant to the trait of sociability?" Then, respondents had to select the most appropriate answer from three options (correct answer in bold):

A. A team member upon whom you rely allows her unanswered emails to accumulate and frustrate your co-workers in the process
B. **You notice that the time has just turned 1pm (which is your lunch time) and you see a few of your colleagues walking to the tea room**
C. You see that you colleague has been working non-stop since the morning

Footnotes

[1]In this paper, we use the term 'judges', 'raters', 'assessors' and 'interviewers' interchangeably.

[2]Note that we assess the conceptual criterion, in other words, theoretical interrelatedness of the three components of dispositional reasoning, in our discussion of positive manifold.

[3]We used Cohen's (1988) guidelines to interpret effect sizes for Pearson correlations, i.e. small (.10), medium (.30) and large (.50) effects. However, to interpret effect sizes for incremental validity, we used Cohen's (1988) guidelines for hierarchical regression f^2 as small (.02), medium (.15) and large (.35) effects.

[4]We calculated internal consistency reliability of the Wonderlic measure as .75, but acknowledge it is not appropriate for speeded measures (Nunnally & Bernstein, 1994).

[5]Because our pilot sample was small, we applied liberal cut-offs in the item analysis that would flag only those items that clearly fell short of our requirements for further consideration.

Chapter 5

Does it take one to know one? Interviewer personality, chronically accessible traits, and trait judgment accuracy[31]

The present study explores the relationship between interviewers' personality and their chronically accessible traits as predictors of trait-specific accuracy criteria. Earlier studies found relatively poor or inconsistent support for personality as a predictor of trait judgment accuracy. The present research has two objectives. First, it uses trait-specific accuracy criteria, as opposed to trait-generic accuracy criterion measures, allowing us to test whether interviewers' personality traits can predict trait judgment accuracy for specific corresponding traits (hypothesis 1), in other words, to determine "does it take one to know one?" The second objective is to establish whether interviewers' chronically accessible traits facilitate accuracy for those specific traits (hypothesis 2). We tested these ideas in two separate studies. In Study 1, college students (N = 183) completed a personality measure and rated the personalities of five hypothetical interview applicants depicted in vignettes constructed to mimic 'realistic' personalities. In Study 2, a field sample (N = 223) of interviewers completed the same measures, as well as a chronic accessibility measure. Results did not support the first hypothesis, as interviewers possessing elevated levels of a trait did not show trait expertise on the same trait. Related to the second hypothesis, accessibility for openness to experience and extroversion predicted judgment accuracy for the same traits. In turn, interviewer openness to experience and agreeableness predicted chronic accessibility for these traits. In sum, our results suggest that 'it may not take one to know one' in interview personality judgments. Moreover, chronically accessible personality traits may deserve a second look in studies of the good judge in personnel selection.

[31] This chapter is in preparation for publication review as:
De Kock, F. S., Born, M. Ph., Lievens, F., Gierdien, Z., & Sait, Z. (2015). *Does it take one to know one? Interviewer personality, chronically accessible traits, and trait judgment accuracy*. Manuscript in preparation.
Earlier versions of the two studies reported in this chapter were completed as Masters dissertations (conducted by the last two authors, under supervision of the first author) at the University of Cape Town, South Africa in 2013 and 2014.

5.1 Introduction

Personality judgments are increasing in importance in personnel selection as they not only underlie ratings used in selection devices such as interviews (Huffcutt, et al., 2001; Lievens, et al., 2001; Van Dam, 2003), but also hold many potential benefits over self-report measures of personality. These so-called 'observer ratings' (such as those by work colleagues or supervisors) can be more predictive of criteria (e.g. academic performance and job performance) and incremental to self-ratings (Connelly & Ones, 2010; Zimmerman, et al., 2010), while also demonstrating show good inter-rater reliability (e.g., corrected meta-analytic estimates of .46 - .62 for Big Five traits; Connolly, et al., 2007). It seems that those around us have 'clearer lenses' for viewing our personality traits (Connelly & Hulsheger, 2012) than we are able to ourselves. As such, observer ratings of personality in interviews are likely to be more construct-valid than self-report personality measures.

For these reasons, it is important to understand the factors that affect the accuracy of observer ratings of personality, such as interviewer characteristics. Although interviewer cognitive factors (Christiansen, et al., 2005; De Kock, et al., 2015) appear to be consistent predictors of personality trait judgment accuracy, interviewer personality traits show relatively inconsistent, poor, and often counterintuitive relationships (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003). The reasons for these findings are counterintuitive as they suggest that interviewer personality plays little to no role in person perception in organizations. However, personality is the predisposition to respond to stimuli in a certain way (John, Robins, et al., 2008) and it affects most areas of functioning, including social functioning and social judgment in the workplace (e.g., Tziner, et al., 2008).

Earlier studies on the relationship between interviewer personality traits and personality trait judgment accuracy have two major drawbacks. First, they often used trait-generic accuracy criteria that measure how well an interviewer is able to infer a complete personality profile, and try to predict this criterion from the interviewer's individual personality traits (as demonstrated in Figure 5.1).

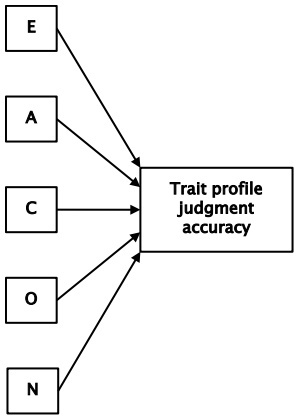

Figure 5.1. Predicting a trait-generic accuracy criterion, as done in earlier judgment accuracy research – Interviewer personality traits and trait profile judgment accuracy. E: Extraversion; A: Agreeableness; C: Conscientiousness; O: Openness; N: Neuroticism.

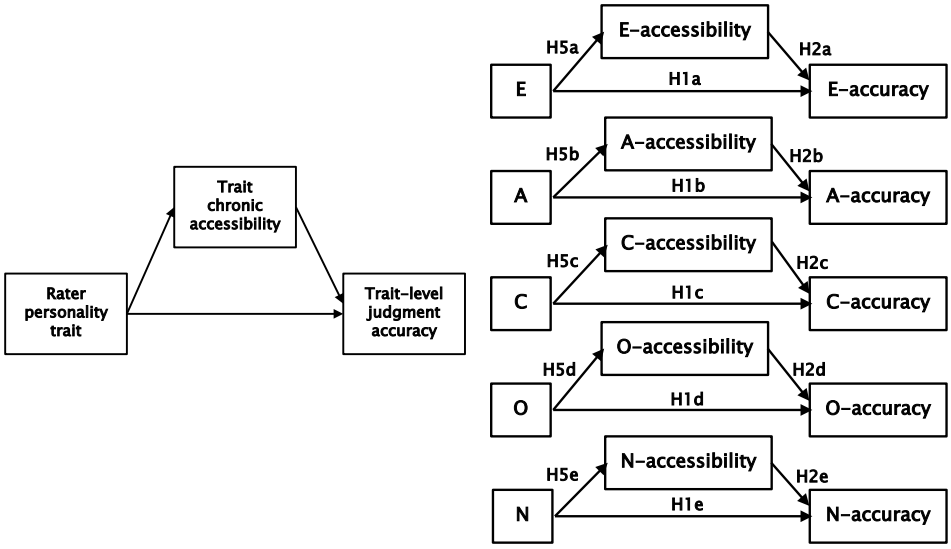

Figure 5.2. Study conceptual framework: Predicting trait-specific accuracy – Interviewer personality traits, trait chronic accessibility, and trait judgment accuracy. E: Extraversion; A: Agreeableness; C: Conscientiousness; O: Openness; N: Neuroticism.

A relatively unexplored avenue in individual difference research is the notion that personality trait judgment accuracy may be *trait-specific*, rather than trait-generic. As predicted by Funder (1995), it is now an established finding that traits differ in their judgability – accuracy scores for judging Big Five traits vary predictably (Allik, Realo, Mõttus, & Kuppens, 2010; Connolly, et al., 2007; Funder & Dobroth, 1987) often with lower accuracy for those traits that would best predict later job success (for example, conscientiousness, r = .16, n.s., and emotional stability, .17, n.s.; Barrick, et al., 2000). Given these findings, we know little about how interviewers' traits may predict trait-specific, or *trait-level* judgment accuracy (see Figure 5.2).

The second drawback of earlier studies was neglecting to offer an explanatory social-cognitive mechanism through which interviewer personality traits may affect their ability to judge others' traits. A growing line of research suggests that the self may be a basis for social cognitive schemas when forming impressions of others (Dunning & Cohen, 1992; Dunning & McElwee, 1995). We do not yet know whether interviewers' person perception processes and expertise for rating specific traits are related to their own personalities. For example, is it possible that extraverts would be better at rating extroversion? In other words, 'does it take one to know one?' in personality trait judgments? Interviewers may become trait experts because they are so familiar with their own traits, that is, they may show higher accuracy for judging traits that correspond with their own.

A plausible rival explanation may be that interviewers' own traits are more salient in their perceptual schemas. Drawing on construct accessibility theory (Higgins, 2012), we determine whether interviewers' chronically accessible traits, defined as the degree to which individuals differ in the readiness with which each construct is utilized in information processing of behavioral stimulus input (Higgins, et al., 1982, p. 45), predict their personality trait judgment accuracy. Moreover, would a more parsimonious account where chronic accessibility mediates (partially) the effect of an interviewers' personality trait on judging the corresponding trait in others be supported? To our knowledge, no earlier studies have tested these ideas empirically.

The present study

The present research has two objectives. First, it seeks to establish whether interviewers' personality traits can predict trait-specific judgment accuracy for conceptually aligned traits. The second objective is to determine whether chronically accessible may predict trait-specific accuracy for corresponding traits, both as direct effects, as well as mediators of, personality traits. To test these ideas, two empirical studies were conducted. In Study 1, we asked college students to infer from behavioral descriptors the personality profiles of hypothetical interview applicants. Students also provided self-reports of their own personalities, allowing

us to determine whether students' trait-specific judgment accuracy can be predicted from their own personality traits. In Study 2, we replicated the first investigation, but in a field setting. We asked interviewers in a large company to complete the same measures as in Study 1. But this time, we also determined their chronically accessible traits. We tested the prediction of trait-level judgment accuracy from interviewers' personality traits and their chronically accessible traits. Our analyses therefore reveal the potential incremental prediction of overall accuracy from interviewers' chronically accessible traits. Together, these studies aim to advance current understanding of the role of interviewer personality and personality trait construct accessibility as individual differences in trait judgment accuracy.

5.2 Study Background

Interviewer personality traits and judgment accuracy: Theory and prior research

Realistic Accuracy Model (RAM)

The process through which interviewer personality traits may facilitate judgment accuracy is explained by Funder's Realistic Accuracy Model (RAM) (Funder, 1995, 1999). Based on earlier models of perception (Brunswik, 1956), RAM proposes that the path to judgment accuracy involves a number of interdependent steps. The behavior displayed by a target serves as a cue and must (a) be *relevant* to the trait being judged, (b) *available* to the judge, (c) *detected* by the judge, and finally (d) correctly *utilized* to form an eventual personality trait impression. Only when all four steps have effectively been achieved, can a perceiver correctly judge another person's personality (Funder, 2012).

By implication, the RAM processes may or may not work effectively – the degree to which personality judgments are accurate is moderated by a number of key factors: good targets, good traits, good information, and finally, good judges. *Good targets* are simply highly judgable people – their behavior is relevant to their underlying personalities and they may be more transparent than poor targets. *Good traits* (e.g. expressiveness) are more visible than others (e.g. deceptiveness) and therefore, are more easily judged. *Good information* is a function of both quantity (e.g. a one-hour interview provides more trait cues than a speed-dating interview) and quality (e.g. when a person is relaxed and responds to good interview questions, higher quality cue information results). Finally, *good judges* are better able to detect and use behavior cues to form an accurate personality trait inference.

In the present study, our interest falls mainly on the individual difference characteristics of the *good judge*; more specifically: how do their personality traits and chronically accessible personality traits predict their judgment accuracy? Judges' personalities may regulate their social functioning in the workplace, including aspects of interpersonal judgment (e.g., Tziner, et al., 2008). RAM suggests

that these specific individual differences may either enhance, or detract from, interviewers' ability to engage the respective processes required for accuracy, especially those related to cue detection and cue utilization (Funder, 1999). Next, we report on prior empirical research and outline their key limitation: a reliance on trait-generic accuracy criterion measures.

Research Findings

Over the last few decades many researchers (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003) have been interested in whether or not certain traits predict overall trait judgment accuracy. In other words, these studies have tried to determine whether, for example, rater extraversion tends to predict accuracy, or not.

Conceptually, many arguments[32] have been advanced to link raters' personality traits to personality judgment accuracy (e.g., see Christiansen, et al., 2005). For example, agreeable individuals show more concern for others' feelings (Digman, 1990) and should, therefore, be more attuned to other individuals. Extraverts are known to seek out social interactions (Costa & McCrae, 1992) and, because of this increased social exposure, are likely to have more opportunity to hone their interpersonal judgments through practice and feedback. But due to their higher tendency to be focused on the self (Goldberg, 1992), extroverts may be less likely to detect behavioral cues signalling others' traits. Conscientiousness manifests in greater detail orientation (Goldberg, 1992) and, therefore, conscientious judges are likely to be more attentive in cue detection and show greater consistency in cue utilization, resulting in more accurate trait inferences. Persons higher in openness to experience are more inquiring and frequently enjoy working with abstract ideas or concepts (Goldberg, 1992), more likely to actively develop mental representations of other's traits and behavior (Kihlstrom & Hastie, 1997), seek patterns of consistencies and inconsistencies, and form and test hypotheses about behavior (Kruglanski & Ajzen, 1983) – all arguments in support of a predictive link between openness and personality judgment accuracy. In addition, openness tends to correlate with need for cognition, cognitive ability, and social intelligence – these characteristics have been linked to higher accuracy (Ackerman & Heggestad, 1997; Palmer & Feldman, 2005; Shafer, 1999). So, conceptually, there are numerous ways in which personality traits could potentially affect trait judgment accuracy.

The empirical evidence seems to suggest otherwise (see Appendix A). In contrast to their conceptual links with accuracy, observers' personality traits and judgment accuracy measures tend to correlate poorly (e.g. $0 < r < .20$, uncorrected) and, where they do show notable effects, results are generally inconsistent between studies (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005;

[32] For a comprehensive review of these arguments, see Chapter 2 of the present dissertation. We summarize the most pertinent ideas here.

Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003). In fact, an evaluation of empirical findings (see Chapter 2 of this dissertation) shows that no personality trait seems to emerge as a consistent predictor of judges' trait judgment accuracy. The bulk of evidence suggests that personality traits matter little in how accurate interviewers are able to produce impressions of others in personnel selection. There also are some traits that may also be detrimental to judgment accuracy, for example aggression (Borman, 1979), being domineering and vindictive (e.g., medium effects, Letzring, 2008), and neuroticism (Gibson, 2006). And counterintuitively, perhaps: judges that are less sociable may be more accurate than sociable individuals (e.g., Ambady, et al., 1995) – these authors suggested that personally and socially vulnerable individuals may be better at decoding nonverbal behavior (p. 526).

Taken together, a considerable gap exists between the conceptual and empirical bases for personality trait-judgment accuracy linkages[33]. On the one hand, clear reasons exist why most personality traits could enhance processes known to affect accuracy (cue detection and cue utilization; Funder, 1999). On the other hand, empirical evidence suggests otherwise. Following Occam's razor, the simplest explanation for the apparent disconnect is that personality is just not important to evaluate personality accurately, but more research is needed to rule out plausible rival hypotheses.

Trait-specific judgment accuracy

The stream of prior research on personality trait predictors of accuracy (e.g., Christiansen, et al., 2005; Powell & Goffin, 2009) has generally used judgment accuracy criteria that were not *trait-specific*. Instead, they have considered how an interviewer, for example, can evaluate interviewees accurately across a set of personality traits (considered together). Accuracy criterion measures like these can be described as *trait-generic*, that is, they are computed across target traits. The implicit assumption in these studies is that trait judgment accuracy generalizes across traits (i.e. people find all traits equally hard to judge) and, therefore, a single overall accuracy criterion may be determined at the personality-profile level.

However, it is now an established finding in empirical literature that traits do, in fact, differ in the degree to which they can be readily judged accurately (e.g., Allik, Realo, Mõttus, & Kuppens, 2010). In fact, in an earlier review of 32 studies, consensus between ratings of different observers looking at the same target person was higher for ratings of extraversion[34] (Kenny, et al., 1994). Also, in a recent study where students had to judge others' personalities from targets' essays, impressions

[33] For these reasons, we do not hypothesize any main effects between interviewers' personality traits and overall trait judgment accuracy.

[34] It must be noted that studies do not always agree on the relative difficulty of judging various Big Five traits.

of Openness to Experience were most accurate (Borkenau, et al., 2015). Such findings would manifest as mean differences in trait judgment accuracy scores for different traits, across a set of interviewers. This growing body of evidence corroborates Funder's (1999) hypothesis that 'good traits' are moderators of accuracy.

A further assumption of earlier investigations that used trait-generic accuracy criterion measures is that interviewers' individual accuracy profiles across a set of traits being judged are linear, and not differentiated. That is, an interviewer tends to be high, medium, or low in accuracy on all traits she is judging, or stated differently, she would show the same accuracy score for different traits judged. However, this assumption is yet to be tested. As traits are not equally well judged (both across, and within, interviewers) it makes sense to consider trait-specific-measures of judgment accuracy, instead of trait-generic measures.

Interviewer Personality Traits and Trait Judgment Accuracy

Trait-specific measures of judgment accuracy allow us to raise new questions about the role of interviewer personality traits in determining rating accuracy outcomes. For instance, would interviewer extraversion predict higher trait judgment accuracy for judging extroversion than for other traits (see Figure 5.2)? Stated differently, is there any substance to the adage that 'it takes one to know one'? As far as we know, our study is the first to explore various personality traits' as potential correlates of judgment accuracy criteria at the trait-specific level (for corresponding traits, that is). This notion is different from the 'similar-to-me'-effect (Rand & Wexley, 1975) where interviewers may form more positive impressions of demographically similar interviewees.

Interviewers may judge others' personality traits on the basis of their own personality traits as the self may be an important basis from which interviewers form impressions of others (Alicke, Dunning, & Krueger, 2005; Markus, Smith, & Moreland, 1985). For example, they may activate their own behaviors as norms when evaluating others: Dunning and Cohen (1992) found that 71% of participants reported comparing a target's observed behavior with their own behavior when forming judgments of targets – evidence that judgments of others' behavior was "egocentrically related to the participants' (judges') own behavior" (p. 213). This may occur through social projection (Krueger, 2007) which is a judgmental heuristic that allows people to make fast and relatively accurate judgments of others. In addition, contrast effects may occur when observers judge the behavior of others relative to their own, primarily to affirm their own self-worth (Beauregard & Dunning, 1998).

The effect of the self on judgments of others may go even deeper: Research shows that not only do observers form idiosyncratic definitions of traits and abilities (Dunning & Cohen, 1992; Dunning & McElwee, 1995) when judging others, they

may also assume that traits correlate in the population in the same way as they do in the self (or 'egocentric pattern projection'; Critcher & Dunning, 2009; Critcher, Dunning, & Rom, 2015). In other words, observers may project their implicit personality theories (IPT) on others. Finally, it seems that people may also build causal theories of how traits cause other traits in ourselves and then project these to targets when evaluating them (Critcher, et al., 2015).

We expect to find that interviewers are better at rating traits they hold themselves, as their familiarity with self-traits may enhance cue detection and cue utilization. This kind of trait interactions, that is, between an interviewer's traits and her ability to judge a specific trait in others, is called *expertise* in RAM. To illustrate trait expertise, an interviewer who is highly conscientious would often plan his schedule carefully (Goldberg, 1992) and, therefore, be able to detect and interpret this kind of behavior, i.e. other conscientiousness-cues, in interviewees' verbal responses to questions. Their enhanced cue detection and utilization ability would therefore lead to higher trait judgment accuracy (Funder, 1999). In sum, we posit:

> *Hypothesis 1*: Interviewers' personality trait levels will be positively correlated with trait judgment accuracy for corresponding traits, for extraversion (H1a), agreeableness (H1b), conscientiousness (H1c), openness to experience (H1d), and emotional stability (H1e)[35].

Chronically Accessible Traits, Personality, and Trait Judgment Accuracy

Personality traits not only regulate observers' social behavior, but they may also influence the perceptual lenses through which they view the social world. Already half a century ago Kelly (1955) stated that personal construct systems are "a kind of scanning pattern which a person continually projects upon his world. As he sweeps back and forth across his perceptual field he picks up blips of meaning" (p. 145). As such, some personal constructs may be employed more readily because of individual differences in the subjective meaning of social events (Mischel, 1973) and these become more salient in social perception. Construct *chronic accessibility* is defined as the degree to which individuals differ in the readiness with which each construct is utilized in information processing of behavioral stimulus input (Higgins, et al., 1982, p. 45). As people tend to encode others' behavior in terms of personality trait concepts (Wyer & Srull, 2014) some personality traits may become more accessible than others, that is, people show individual differences in accessibility for different constructs (Higgins, et al., 1982). From this perspective, some judges may be described as *chronics* or *non-chronics* for a particular trait. For example, a manager who tends to use mostly extroverted-related descriptors for

[35] We formulate hypotheses for all Big Five traits, regardless of the findings that show that some traits are easier to judge than others. This decision is based on our observation that studies do not seem to agree on which traits are considered 'good traits'.

applicants (for example, "…she was very outgoing" and "this candidate tended to initiate conversations during the exercise") would be described as an 'extroversion-chronic'.

Our conceptual model (see Figure 5.2) suggests a link between interviewers' personalities and their chronically accessible traits. Interviewers' chronically accessible constructs may partly stem from our own personality types[36] if the self is a basis from which people form impressions of others (Alicke, et al., 2005; Markus, et al., 1985). If true, it would manifest when neurotic interviewers, for example, also tend to have neuroticism as chronically accessible traits. In sum, we expect:

> *Hypothesis 2*: Interviewers' personality trait levels will be positively correlated with personality trait chronic accessibility for corresponding traits, for extraversion (H2a), agreeableness (H2b), conscientiousness (H2c), openness to experience (H2d), and emotional stability (H2e).

Chronic trait accessibility affects our social information processing as observed stimulus information that is related to an interviewer's accessible constructs would be more readily encoded, processed and retained than information that is related to inaccessible constructs (Higgins, 2012; Srull & Wyer, 1979; Wyer & Srull, 2014). Practically, it means that cues that are exhibited by applicants may be more readily remembered when they are related to the observer's chronically accessible traits. Also, "subjects make quicker judgments about a stranger along dimensions that are highly self-relevant (i.e. highly accessible) than along dimensions that are only moderately self-relevant"(Higgins, et al., 1982, p. 45). Construct accessibility is most likely a form of procedural memory (Smith & Branscombe, 1988).

When an interviewer has extremely high accessibility for a given trait (e.g. conscientiousness) then their ability to detect and use cues related to the same trait (e.g. going the extra mile in a work task, or being very lazy) may be higher. According to the Realistic Accuracy Model (RAM; Funder, 1999) better cue detection and use should result in higher trait-specific judgment accuracy for the same trait (e.g. higher conscientiousness judgment accuracy). Further, Wyer and Srull's 'storage bin' model suggests that schemas are used in impression formation (Wyer & Srull, 2014) and, when called upon to interpret behavior cues, people will access the most salient mental programs that lie at the very top of their storage bins (Uleman & Bargh, 1989). As such, an interviewer who is a conscientiousness-chronic (in other words, such an interviewer tends to think of others in terms of conscientiousness-related terms) may be better at correctly detecting and using cues related to conscientiousness (e.g. 'this person is always late for meetings'; Goldberg,

[36] Note that this argument is not analogous to the concept of *social projection*, which refers to the judgmental heuristic that manifests in the tendency to expect similarities between oneself and others (Robbins & Krueger, 2005).

1992) as information related to this schema is more accessible. In turn, it should also result in higher accuracy for rating conscientiousness.

Empirical research in performance ratings shows that, first, raters may differ in the degree to which performance-related dimensions are accessible for use, and second, dimensional performance evaluations will also be more accurate when the corresponding performance appraisal dimensions are accessible (Woehr, 1992). No earlier studies have considered the role of accessibility for individual personality traits and trait specific judgment accuracy, however. Therefore, we posit:

> *Hypothesis 3*: Interviewers' trait accessibility for each trait will be positively correlated with accuracy at judging corresponding traits, including extraversion (H3a), agreeableness (H3b), conscientiousness (H3c), openness to experience (H3d), and emotional stability (H3e).

Although we expect that accessibility for a trait would enhance judgment accuracy for the *same trait*, it should not extend to other Big Five traits where there is no conceptual alignment. In this line of argument, for example, higher accessibility for neuroticism would not enhance the detection and utilization of cues related to other personality traits (e.g. extraversion). As such, we expect an independence of trait accessibility and accuracy for judging non-corresponding traits. Put differently, accessibility for one trait should not be related to accuracy on other traits.

Given the arguments outlined above, our conceptual model posits chronic accessibility as a mediator of the influence of a trait on its corresponding trait judgment accuracy. Chronically accessible traits are more salient in interviewers' perceptual mechanisms and, therefore, may be considered the 'vehicle' through which interviewees' behavior cues are better detected and utilized. However, we think this mediating effect would be partial. Thus,

> *Hypothesis 4*: Interviewers' trait accessibility will partially mediate the effect of judges' traits on the corresponding trait-level accuracy, for extraversion (H4a), agreeableness (H4b), conscientiousness (H4c), openness to experience (H4d), and emotional stability (H4e).

Incremental validity of personality trait chronic accessibility

We are also interested in seeing whether or not chronically accessible traits can predict accuracy when controlling for the rater's personality traits. Such a finding would further support the potential importance of accessible traits as individual differences that may explain trait judgment accuracy:

> *Hypothesis 5*: Individual differences in interviewers' chronic accessibility for the Big Five traits would increment the validity of their traits to predict trait judgment accuracy.

5.3 Study 1 Method

Participants

The last author recruited 186 college students in an introductory psychology course to complete the measures after tutorial group meetings in 2013. Of these, 83.6% were female and 16.3% were male. Participants reported their race as White (54.1%), Black, (18%), Mixed Race (14.8%), Indian (4.9%), or Chinese (2.6%). A small percentage (4.4 %) preferred not to indicate their race, or indicated it as 'other' (2.2%). They were between 18 and 29 years old (M = 19.5 yrs, SD = 1.7 yrs). Students were from various faculties, including Social Sciences (35%), Science (32.8%), Arts (8.2%), Social Work (6%), and Commerce (2.76%). Some (15.3%) chose not to indicate their study directions. English was the official language of instruction at the South African university.

Procedure

The questionnaire administrators (course tutors) were briefed prior to the data collection. After concluding their tutorial group meetings, they introduced the research as a study of the relationship between assessor personality and trait judgment accuracy. Study participation was voluntary and participants were informed of their rights (to withdraw, anonymity, and confidentiality). Participation was incentivized by a prize draw (of roughly $30). Participants could choose to disclose their e-mail address in order to be contactable for awarding of the prize. Finally, they independently rated the five hypothetical candidates depicted in the experimental vignettes (discussed next) and filled in the individual difference measures (e.g. personality) and demographics measure before being debriefed and thanked for their participation. Most participants completed the measures in less than 30 minutes.

Materials

Applicant Vignettes

We decided to use written descriptions of five hypothetical applicants' personalities as stimuli in the rating task in our study. These vignettes were developed using guidelines for experimental vignette methodology (EVM) by Aguinis and Bradley (2014). Appendix D provides a detailed description of the development, pilot testing, and manipulation check of the experimental vignettes. Each vignette contained a profile of a mock interview target revealing behavioral cue descriptors related to traits of the Big Five personality dimensions to varying degrees. An example of one applicant ('Person A') can be found in Appendix B along with the underlying profile for this applicant. The applicant was described in a paragraph (of about ten sentences in length) containing behavior descriptors (key words and

phrases). For example, Person C[37] was described as 'At work, C is particularly detail-oriented and always strives for perfection. C loves order and regularity.' Also, 'However, C isn't necessarily comfortable amongst strangers and avoids excessive attention.' Where appropriate, vignettes contained intensity descriptors, such as 'always', 'often', 'occasionally', 'sometimes', and 'hardly' to denote the extent to which the target exhibited a particular personality dimension. To balance the full spectrum of target personalities in the stimuli, the five targets were designed to each have a single 'dominant' trait, for example, Person A was neuroticism-dominant ('5') and also low ('1') on agreeableness. Instead of designing random profiles, we relied on empirical evidence (e.g., meta-analytic evidence of Big Five trait intercorrelations; Goldberg et al., 2006) to avoid unrealistic personality profiles.

The judgment task consisted of two steps. First, participants were given a list of Big Five traits and descriptions of each trait (See Appendix B) in the form of adjectives, which they had to study. For example, 'Conscientiousness' was described as: 'Those high in conscientiousness are strong-willed and determined. They are also well-organized and have high aspiration levels. Those low in conscientiousness tend to procrastinate, may be unreliable, and are not very methodical.' Next, they were given the set of five target descriptions, with instructions to 'form an impression of each person's personality within the workplace context'. Students with no prior work experience[38] were asked to 'think of these behaviors in any study- or task-related role and not in a personal context'. For each target, they were given a list of all five traits, and asked to indicate the level of personality trait exhibited on a scale of 1 to 5 (1 = low indication of trait; 5 = strong indication of trait). They could refer back to the personality descriptions if they wanted to. We decided not to provide a position (job) description within the rating task as respondents' implicit trait policies (Motowidlo, Hooper, & Jackson, 2006) and idiosyncratic theories of performance (Hauenstein & Alexander, 1991) may have adversely affected our results; we wanted participants to focus on inferring the personality profiles of targets (i.e. a diagnostic judgment, instead of a predictive judgment) without a position in mind, as their implicit ideas about which personality traits may be relevant to the position may have been a confound in our study, if left uncontrolled.

Criterion Measures

Accuracy scores served as dependent variable in our study. Consistent with recent trait judgment accuracy studies (Christiansen, et al., 2005; De Kock, et al., 2015; Powell & Goffin, 2009) we computed an accuracy score for each participant, derived

[37] 'C' was used as a non-descript designator for each target as we were concerned that providing a name that provides gender or ethnicity cues may affect respondents' evaluations (Letzring, 2010).
[38] We did not ask respondents to report their prior work experience, but anecdotal evidence suggests that most students in this particular course have some degree of casual or formal work experience, albeit temporary positions as 'vacation work'.

from within-person profile correlations (in other words, between the profile inferred by the rater and the accuracy criterion profile of the target) (see Borman, 1977) at the dimension level, with an *r*-to-Fisher's-*z* transformation. This method assesses the congruence (see Funder & Colvin, 1997) between the complete set of judgments made by an interviewer and the target's accuracy criterion profile. An example of a hypothetical personality profile is listed at the end of Appendix A. The development and evaluation of the accuracy criterion profile are explained in Appendix D.

We used accuracy scores at two different levels: A trait-specific judgment accuracy score was calculated as the correlation between interviewers' ratings for a trait (across targets) and the pre-determined 'true scores' for the same trait, whereas a trait-generic (profile) judgment accuracy criterion was calculated as the correlation between interviewers' ratings for all traits (across targets) and 'true scores' for all traits.

Predictor Measures

Personality. All participants completed the 44-item Big Five Inventory (BFI) (John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008). Participants had to indicate agreement with the statements on a 5-point Likert-type scale (*strongly disagree* to *strongly agree*. An example of an item was "Is relaxed, handles stress well." Earlier studies (John, et al., 1991; John, Naumann, et al., 2008) reported satisfactory reliability, convergent validity and discriminant validity.

Demographic Measure. In addition, participants completed a basic demographic measure, including gender, race, first language, study direction/faculty, age, etc. Participants were asked to indicate their informed consent for voluntary participation in the study by ticking a box on the questionnaire cover sheet. A category 'prefer not to answer' was added to the gender and race items, as required by the relevant university ethics protocol.

5.4 Study 2 Method

Participants

Our second study used 223[39] respondents from a large organization in the financial services sector in South Africa. Data for this study were collected in 2014 by the fourth author, who was an HR practitioner within the organization at the time. Of these, 62.3% were female and 35.4% were male, with the remainder (2.2%) not indicating their gender. Participants reported their race as White (46.6%), Mixed Race (27.8%), Black (11.2%), Indian (9.4%), other (.4%), or did not specify (4.4%). The participants' ages in years ranged from 21 to 65 (M = 38.16 yrs; SD = 9.4 yrs).

[39] Of 551 participants that opened the survey link, only 223 participants (40.5%) completed the full set of measures; judging from case demographic information, no obvious withdrawal pattern was evident and, therefore, incomplete responses were deleted from further analysis.

Although English was the official workplace language, the first language was reported as English (52.9%), Afrikaans (33.2%), Xhosa (6.7%), other (4.9%), or not specified (2.2%). Participants were relatively well-educated (postgraduate degree: 38.1%; graduate degree: 36.8%; high school matriculation certificate: 22.9%; not reported: 2.2%) and represented staff at all organizational levels (junior management/staff: 57%; middle management: 28.3%; senior management: 11.7%; top management: .9%; not specified: 2.2%). A large portion of the participants (54.3%) often completed performance appraisals of others and a substantial number (22%) often conducted job interviews.

Procedure

An e-mail invitation to participate in the survey was distributed to all staff in the organization. The invitation included a cover letter that explained the nature and aim of the study. After accessing the web-link to our online survey platform (Qualtrics Development Company, 2015) participants were informed of their rights (voluntary participation, right to withdraw, confidentiality, anonymity) and asked to indicate their informed consent. Participation was incentivized by a prize draw (shopping voucher to the value of approx. $50). Participants could choose to disclose their e-mail address in order to be contactable for awarding of the prize. Finally, they completed the survey measures before rating the five hypothetical candidates depicted in the experimental vignettes (described in Materials). Anecdotal evidence[40] suggests that most respondents completed the measure during work time at their place of work.

Materials

Study 2 used the same materials as Study 1 – BFI and demographic measure – with the addition of a measure of chronic accessibility.

Chronic Accessibility. We elicited respondents' accessible traits with a free-response measure of accessible constructs (Higgins, et al., 1982). The full measure is presented in Appendix C. Respondents were requested to describe five 'persons': (a) a type of person they sought out, (b) a type of person they avoided, (c) a type of person they liked, (d) a type of person they disliked, and (e) a type of person they frequently encountered. Respondents could free-recall and write down (up to nine) characteristics that best described each person. These descriptors were coded into Big Five trait category terms and scored using a variation of Higgins' (1982) output primacy operationalisation to yield a trait accessibility score for each trait.

To code participants' responses, the last two authors exported trait descriptors provided for each 'person' into an Excel spreadsheet. They compiled a database of Big Five personality descriptor terms from published empirical research (e.g.,

[40] We inspected the survey login data on Qualtrics that reported login time and geographic location as determined by IP addresses. Results showed that 96.4% of respondents logged on for the survey before 6pm. Also, as the study invitation was sent in an email to participants' work email addresses, they often completed the measure within their work environment.

Goldberg, 1990, and others; Hofstee, de Raad, & Goldberg, 1992; John, 1990; Saucier & Goldberg, 1996) which would serve as an objective coding scheme for trait descriptors. The coders, two separate master's students in I-O Psychology, coded participants' trait descriptors into a single Big Five trait category by referencing words against the coding scheme spreadsheet. The two individuals coded the first 15% of trait terms together in order to establish a common process and interpretive standard. As a large number (5390) of trait terms had to be coded, the coders randomly split the remaining trait descriptors among themselves and independently coded these. Where exact matches for terms in the coding spreadsheet were not found, synonyms and antonyms were sourced from a Thesaurus. Failing subsequent further matches using alternative terms for a yet unmatched descriptor, coding was discussed between coders in relation to the original trait definition (and its sub-facets) until consensus was reached. Some descriptor terms could not be coded into the Big Five typology, for example physical characteristics (e.g. 'old people'), roles ('my boss'), and categories of people ('soccer players') – these were coded as missing values. Overall, 5 390 trait descriptors were coded into the Big Five traits. The average number of trait descriptors given by each respondent was 20.91 (*SD* = 7.91). At most, they could list a maximum of 45 descriptors (five 'persons' to be described and nine descriptors for each). The various hypothetical persons received a varying proportion of overall descriptors provided, 'liked'= 24.7%, 'disliked' = 20.6%, 'frequently encountered' = 17.3%, 'avoided' = 17.2%, and 'sought out' = 20.1%.

After coding the individual descriptors into personality traits, we derived a respondent's chronic accessibility score for each trait. To this end, we relied on an operational definition of chronic accessibility based on output primacy (Higgins, et al., 1982), that is, one that reflects a given trait's order of appearance (i.e. the first trait that comes to mind when describing a person). Practically, a person's accessible traits were only those they listed first in response to the five 'persons' to be described. Although some studies have demonstrated the construct validity of this operationalisation (Bargh, Bond, Lombardi, & Tota, 1986; Bargh & Thein, 1985; Higgins, et al., 1982) it results in a dichotomisation where each trait is described as accessible or not. In our view, a more nuanced operationalisation is needed where traits' accessibility can be described on a continuum of low to high, which would be more consistent with our understanding of trait accessibility as the relative salience of various traits in interviewer's perceptual schemas. As such, we also determined the frequency[41] that a trait was indicated first (i.e. output primacy) but across the five 'persons' described. For example, a respondent that listed agreeableness-related descriptors (e.g. 'a warm person') for all five persons would receive a score of 5, whereas the original Higgins (1982) operationalisation would yield an accessibility

[41] Despite our reservations about the original output primacy scoring approach, we calculated scores for each participant using both operationalisations. The correlation was high (mean r = .80 across Big Five traits, uncorrected for unreliability, nor range restriction).

score of '1' (i.e. agreeableness-accessible). In contrast, for example, the original Higgins' operationalisation would not reflect the relative differences in accessibility for two respondents, one listing 'agreeableness' first for two of the 'persons', and another, who lists agreeableness first for all five of the 'persons'.

5.5 Results

Descriptive Statistics

Table 5.1 presents the means, standard deviations, and correlations of the variables for Study 1. Table 5.2 presents the same information for Study 2. Figures 5.3 and 5.4 present the central tendency and variability, respectively, for trait accessibility and trait judgment accuracy for each trait, as found in Study 2.

In both studies, our results indicate that participants found it easier to judge conscientiousness (Mean Fisher's z accuracy scores[42] $M_{study\ 1}$ = 1.10 ;$M_{study\ 2}$ = 1.00), agreeableness ($M_{study\ 1}$ = 1.06; $M_{study\ 2}$ = .96), and neuroticism ($M_{study\ 1}$ = .96; $M_{study\ 2}$ = .88) from the vignettes than they did for openness ($M_{study\ 1}$ = .85; $M_{study\ 2}$ = .77) and extroversion ($M_{study\ 1}$ = .79; $M_{study\ 2}$ = .76), in that order, across the five target persons being rated on a response scale from 1 = low indication of trait to 5 = strong indication of trait). These means are illustrated graphically in Figure 5.3. Earlier studies reported relatively small differences in observer agreement for judging different Big Five traits (for a review of various studies on this issue, and important note on statistical artefacts, see Allik et al., 2010). Also, the standard deviation of the accuracy scores (see Figure 5.4) for the different traits suggests sizeable individual differences (.48 < Mean Fisher r_z < 1.34).

In study 2, our results indicate that participants showed comparatively higher accessibility (see Figure 5.3) for agreeableness (M = 2.35) and extroversion (M = 1.01), across the five targets being rated, and had the lowest accessibility for openness (M = .35). So, it seems like managers and professional staff (in our sample) showed a tendency to use behavior descriptors associated with how agreeable others are (or are not), rather than using descriptors associated with how open to experience others are (or are not). Considering the variability in trait accessibility (see Figure 5.4), exactly the same pattern emerged, that is, the sample was most heterogeneous in their accessibility for agreeableness (SD = 1.46), than for example, openness to experience (SD = .67), which showed the least variability of the accessibility scores. It is also noteworthy that, per definition, the intercorrelation between trait accessibility scores for the various Big Five traits are mostly negative (-.38 < r < .03; see Table 5.2).

[42] These accuracy score means are Fisher-transformed profile correlations ($r_{z,}$) between judges' ratings for a trait, across five targets, and 'true' scores.

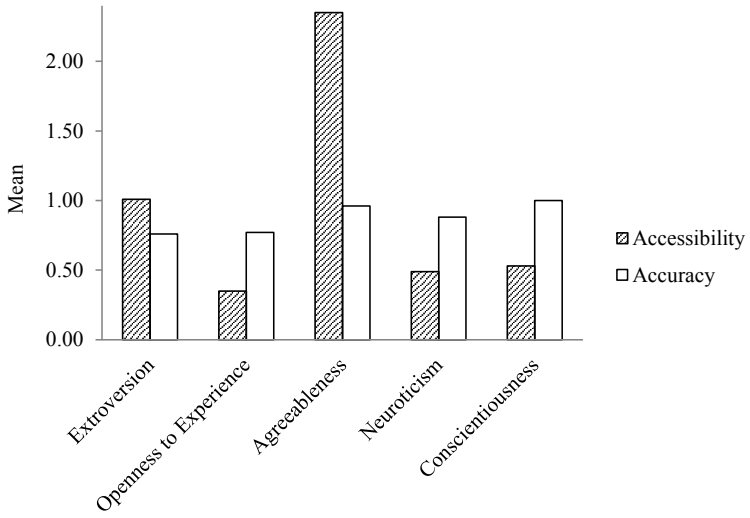

Figure 5.3. Central tendency (Mean) scores for trait judgment accuracy (Mean Fisher's r_z) and trait chronic accessibility (Mean proportional frequency of descriptors) by trait of managers/staff (in Study 2).

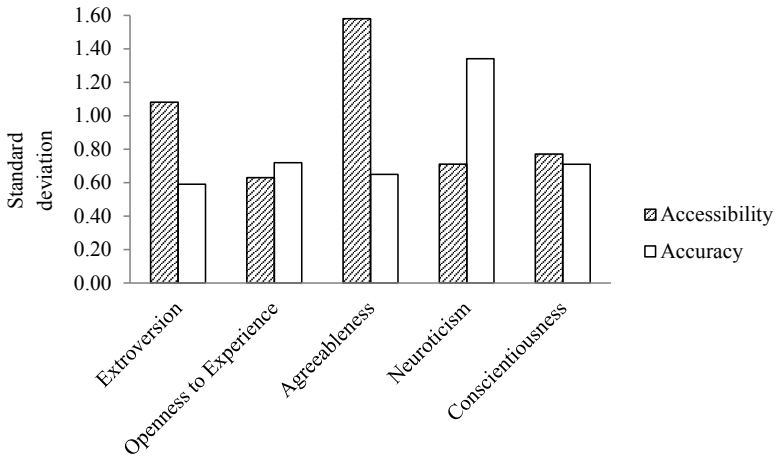

Figure 5.4. Variability (Standard Deviation) of trait judgment accuracy scores (SD of Fisher's r_z) and trait chronic accessibility (Mean proportional frequency of descriptors) by trait of managers/staff (in Study 2).

Table 5.1

Study 1: Descriptive Statistics and Intercorrelations of all Study Variables

Variable	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
1. Gender[a]	1.84	.37	-												
2. Age	19.54	1.67	-.16*	-											
3. Assessor agreeableness	3.70	.54	.01	.01	-										
4. Extroversion	3.30	.71	.04	-.01	.19*	-									
5. Conscientiousness	3.34	.65	.18*	-.04	.39**	.25**	-								
6. Neuroticism	3.02	.72	.08	-.01	-.27**	-.27**	-.05	-							
7. Openness	3.66	.51	-.10	.05	-.02	.10	.06	-.07	-						
8. Agreeableness Accuracy[b]	1.06	.68	.16*	-.06	.04	-.04	-.05	.01	-.09	-					
9. Extroversion Accuracy[b]	.79	.60	.17*	-.28**	.00	-.21**	.01	.03	.01	.33**	-				
10. Conscientiousness Accuracy[b]	1.10	.68	.14	-.12	-.07	-.16*	-.05	.14	-.02	.41**	.27**	-			
11. Neuroticism Accuracy[b]	.96	1.06	.10	-.18*	-.05	-.06	-.11	.01	.03	.18*	.20**	.22**	-		
12. Openness Accuracy[b]	.85	.71	-.05	-.15*	.07	-.10	.01	.05	.10	.23**	.32**	.31**	.22**	-	
13. Overall Accuracy[c]	.96	.48	.15*	-.25**	-.01	-.16*	-.06	.06	.02	.63**	.61**	.66**	.67**	.63**	-

Note. $N = 183$.

[a]Gender was coded such that men were 1 and women were 2. [b]Trait-specific accuracy scores were Fisher transformed (r to z) profile correlations between participants' ratings at item level and true scores. Higher scores denote higher trait judgment accuracy. Overall Accuracy is the linear composite (mean) of the five trait-specific accuracy scores for the Big Five traits.

* $p < .05$. ** $p < .01$ (two-tailed).

Table 5.2

Study 2: Descriptive Statistics and Intercorrelations of Study Variables

Variable	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1. Gender[a]	1.64	.48	-																
2. Age	38.16	9.37	.00	-															
3. Agreeableness	3.96	.47	.24**	.11	-														
4. Extroversion	3.23	.68	.01	-.05	.07	-													
5. Conscientiousness	4.01	.52	.23**	.11	.43**	.17*	-												
6. Neuroticism	2.51	.59	.08	-.19**	-.48**	-.34**	-.35**	-											
7. Openness	3.54	.48	-.22**	-.01	-.07	.31**	.04	-.21**	-										
8. Agreeableness CA[b]	2.35	1.46	.01	.03	.14*	.02	.08	.05	-.16*	-									
9. Extroversion CA[b]	1.01	1.10	.11	-.12	-.10	-.13	-.06	.09	.00	-.38**	-								
10. Conscientiousness CA[b]	.53	.81	-.09	.03	.05	.06	.08	-.13	.03	-.33**	-.20**	-							
11. Neuroticism CA[b]	.49	.75	.08	.08	.02	-.06	-.06	-.03	-.04	-.33**	-.11	-.04	-						
12. Openness CA[b]	.35	.67	-.21**	-.01	-.13	.11	-.13	-.12	.34**	-.26**	-.14*	.03	-.02	-					
13. Agreeableness Accuracy[c]	.96	.65	.01	.11	.00	.05	.10	-.07	.00	-.01	.11	-.01	.07	.00	-				
14. Extroversion Accuracy[c]	.76	.49	-.10	.05	-.12	.02	-.02	-.01	.11	-.03	.19**	-.13	.08	.04	.25**	-			
15. Conscientiousness Accuracy[c]	1.00	.61	.02	.00	-.12	-.09	.09	.07	.05	.11	.03	-.13	.04	-.02	.26**	.17*	-		
16. Neuroticism Accuracy[c]	.88	1.34	-.10	.00	-.13	.04	-.04	.03	.14*	-.07	.16*	-.07	.11	.04	.15*	.28**	.24**	-	
17. Openness Accuracy[c]	.77	.64	-.08	.05	.01	.05	-.03	-.14*	.04	-.06	.03	.00	-.01	.21**	.19**	.19**	.20**	.17*	-
18. Overall Accuracy[d]	.87	.48	-.10	.07	-.12	.03	.00	-.01	.11	-.04	.19**	-.11	.11	.09	.56**	.54**	.57**	.78**	.52**

Note. N = 181 to 223.

[a]Gender was coded such that men were 1 and women were 2. [b]Chronic accessibility was assessed using a variant of Higgins et al.'s (1982) method. Higher scores denote higher trait accessibility. [c]Trait-specific accuracy scores were Fisher transformed (*r* to *z*) profile correlations between participants' ratings of all targets at dimension level and true scores. Higher scores denote higher trait judgment accuracy. [c]Overall Accuracy is the linear composite (mean) of the five trait-specific accuracy scores for the Big Five traits.

*$p < .05$. ** $p < .01$ (two-tailed).

Tests of Hypotheses

H1: Interviewer personality traits and trait-level judgment accuracy

Hypothesis 1 proposed that interviewers' trait levels will be positively correlated with trait-level judgment accuracy for corresponding traits. The correlations between interviewers' traits and accuracy for judging the corresponding traits revealed generally trivial to small effects (Cohen, 1988). For example, in Study 1 (see Table 5.1), these were non- significant and negligible for agreeableness (.04), conscientiousness (-.05), neuroticism (.01), and openness (.10) (all $p > .05$). The single exception was extroversion, which showed a small-to-medium effect, but it was in fact associated with *lower* extroversion accuracy ($r = -.21$, $p < .01$). Similarly, in Study 2 (see Table 5.2), no traits correlated significantly with accuracy for the same trait ($.04 < r < .09$) (all $p > .05$). Therefore, Hypothesis 1 was not supported.

H2: Interviewer personality traits and chronic accessibility for corresponding traits

Hypothesis 2 proposed that interviewers' personality trait levels would be positively correlated with trait accessibility for corresponding traits. As shown in Table 5.2, openness (.34, $p < .01$) and agreeableness (.14, $p < .05$) showed small-to-moderate effects on accessibility for the same trait, but the effects for conscientiousness (.08, $p = .24$), neuroticism (-.03, $p = .68$), and extroversion (-.13, $p = .07$) were negligible to small. Therefore, Hypothesis 2 was partially supported.

H3: Trait chronic accessibility and trait-level judgment accuracy

Hypothesis 3 stated that interviewers' trait accessibility will be positively correlated with trait-level judgment accuracy for corresponding traits. Table 5.2 shows that the findings partly confirmed Hypothesis 2. Interviewers who had openness ($r = .21$, $p < .01$) and extroversion ($r = .19$, $p < .01$) as accessible traits had higher accuracy for the same trait but not for other traits. Effects for other personality traits did not reach significance, for example neuroticism ($r = .11$, $p = .13$), agreeableness ($r = -.01$, $p = .85$), and conscientiousness ($r = -.13$, $p = .08$). In sum, evidence regarding the links between a trait's accessibility and its trait judgment accuracy is divided. Therefore, our results partially supported Hypothesis 3. We also expected that interviewers' trait chronic accessibility would not be positively correlated with trait-level judgment accuracy for non-corresponding traits. Table 5.2 shows support for this idea: almost none of the off-diagonal correlations between accessibility and accuracy for a trait reached statistical significance, with the exception of one (extraversion chronic accessibility and neuroticism trait judgment accuracy; a small effect: $r = .16$).

H4: Mediating effects of trait chronic accessibility

Hypotheses 4 proposed that interviewers' trait accessibility will partially mediate the effect of their traits on the corresponding trait-level accuracy criteria, for extraversion (H4a), agreeableness (H4b), conscientiousness (H4c), openness to experience (H4d), and emotional stability (H4e). Table 5.3 summarizes the results of the mediation analyses. In order to test the hypothesis that trait chronic accessibility would mediate the relationship between a trait and accuracy for judging the

corresponding trait, we conducted mediation analysis (Baron & Kenny, 1986) with standard regression techniques and calculated the normal theory (i.e. Sobel, 1982) test for the total and specific indirect effect. In addition, we also calculated percentile-based, bias-corrected, and bias-corrected and accelerated bootstrap confidence intervals for the indirect effects using resampling (Preacher & Hayes, 2008). A SPSS macro and script (Hayes, 2014; Preacher & Hayes, 2004) were used that generated 95% bias-corrected and accelerated bootstrap confidence intervals for all indirect effects and contrasts of indirect effects using $z = 1000$ bootstrap samples.

The results showed that the total indirect effect of personality traits on trait-specific accuracy through chronic accessibility was not significant for all traits. There was a significant indirect effect of openness on openness judgment accuracy through openness accessibility, $b = .107$, Bootstrap estimate BCa CI [.040, .195]. As the confidence interval does not include zero, it means that this is likely to be a genuine effect, that is, openness accessibility is a mediator of the relationship between openness and openness judgment accuracy. However, this represents a small effect size, $\kappa^2 = .08$, 95% BCa CI [.030, .146]. The mediation results for the remainder of the personality traits are reported in Table 5.3. None of these trait accessibility scores were significant mediators of their corresponding traits on trait judgment accuracy.

Table 5.3

Results of Mediation Analysis of the Indirect Effect of Interviewers' Personality Traits on Trait-specific Judgment Accuracy[a] through Trait Chronic Accessibility[b]

Predictor	Estimates			Effect Size
	β	SE	95% BCa CI	κ^2 and 95% BCa CI
Agreeableness	-.003	.017	[-.042; .026]	.002 [.000; .008]
Extraversion	-.017	.011	[-.046; -.002]	.026 [.004; .065]
Conscientiousness	-.013	.014	[-.061; .002]	.012 [.000; .051]
Neuroticism	-.008	.019	[-.066; .019]	.003 [.000; .020]
Openness	.107	.041	[.040; .195]	.080 [.030; .146]

Note. $N = 191$ (listwise). β = Estimate of indirect effect. SE = bootstrap estimate of standard error. BCa CI = Bias-corrected and accelerated confidence interval (95%). κ^2 [CI] = Preacher and Kelly (2011) Kappa-squared. [a]Accuracy scores are Fisher transformed (r to z) profile correlations between participants' ratings of targets' overall personality profiles and subject matter expert true score estimates. [b]Chronic accessibility was assessed using a variant of Higgins et al.'s (1982) method. Our frequency-scoring approach is interpreted such that higher scores denote higher personality trait chronic accessibility.

*$p < .05$. **$p < .01$.

H5: Incremental validity of trait chronic accessibility

Hypothesis 5 posited that interviewers' chronic accessibility for the Big Five traits would increment the validity of their traits to predict trait-specific accuracy for corresponding traits. In Step 1, the Big Five trait measure score was entered. In Step 2, we entered the corresponding trait chronic accessibility measure score. Inconsistent with our hypothesis, when trait accessibility measures were added in Step 2 for each trait, results revealed statistically *in*significant increments in the ability to explain trait-specific judgment accuracy for agreeableness (ΔR^2 = .00, p = .76), conscientiousness (ΔR^2 = .01, p = .14), neuroticism (ΔR^2 = .01, p = .13), and openness (ΔR^2 = .00, p = .51). In contrast, the incremental validity for extraversion (ΔR^2 = .04, p = .01) accessibility was significant with a small incremental effect size[43] (Cohen's f^2=.04). So, the addition of chronic accessibility to the equation with personality traits generally resulted in statistically insignificant increments in R^2, with the exception of extraversion. These trivial effect sizes that we observed for the incremental effect of trait accessibility did not support Hypothesis 5 for most traits.

Additional Analyses

Interviewer Personality Traits and Overall Trait-Generic Judgment Accuracy

We also used the trait-generic measure[44] of judgment accuracy as criterion and correlated these with interviewers' self-reported personality trait scores. Tables 5.1 and 5.2 report the correlations for Study 1 and 2, respectively. Personality traits of the interviewers were generally unrelated ($-.16 < r < .11$) (all $p > .05$, except for extroversion in Study 1) to overall trait judgment accuracy. In Study 1, extroversion was in fact associated with *lower* overall trait-generic judgment accuracy (r = -.16, p < .05).

Incremental Validity Analyses for Trait-Generic Accuracy

In addition the trait-specific incremental validity analyses, we were also interested to see whether interviewers' chronically accessible traits would predict a trait-generic accuracy criterion and increment personality traits in doing so. Out of all five trait accessibility scores, only extroversion accessibility predicted the trait profile judgment accuracy criterion (r = .19, p < .01; Study 2, Table 5.2).

Table 5.4 summarizes the results of the hierarchical regression analyses. In Step 1, the Big Five trait scores were entered. In Step 2, we entered chronic accessibility measures as a set. Results revealed a significant increment in the ability to explain trait profile judgment accuracy (ΔR^2 = .08, p = .006) when trait accessibility measures were added in Step 2. So, the addition of chronic accessibility to the equation with

---

[43] The effect size for hierarchical multiple regression (Cohen, 1988) calculated as $f^2_B = (R^2_{AB}-R^2_A)/(1-R^2_{AB})$, where R^2_A is the variance accounted for by the set of personality trait independent variables A, and R^2_{AB} is the combined variance accounted for by A and the additional set of accessibility variables for each trait B. To interpret effect sizes for incremental validity we used Cohen's (1988) guidelines for hierarchical regression f^2 as small (.02), medium (.15) and large (.35) effects.

[44] Operationalized as a linear composite (mean) of accuracy scores across all Big Five target traits.

personality traits resulted in a statistically significant increment in R^2 when predicting an interviewers' accuracy for judging a profile of traits (small to medium effect size[3]; Cohen's f^2=.09).

Table 5.4

Results of Hierarchical Regression Analyses of Interviewers' Overall (Profile) Judgment Accuracy[a] on Personality of the Interviewers and Chronic Accessibility of the Interviewers

Predictor	Step 1		Step 2	
	β	*t*	*β*	*t*
Step 1				
Agreeableness	-.16	-1.79	-.17	-1.93
Extraversion	-.05	-0.68	-.02	-0.26
Conscientiousness	.02	0.21	.05	0.60
Neuroticism	-.03	-0.34	-.03	-0.32
Openness	.13	1.61	.11	1.41
Step 2				
Agreeableness accessibility[b]			.26*	2.31
Extraversion accessibility			.33**	3.42
Conscientiousness accessibility			.04	0.48
Neuroticism accessibility			.24**	2.89
Openness accessibility			.15	1.74
Total R^2	.04		.12**	
ΔR^2	.04		.08**	

Note. N = 191 (listwise). The effective sample size was reduced due to some participants that did not provide a sufficient number of trait descriptors to calculate a personality trait chronic accessibility score.

[a]Accuracy scores are Fisher transformed (r to z) profile correlations between participants' ratings of targets' overall personality profiles and subject matter expert true score estimates. [b]Chronic accessibility was assessed using a variant of Higgins et al.'s (1982) method. Our frequency-scoring approach is interpreted such that higher scores denote higher personality trait chronic accessibility.
*p < .05. **p < .01.

5.6 Discussion

Main conclusions

The present study attempted to further the search for characteristics of the 'good judge' of personality in personnel selection. More specifically, it investigated the role of the interviewers' personality traits and chronically accessible traits in personality trait judgment accuracy. A major departure from earlier work that relied on trait-generic accuracy criteria was that we studied individual difference predictors of trait judgment accuracy at the level of individual target traits, or trait-specific accuracy criteria. This enabled a test of how interviewer traits may predict trait judgment accuracy of the corresponding traits of targets. Our study also introduced chronically accessible traits (Higgins, 2012) as potential explanatory mechanisms to explain why it may "take one to know one" in personality judgment.

At the most basic level, our findings are in line with earlier studies (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003) that found little evidence for interviewer personality traits as correlates of personality judgment accuracy. However, our two studies provide early support for the possibility that this may extend to trait-specific accuracy criteria. We expected that our study participants would be more accurate at judging traits they shared with targets, but in both studies this was not the case (i.e. all trivial to zero effects, Cohen, 1988) suggesting that it is possible that being elevated on a trait may not necessarily make one more adept at judging it.

Further, our research drew on construct accessibility theory (Higgins, 2012) and hypothesized that interviewers' chronically accessible personality traits may predict corresponding trait-specific accuracy criteria. As trait chronic accessibility represents the degree to which interviewers differ in the readiness with which each trait is utilized in behavioral information processing (Higgins, et al., 1982) it would make sense that trait information that is salient in the interviewers' perceptual schema would enhance the detection and use of cues related to traits that are accessible to the interviewer. Our hypothesis received partial support, as openness to experience and extraversion accessibility predicted trait judgment accuracy for the same traits. This finding did not extend to the remaining Big Five traits – accessibility for agreeableness, conscientiousness, and neuroticism did not predict accuracy for corresponding traits. In sum, trait construct accessibility may contribute to interpreting cues for the same trait, but much more research is needed before a firm conclusion may be reached. In our results, accessibility for openness to experience and extraversion may have featured more strongly than others in our results as they may be 'good traits' (Funder, 2012) that are more visible and therefore easier to judge (Funder & Dobroth, 1987).

Further, accessibility for some traits (e.g. openness and agreeableness) appeared to 'spill over' from the interviewers own corresponding personality traits: interviewers high on agreeableness also tended to describe others in agreeableness-related terms. Taken together, our finding supports the view that the self may hold potential clues for how we judge others (Dunning & McElwee, 1995; Motowidlo, et al., 2006). As far as we know, this is a novel finding that may suggest a possible link between observers' personalities and the social-cognitive schemas they employ.

Whether trait accessibility tells us more about interviewer's judgment accuracy, than knowing about their personality alone, is still an open question. Our trait-level analyses showed that interviewers' chronically accessible traits did not increment personality traits in predicting trait-specific accuracy, running counter to the idea that salient social-cognitive information (due to higher accessibility for a trait; Higgins, 1982), when separated from the interviewers' personality, may enhance cue detection and utilization for the same trait. In addition, our mediation analysis showed that accessibility only interviewer openness mediated the effect of openness on accuracy for judging the same trait (i.e. openness to experience). For other traits, accessibility did not act as a mediator. More work is needed to explore the strength of these findings before conclusions may be drawn about them.

Limitations

Our study had some limitations relating to generalizability issues. We used stimuli that were not real people and, therefore, the generalizability of our results to field settings must be established. Overall, the benefits of using experimental vignette methodology (EVM; Aguinis & Bradley, 2014) outweighed its limitations. For example, using standard stimulus cues, with pre-determined true score estimates on the trait dimensions, allowed the determination of accuracy scores for a large group of respondents – this would have been very difficult to achieve in a field setting where standardized performance information cannot readily be presented to all interviewers.

Other advantages to EVM in our study involved standardisation and control: By presenting the same stimulus materials to all respondents we could hold trait-related information constant and minimize error variance related to using real interviewees. We tried to enhance the realism of the study in a number of ways. For example, the vignettes contained profiles that were constructed from empirical meta-analytic estimates of how personality traits are correlated in the real world. They also contained behavior cues that have been empirically linked to each Big Five trait. The resulting vignettes were manipulation-checked for realism and judgability. Our second study also used field-sample respondents who routinely conduct job interview ratings and they completed our measures in their work environment. However, future studies should also try to replicate our study in contexts closer to the field setting, for example, using video-based stimuli of interviewee behavior.

Implications for Theory and Future Research

Our studies presented here may have implications for theory and future research. Although the 'good judge' may be an important moderator[45] of accuracy (Funder, 1995, 2012) our results suggest that personality-related characteristics of interviewers are less relevant to accuracy outcomes than their cognitive counterparts, such as cognitive ability and dispositional reasoning (*cf.* Christiansen, et al., 2005). The effect sizes for cognitive predictors of accuracy are generally moderate-to-large (See our review in Chapter 2) whereas the two studies presented here revealed that, in line with the earlier work summarized in Appendix A, interviewer personality traits had trivial-to-zero effects (Cohen, 1988) when we used both trait-generic and trait-specific accuracy measures.

Our results add to the growing evidence that the empirical basis for personality as a characteristic of the 'good judge' is not strong. We are not sure if these findings would extend to narrow trait-predictors of accuracy, though. For example, in some other contexts (e.g., predicting academic criteria; De Vries, De Vries, & Born, 2011) narrow traits were better predictors of criteria than broad traits.

Another interesting avenue for research is to determine how accessible constructs relate to the cognitive predictors of trait judgment accuracy outcomes, such as trait induction (De Kock, et al., 2015), defined as the ability to infer the traits that underlie behavior cues? For example, would an interviewer with low accessibility for neuroticism (i.e. neuroticism is not salient in their perceptual field) also struggle to detect and use appropriately the behavior cues related to neuroticism that an interviewee emits during an interview (i.e. low trait induction for neuroticism)?

Implications for Practice

Given our findings it would be too early to make recommendations for practice, pending further research. However, practitioners may consider using the chronic accessibility measure in interviewer training to highlight their accessible traits and potential blind spots for inaccessible traits when interpreting others' behavior.

5.7 Conclusion

This study explored the role of interviewers' personality and chronically accessible traits in the accuracy of their judgments of others' specific personality traits. Results from our trait-specific approach found that it does not necessarily "take one to know one" in personality judgments. That is, interviewers' own personality traits did not seem to affect their ability to effectively judge corresponding traits. Further,

[45] According to RAM, 'good judges' are moderators of accuracy in addition to 'good targets', 'good traits', and 'good information' (Funder, 2012).

interviewers' chronically accessible traits for selected traits (for example, extraversion and openness in the present study) may be useful predictors of trait-specific judgment accuracy. As such, they deserve a closer look in studies of the good judge of personality traits in personnel selection.

Appendix A

Research Evidence[a] on Individual Differences Constructs Predicting Judgment Accuracy in HRM

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Borman (1979)	146	Personality	Aggression	Small
Letzring (2008)	142	Personality	Big 5	Small-to-medium
Letzring (2008)	138	Personality	Big 5	Small-to-medium
Janovics (2003)	410	Personality	Big 5	Not avail. (dissertation abstract)
Gibson (2006)	*nr*	Personality	Big 5	Not avail. (dissertation abstract)
Christiansen et al. (2005)	122	Personality	Big 5	Small-to-medium
Powell (2008)	164	Personality	Big 5	Various (Small-to-medium)
Davis (1999)	82	Personality	Conscientiousness	Not avail. (dissertation abstract)
Borman (1979)	146	Personality	Detail orientation	Small-to-medium
Borman et al. (1991)	79	Personality	Detail orientation	Small
Letzring (2008)	138	Personality	Interpersonal problems	Small-to-medium
Lippa et al. (2000)	109	Personality	Masculinity/femininity	Small
Letzring (2008)	138	Personality	Narcissism	Small-to-medium
Davis (1999)	82	Personality	Need to evaluate	Not avail. (dissertation abstract)
Gibson (2006)	*nr*	Personality	Need to evaluate	Not avail. (dissertation abstract)
Borman et al. (1991)	79	Personality	Personal adjustment	Small
Lippa et al. (2000)	109	Personality	Personality (Big 5)	Various (Up to small-to-medium)
Vogt et al. (2003)	102	Personality	Psychological communion	Medium
Letzring (2008)	138	Personality	Psychological well-being	Various (Small-to-medium)
Human et al. (2011)	380	Personality	Psychological well-being and adjustment	Various
Borman (1979)	146	Personality	Self-control	Small-to-medium
Borman et al. (1991)	79	Personality	Self-control	No/negligible
Borman (1979)	146	Personality	Self-monitoring	No/negligible
Davis (1999)	82	Personality	Self-monitoring	Not avail. (dissertation abstract)

Author(s)	*N*	Cluster	Predictor	Effect size[b]
Borman (1979)	146	Personality	Sociability	No/negligible
Borman (1979)	146	Personality	Tolerance	Small
Borman (1979)	146	Personality	Various traits	Various (Up to Small-to-medium)
Letzring (2008)	138	Personality	Various traits	Small-to-medium
Hjelle (1969)	72	Personality	Various traits	Various (Small-to-large)
Ambady et al. (1995)	90	Personality	Various traits	Various (Small-to-medium)
Adams (1927)	80	Personality	Various traits	Questionable method
Borman (1979)	146	Personality	Various traits	Various (Neglible-to-Medium)

Note. k = 15 reported studies. The actual number of effects is much 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. [b]We used Cohen's (1988) guidelines to interpret effect sizes, i.e. no/trivial (.00), small (.10), medium (.30) and large (.50) effects. An effect-size interval of .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 is not reported for some studies because it was not available (typically when results were drawn from dissertation abstracts and the original dissertation could not be sourced).

Appendix B: Example Items from the Applicant Bio Rating Task

'Reading' applicants' personalities

Listed below are descriptions of five personality traits. Each description lists adjectives that describe people high and low on the trait. Please read each description carefully. You will use these descriptions in a subsequent rating activity.

	Behavior Description	
Trait	***High (+)***	***Low (-)***
1. Agreeable	Altruistic Humble Trust people	Sceptical Does not get involved with the problems of others
2. Conscientious	Strong willed Determined Well-organized	Procrastinate Unreliable Not very methodical
3. Extroversion	Likes people Active Warm	Reserved Independent Low need for thrills
4. Open to experience	Open to new experiences Curious Imaginative Appreciate art and beauty	Find change difficult Prefer to stick with the tried and true
5. Neurotic	Anxious Hostile Self-conscious	Calm Even-tempered Handle themselves well in stressful situations

Instructions

Next, we describe five interview applicants in terms of their personalities on each of the traits that were just described to you. Try your best to form an impression of each person's personality within the workplace context. Please indicate the level of personality trait exhibited by each person by selecting a number from 1 to 5 (1 = low indication of trait; 5 = strong indication of trait). You may refer to the personality descriptions listed earlier.

Person A

Person **A** is not really interested in others and shows little concern for others' problems. **A** also tends to insult people frequently. **A** doesn't particularly like structure and only sometimes does things according to plan. At work, **A** wouldn't necessarily be one to initiate conversations, but wouldn't bottle up feelings either. This person sometimes comes up with workable ideas for doing things better, although doesn't have a particularly good imagination. Person **A** is easily irritated and has frequent mood swings and often feels blue. **A** takes offence easily.

Please rate Person A on each trait by making a selection in the appropriate circle:

	1	**2**	**3**	**4**	**5**
1. Agreeableness					
2. Conscientiousness					
3. Extroversion					
4. Openness to Experience					
5. Neuroticism					

Hypothetical Personality Profile ('True Scores')

For interviewee **Person A**

(the information below was not visible to participants)

Person A – Neurotic Dominant	
Trait	Level (1-5)
Agreeableness	1
Conscientiousness	2
Extroversion	3
Neuroticism	**5**
Openness to Experience	3

Appendix C: Chronic Accessibility Measure (Higgins, 1982)

People that You Know

On this page describe five typical persons, using as many characteristics as you like to describe them. Use at least three word descriptions for each person.

First, describe a person that you **liked:**

…… …… ……

…… …… ……

…… …… ……

Next, describe a person that you **disliked:**

…… …… ……

…… …… ……

…… …… ……

Next, describe a person that you frequently **encountered:**

…… …… ……

…… …… ……

…… …… ……

A type of person that you **avoided:**

…… …… ……

…… …… ……

…… …… ……

A type of person that you **sought out** (in other words, you looked for their company):

…… …… ……

…… …… ……

…… …… ……

Appendix D: Development of Applicant Vignettes and True Scores

Vignette Development

To design and develop the applicant vignettes that served as stimulus materials, we relied on best practice recommendations for Experimental Vignette Methodology (EVM), which consists of "presenting participants with carefully constructed and realistic scenarios to assess dependent variables… thereby enhancing experimental realism and also allowing researchers to manipulate and control independent variables, thereby simultaneously enhancing both internal and external validity" (Aguinis & Bradley, 2014, p. 351). EVM was considered suitable as EVM "allows researchers to include factors that are relevant to the research question while excluding those that might confound the results. This amount of control helps to test causal hypotheses that would otherwise be difficult."(p. 357) Within EVM, we chose to use the 'paper people approach'. In addition to being the most often used EVM approach, this method was is appropriate when the goal is to assess explicit responses and outcomes (Aguinis et al., 2014), such as rating accuracy in our study.

Advantages of EVM. For our study purposes, there were various advantages of this method. First, a study that uses a fully-crossed design, where each participant views the same set of vignettes, can help the researcher to uncover the judgment processes and outcomes of an individual rater (Atzmüller & Steiner, 2010; Putka, Le, McCloy, & Diaz, 2008). Second, this approach allows for the presentation of the same stimuli to all participants and, therefore, helps control the potential confounding effect of irrelevant factors on judgments, such as interviewee race, gender, age, or attractiveness (Letzring, 2010; Sheppard, Goffin, Lewis, & Olson, 2011). To this end, we ensured that our target descriptions contained no reference to characteristics other than personality. For example, targets were referred to as 'Person A' and all descriptions contain neutral target descriptions related to personality only. Third, this method allowed us to control the level of immersion to maintain realism without introducing potential experimental 'noise' that accompanies video-stimuli. Finally, the use of text-based stimuli in accuracy studies has been associated with higher accuracy, compared to video- and live interpersonal interactions (Borkenau, et al., 2015).

Targets' Trait Profiles. We constructed a hypothetical personality profile for each target. For example, the trait profile for the applicant in the first vignette ('Person A') is described in Appendix B. In line with earlier studies (Byron, 2008) these profiles were developed with consideration of actual empirical correlations between personality traits (e.g., meta-analytic evidence of Big Five trait intercorrelations; Goldberg, et al., 2006) to ensure that the resulting profiles were realistic. An example of a profile for Target C is: agreeableness (4), conscientiousness (5), extraversion (3), neuroticism (or emotional stability) (2), and openness to experience (3). To ensure a balance of personality traits in the vignettes, each target was designed with a single dominant trait (e.g. Target A was neuroticism-dominant, or a '5' on a scale of 1-5).

Finally, the next step was to populate each vignette stimulus with trait-related cues. For each applicant, we developed a descriptive paragraph (of about ten sentences in length) containing behavior descriptors (key words and phrases) from a database of Big Five descriptor terms that we compiled from published empirical research (e.g., Goldberg, 1990, and others; Hofstee, et al., 1992; John, 1990; Saucier & Goldberg, 1996). For example, Person C[46] was described as 'At work, C is particularly detail-oriented and always strives for perfection. C loves order and regularity.' Also, 'However, C isn't necessarily comfortable amongst strangers and avoids excessive attention.' Where appropriate we inserted intensity descriptors (such as 'always', 'often', 'occasionally', 'sometimes', and 'hardly') to denote the extent to which the target exhibited a particular personality dimension.

Manipulation Check. It was important to establish whether the vignettes were adequate stimulus samples (Highhouse, 2009) of personality. In order to determine whether the hypothetical profiles were observable from the trait-cues that we inserted in each vignette we asked three professors in I-O psychology with knowledge of personality to rate them on each of the Big Five dimensions. Two-way, random effects intraclass correlations (ICC; Shrout & Fleiss, 1979) were conducted in order to ascertain the level of inter-rater reliability and agreement between subject matter experts' ratings of the set of vignettes. As the ICC adjusts for chance agreement and systematic differences between raters it is often preferable to other indices of inter-rater reliability (Hoyt, 2010). Analysis of the agreement between professors' ratings of the personality profiles showed acceptable (ICC = .75; Cicchetti, 1994) agreement that was higher than typical observer agreement for personality traits (see the following studies for reviews of observer agreement for personality trait judgments: Allik, Realo, Mõttus, Esko, et al., 2010; Allik, Realo, Mõttus, & Kuppens, 2010; Connelly & Ones, 2010; Mõttus, McCrae, Allik, & Realo, 2014). As such, the vignettes were considered highly 'judgeable'. To ensure that our applicant descriptions were realistic, we asked ten Masters students in IO-psychology to provide realism ratings on a scale of 1 'totally unrealistic' to 10 'completely realistic'. Across the five targets, the vignettes were rated as quite realistic (M = 8.8 out of 10, SD = 1.22). No realism ratings lower than 8 were received.

[46] Alphabetic designators (e.g., 'C') were used as a non-descript labels for each target as we were concerned that providing a name that provides gender or ethnicity cues may affect respondents' evaluations (Letzring, 2010).

Chapter 6

Summary and discussion

The current dissertation presents four studies investigating individual differences in the accuracy of judgments in subjective personnel selection measures. These individual differences are general mental ability, dispositional reasoning, Big Five personality traits, chronically accessible traits, and demographic factors. From the existing literature, little is known about why individual differences exist in judgment accuracy. In the study reported in the present dissertation, the characteristics of assessors were used as a vantage point to understand assessor accuracy. In this chapter, a concise summary of the main research findings is given, followed by a discussion to embed these findings in the literature. Finally, implications for practice and avenues for future research are carved out.

6.1 Summary of Main Findings

In the introductory chapter, our main research question as a central theme for this dissertation was formulated as:

> *To what extent do assessor constructs explain differences in their judgment accuracy in subjective rating measures used in personnel selection?*

In pursuit of answers to this question, four specific research questions (RQs) were raised that directed the studies presented in this dissertation. These related to the degree to which assessor constructs addressed in prior empirical literature (RQ1) and in the present study (RQ2-RQ4) are able to explain differences in assessors' accuracy in judgments in personnel selection ratings. The broad collection of assessor constructs that were investigated in the respective studies in this dissertation are highlighted in Figure 6.1.

Taken together, the research questions sought to determine whether there is empirical support for the theory-driven notion of the 'good judge' as a moderator of accuracy (see Figure 6.2), as predicted by the RAM (Funder, 1995, 1999, 2012). According to the RAM, it was proposed that so-called 'good judges' contribute to judgment accuracy (along with 'good targets', 'good traits', and 'good information'). Judges are required to detect and use behavioral cues effectively in order to produce accurate judgments. Using the research questions as signposts, the main findings of the five studies are outlined next.

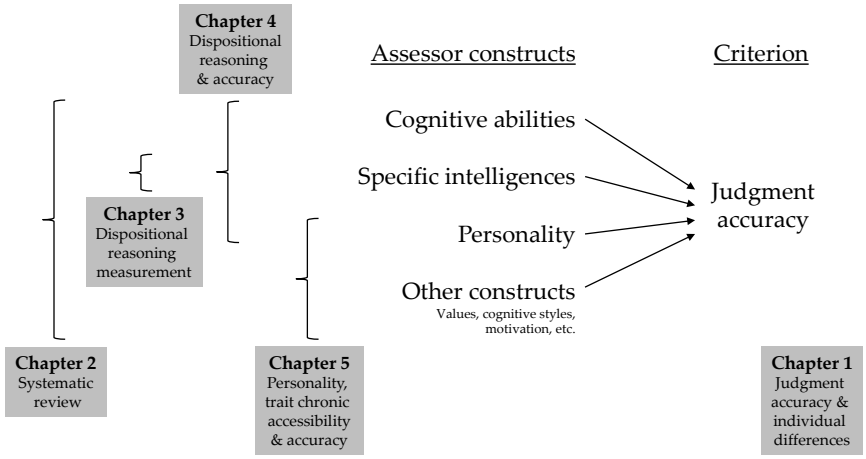

Figure 6.1. Assessor individual differences framework and visualising the linkages between studies in the present dissertation. Adapted from Farr, J. L., & Tippins, N. T. (Eds.). (2010). *Handbook of employee selection*. New York, NY: Routledge.

> Research Question 1: From a systematic review of the empirical HRM literature, which individual differences explain judgment accuracy in personnel selection, in other words, what is the profile of the good judge?

As a starting point for subsequent empirical research, it was necessary to conduct a systematic review of the extant HRM literature base on individual differences in judgment accuracy. In **Chapter 2**, prior individual difference research was weighed to establish what we know, and do not yet know, about the profile of the good judge. To this end, empirical studies over more than 60 years were reviewed to identify, summarize, analyse and evaluate evidence in support of various individual differences thought to explain accuracy. In this review, the overall aim was to construct a profile of the 'good judge' and chart directions for future research on individual differences that may explain judgment accuracy.

Our review showed that more is known about the 'good judge' than earlier thought (Funder, 1999; Guion, 1999). The review shows 126 individual effects reported in 48 works (published articles and unpublished dissertations and theses). The studies that met our inclusion criteria were conducted between 1953 and 2011. Of these, the majority were college studies (79.3%) with only a small proportion in field samples (14.9%) or mixed groups (5.8%). The mean sample size for the studies reviewed was approximately 166 participants (SD = 116; min = 44; max = 898). Apparently, the majority of studies were conducted in North America, although many studies did not reveal the location of the research.

Our review results are summarized in Figure 6.2. Overall, empirical evidence suggests that cognitive factors appear to play a dominant and consistent role in judgment accuracy. For example, it appears that accurate assessors in HRM studies are not only more intelligent (i.e. they have higher general mental ability) but a few recent studies show they may also have higher dispositional reasoning. Dispositional reasoning is defined by Christiansen et al. [2005] as the complex understanding of others' behaviors, traits, and situations' potential to elicit traits into manifesting themselves. The effect sizes (uncorrected) for cognitive ability and dispositional reasoning predictors of judgment accuracy are in the moderate-to-large (Cohen, 1988) range. These findings support the view that cognitive ability enables cue utilization in the rating task encountered in personnel selection devices such as interviews. In contrast, the empirical research base suggests that accurate assessors do not necessarily share a prototypical personality type, that is, personality traits tended to be poor predictors ($0 < r < .20$; small effect size, Cohen, 1988) and inconsistent predictors of judgment accuracy. Our review showed that none of the broad Big Five traits were reliable predictors of accuracy outcomes in HRM studies.

However, the review highlighted assessor constructs that hold potential for advancing understanding of individual differences in judgment accuracy. These constructs were addressed in subsequent empirical research studies – focal

constructs that were studied included dispositional reasoning (RQ2 and RQ3), as well as personality and chronically accessible personality traits (RQ4).

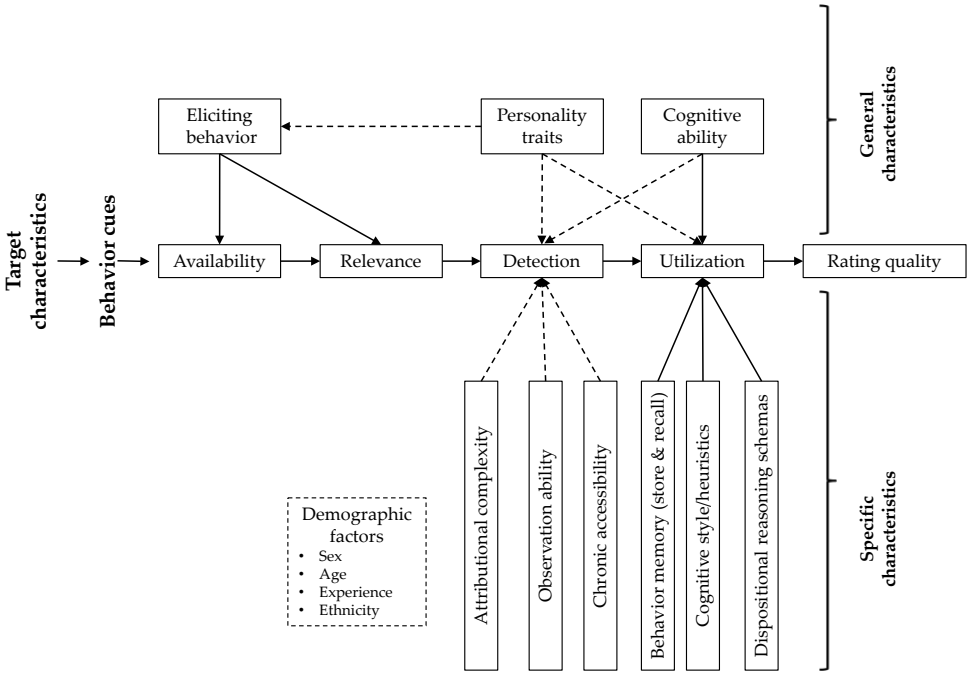

Figure 6.2. A 'Good Judge' Model of Individual Differences in Judgment Accuracy. Solid arrows indicate relationships with empirical research support, whereas dotted arrows indicate relationships with limited/inconsistent research support.

Research Question 2: How can dispositional reasoning be measured reliably and with measurement validity at the component level?

In the aforementioned review chapter, it was determined that dispositional reasoning (Christiansen et al., 2005) may hold potential as a useful predictor of judgment accuracy in subjective rating methods. Dispositional reasoning was earlier defined as complex knowledge and understanding of traits, behaviors, and the potential of situations to manifest traits into behaviors (Christiansen et al., 2005). According to the dispositional reasoning framework (see Figure 6.3), interviewer judgment accuracy may depend on three components, namely *trait induction* (the ability to know how traits manifest themselves in behavior), *trait extrapolation* (an understanding of how traits and their behavioral manifestations naturally co-vary), and *trait contextualization* (the ability to identify situations that are relevant to different traits).

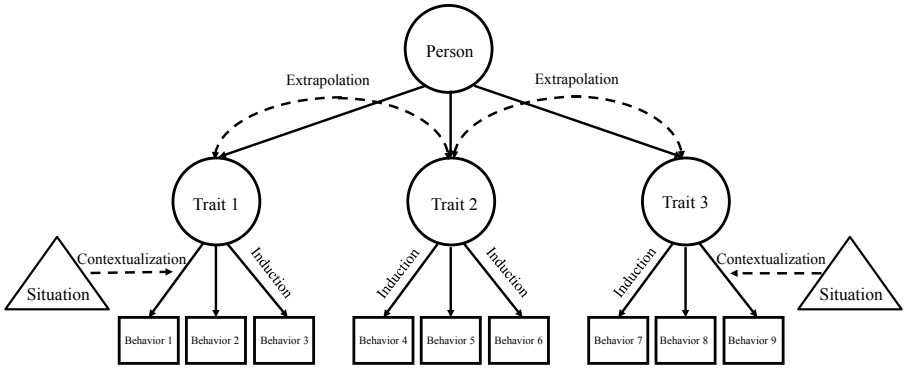

Figure 6.3. Understanding the components of dispositional reasoning: trait induction, trait extrapolation and trait contextualization.

Although dispositional reasoning was initially conceptualized (Christiansen, et al. 2005) as a broad set of three theoretically distinguishable components (induction, extrapolation, and contextualization), insight into the suggested factor structure of dispositional reasoning has been hampered by a lack of reliable subscale scores to measure the components. As a result, earlier studies (e.g. Christiansen et al., 2005; Powell & Goffin, 2009) have not been able to test the internal validity of the component-view of dispositional reasoning. By implication, it was not possible to further investigate the individual components in new studies.

In **Chapter 3**, the internal componential structure of dispositional reasoning was discussed. We addressed the need for a component-level measure of dispositional reasoning by developing[47] the revised interpersonal judgment inventory (RIJI) (De Kock et al., 2015) to yield reliable component scores. The component measures were then administered to two samples drawn from different populations of assessors (managers, $N = 160$; psychology students, $N = 161$) in order to test competing models of underlying factor structure of dispositional reasoning scores. These groups represented parts of the pool of assessors who typically receive rater training in HRM (Krause & Thornton, 2009). After data collection, we compared a general factor model (M1), a three-component model (M2), as well as a second-order model (M3), which combines three components (at level 1) with a higher-order general dispositional reasoning factor (at level 2). These analyses were conducted to seek evidence of internal measurement validity of the revised measure. Furthermore, as the measure was administered to different assessor types, it was

[47] The development of the measure is described in detail in Chapter 3. For logical reasons, the factor analytic study (Chapter 2) was conducted before the external validity study reported in Chapter 3.

possible to compare the factor structures underlying the dispositional reasoning measure in these groups.

Results showed that the RIJI (De Kock et al., 2015) showed reliable and measurement valid scores in both samples that were tested. The confirmatory factor analysis supported the three-component theoretical structure proposed by Christiansen et al. (2005), but this model was not invariant between the manager and psychology student groups. Moreover, a hierarchical model of dispositional reasoning (with a higher-order general latent factor, influencing three lower-order specific components) showed good fit in the combined sample, did not fit significantly worse than a three-component model, showed acceptable fit in both manager and student samples, but showed only configural invariance between these groups. Together, these results show acceptable measurement properties of the dispositional reasoning measure in the populations that we sampled from, namely managers and psychology students.

> Research Question 3: Does dispositional reasoning meet the classic criteria for an intelligence measure, considering the relationship between interviewers' dispositional reasoning components, general mental ability, personality, and their judgment accuracy for rating interview dimensions?

The systematic review of empirical studies reported in Chapter 2 identified dispositional reasoning as a promising individual difference construct that might aid further attempts to explain judgment accuracy in personnel selection. In **Chapter 4**, we took an in-depth look at dispositional reasoning as a predictor of interview judgment accuracy. The present study departed from earlier dispositional reasoning research in two ways. First, it took a componential view of dispositional reasoning, that is, it explored the role of the subcomponents of trait induction, trait extrapolation and trait contextualization in predicting accuracy. To this end, we extensively revised an earlier measure of dispositional reasoning, the interpersonal judgment inventory (Christiansen, et al., 2005) to yield reliable subscale scores. The measure was administered, along with other individual difference measures, including general mental ability and personality traits, to a field sample (N = 146) of mid-level managers in a police services organization. At the same time, respondents completed measures of general mental ability and personality, as well as a judgment accuracy task consisting of rating high-structure videotaped interviews of eight applicants.

By studying dispositional reasoning at the component level, it was possible to address two issues. First, we wanted to determine the nature of dispositional reasoning as a construct. The question that was raised and tested here was, "Is there enough evidence to support the view of dispositional reasoning as an intelligence, or not?" To address this question, we tested whether dispositional reasoning meets several conceptual and empirical criteria (Carroll, 1993; Flanagan, et al., 1997; Mayer,

et al., 1999). A corollary of this series of tests was that it also yielded evidence of construct validity issues by indicating the nomological placement of dispositional reasoning and its components (induction, extrapolation and contextualization) within an assessor individual differences framework.

Results of this study suggested that the dispositional reasoning components generally correspond to the characteristics of an intelligence measure as they broadly adhered to the conceptual and correlation criteria that we tested. First, the components converged with one another and with general mental ability, reflecting evidence of positive manifold (Horn & Cattell, 1966). Second, the components also predicted ($.14 < r < .33$; uncorrected) our accuracy criterion, namely interview judgment accuracy. In addition, the components showed incremental validity (with small effect) (Cohen, 1988) to general mental ability in predicting accuracy, and offered evidence of discriminant validity with personality constructs. In short, these results provided evidence for a nomological network with dispositional reasoning positioned as a form of intelligence. That is, our results provide early support that dispositional reasoning may be a specific mental ability that good judges employ to process behavioral information in selection interviews. In the study reported here, managers who were better at extrapolation and contextualization specifically were more accurate judges of interview dimensions than those with inaccurate implicit personality theories (low extrapolation) and poor understanding of trait activation (low contextualization). The induction component seemed to be a less important contributor to interview judgment accuracy, suggesting that knowledge of which traits are indicated be particular behaviors may not play a major role in judgment accuracy in the high-structure interviews that we used. Further replication of these findings is necessary, however.

> Research question 4: Would judges' personality traits and their chronically accessible traits predict trait-specific personality judgment accuracy?

Our final research question departed from the cognitive-constructs theme of Chapters 3 and 4. In **Chapter 5**, attention shifted to assessors' personality-related characteristics as predictors of their trait judgment accuracy. Our earlier review of individual difference predictors of accuracy (see Chapter 2) concluded that personality traits compared relatively poorly to cognitive factors in how well they explain variance in accuracy measures, despite their potential theoretical relevance to judgment tasks. We argued that poor support that personality traits have received as predictors of personality judgment accuracy may have resided in the fact that they generally used trait-generic accuracy criteria, with poor conceptual alignment between the judge's respective personality traits and an accuracy criterion that is generic across traits. At the same time, emerging research (Allik, Realo, Mõttus, & Kuppens, 2010) indicates that traits differ in terms of how accurately they can be judged. Taken together, these arguments build the case for the use of trait-specific accuracy criterion measures. To address this section of the literature, Chapter 5

adopted a trait-level approach to accuracy: we calculated accuracy criterion measures separately for each of the Big Five personality traits that research participants had to judge. As emerging research findings suggest that the *self* may be an important basis from which judges form impressions of others (Alicke, et al., 2005), we hypothesized that trait expertise for judging specific traits may emanate from the interviewer's own traits, that is, by holding high levels of the trait being judged. Stated differently, perhaps assessors are better at judging traits they are elevated on, which would be the case when 'it takes one to know one'. The RAM (Funder, 1999) supports this idea, as cue detection and cue utilization may be easier for traits with which interviewers are more familiar if these traits are part of their own personality profile. Drawing on construct accessibility theory (Higgins, 2012), we hypothesized that trait expertise may arise from the heightened salience of interviewers' own traits in their perceptual filters, that is, through their chronically accessible traits. For instance, a person with conscientiousness as an accessible trait would tend to describe others more in conscientiousness-related descriptors than for other traits. To address this research question, a self-report inventory measure of personality was administered along with a personality judgment task in two separate studies (Study 1: students, $N = 183$; Study 2: managers, $N = 223$). In addition, respondents in our field sample (Study 2) completed an open-ended free-response measure of construct accessibility (Higgins, 1982), from which we could derive their accessible Big Five traits. To our knowledge, no earlier studies have investigated the relationship between interviewers' chronically accessible traits, their own traits, and trait-level rating accuracy.

The results of the final empirical study revealed that being elevated on a trait (for example, 'being a very outgoing person', indicative of extraversion) (Goldberg, 1992) may not necessarily make one more adept at judging that trait. For example, extraverts in our studies were not better than introverts in detecting and using extroverted-related behavior cues in the target stimuli we presented to our participants.

Furthermore, our results replicated those of earlier studies (e.g., Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003) by finding that interviewers' personality traits did not predict their personality trait judgment accuracy. Even when using trait-level accuracy measures in our study, we did not find an empirical link between assessor personality traits and accuracy. However, interviewers who were more likely to describe others in terms of openness to experience and extraversion were also more accurate at judging these traits in others, lending partial support to our hypothesis that trait chronic accessibility may predict trait judgment accuracy for certain traits.

6.2 Discussion, Practical Implications, and Future Research

The main purpose of the present study was to extend the empirical literature about potential individual differences explaining assessor judgment accuracy in personnel selection. As subjective rating methods that utilize assessors (e.g. interviews or ACs) remain at the core of modern personnel selection programmes (Dipboye, et al., 2012; Ryan, et al., 1999) a closer look at assessor factors that explain rating accuracy outcomes was warranted. At the same time, calls for a better understanding of how assessor constructs (Jones & Born, 2008; Landy & Farr, 1980; Lievens, et al., 2009; Nathan & Alexander, 1985) may affect rating quality have been largely left unanswered. As such, the field of HRM still knows relatively little about what makes a good judge. If individual differences that consistently predict accuracy could be found, it would support the practice of screening assessors on measures of these constructs, or to use them in assessor training and development interventions. In the following sections, the main research findings will be synthesized and embedded in the extant literature. Furthermore, practical implications and directions for future research will be carved out.

Cognitive Factors

In three of the four research studies reported on in this dissertation, cognitive predictors of judgment accuracy in personnel selection received considerable attention. Overall, empirical evidence suggested that cognitive factors play a dominant and consistent role in judgment accuracy. For example, the good judge in our research was intelligent (i.e. high cognitive ability) and also had well-developed understanding of others' behaviors, traits, and situational activation of traits (i.e. higher dispositional reasoning) (Christiansen et al., 2005). Both these constructs showed consistent small to medium effects (uncorrected $.14 < r < .33$) (Cohen, 1988) with our accuracy measures. It must be pointed out that the effect size for overall dispositional reasoning as a predictor of interviewer judgment accuracy in our study (uncorrected $r = .34$; moderate) was not as large as the effect size ($r = .42$) reported in Christiansen et al. (2005). Nevertheless, our results suggest that better social information processing allow accurate assessors to process, store and recall targets' behavior better than poor assessors are able to. Our systematic review of the individual differences literature shows that effect sizes for cognitive ability factors overall are moderate (and sometimes large) and these appear to be relatively consistent. The important role of cognitive factors in judgment accuracy is logical: the accuracy of interpersonal judgments in personnel selection devices relies heavily on cognitive processes (DeNisi, et al., 1984; Lance, et al., 2004) and the individual differences that drive them (Jones & Born, 2008).

In Chapters 3 and 4, two studies were reported with dispositional reasoning as focal construct. In the first (**Chapter 3),** we were able to explore the internal construct validity of our revised measure of dispositional reasoning (RIJI) (De Kock et al.,

2015) by assessing its factor structure. In the second investigation, we built on these findings by conducting empirical tests of dispositional reasoning against classic criteria for an intelligence measure (**Chapter 4**). Results showed that together, trait induction, trait extrapolation and trait contextualization are more than likely to be specific intelligence measures that reside in the social-cognitive domain. Together, these components also predicted rating accuracy when controlling for general mental ability. Dispositional reasoning components were also relatively uncontaminated with personality as our analyses revealed discriminant validity of the components with personality measures.

In summary, our results suggest that dispositional reasoning may be a useful predictor of judgment accuracy of interview dimensions in the field. In doing so our findings replicate those of earlier investigations (Christiansen, et al., 2005; Powell & Goffin, 2009) showing the utility of dispositional reasoning in explaining rating accuracy outcomes. When embedding the results of our study in the growing literature on dispositional reasoning, our component-approach may add a new angle to the study of this construct as a predictor of judgment accuracy in personnel selection ratings.

The conclusion that dispositional reasoning resembles the characteristics of an intelligence construct is underlined by our confirmatory factor analysis results (**Chapter 3**) – the factor structure underlying dispositional reasoning scores from our study appears well-represented as a hierarchical model, that is, one with an underlying general factor influencing the components. In doing so, the internal composition of dispositional reasoning observed in the results of our analysis follows in the footsteps of intelligence constructs (Carroll, 1993, 2003) that are often constituted in different strata. Taken together, our two studies on dispositional reasoning built a case for it as nomologically placed squarely in the cognitive ability domain. As a result of its stability as predictor of accuracy outcomes, the prominence of dispositional reasoning and its components in individual differences research can only grow further.

Regarding the role of specific components of dispositional reasoning, our results reported in **Chapter 4** point out that interview dimension judgment accuracy may depend more upon understanding how traits co-vary (extrapolation) and how situations affect trait expression (contextualization), than on knowing which traits are signalled by behavioral cues (induction). For example, the relative weights analysis confirmed that extrapolation and contextualization exerted the strongest influence in predicting accuracy. Practically, this means that interviewers who are better able to fill in missing information about an interviewee by extrapolating information about the interviewee's other traits (using their implicit personality theories) were also more accurate judges of interview dimensions. Although we are not suggesting that interviewers should deviate from strict reliance on observable behavior to infer the characteristics of the interviewee (Gatewood & Field, 2011), our

results suggest that more developed implicit personality theories may assist the interviewer to develop a cohesive mental picture of the interviewee.

Personality Factors

Neither our review of the extant empirical literature nor our empirical studies reported on in this dissertation could find support for the view that personality traits are important in interpersonal judgment. There is, however, room for more work to explore more complex hypotheses. Personality may be more complexly related to accuracy, for example as a moderator of the effect of cognitive variables on judgment accuracy (e.g. Christiansen et al., 2005). Or, personality may influence accuracy indirectly through its effect on the assessors' behavior during the interview. For example, Letzring (2008) reported that accurate judges used certain behaviors related to cue elicitation. It is likely that these behaviors are a function of the judges' personality, but why traits themselves do not seem to predict accuracy, is a question that remains unanswered.

In our final study (**Chapter 5**), we introduced chronically accessible personality traits as potential predictors of trait-level accuracy. We expected that interviewers would be more accurate at reading traits in others if they had elevated levels on these traits themselves. Stated otherwise, we wanted to determine whether it truly 'takes one to know one' in personality judgments. As chronically accessible traits represent more salient behavior information in the judges' perceptual schema (Higgins, 2012), we reasoned that interviewers would be more adept at detecting and using behavior cues that represent these traits. In turn, the enhanced cue detection and utilization this affords should promote accuracy (Funder, 2012).

Our hypothesized predictive, incremental and mediating role for chronic accessibility (in conjunction with personality) in judgment accuracy received only partial support. Openness to experience and extraversion accessibility predicted trait judgment accuracy for the same traits. But this finding did not extend to the remaining Big Five traits – accessibility for agreeableness, conscientiousness, and neuroticism did not predict accuracy for corresponding traits. As such, trait construct accessibility may contribute to interpreting cues for the same trait, but much more research is needed before a firm conclusion can be reached. Further, accessibility for openness, emotional stability and agreeableness appeared to spill over from the interviewers' own corresponding personality traits: for example, interviewers high on agreeableness also tended to describe others in agreeableness-related terms. Taken together, our finding lends limited additional support for the view that the self may influence how we judge others (Dunning & McElwee, 1995; Motowidlo, et al., 2006).

At the broadest level, the research findings presented in this dissertation are in line with theories of judgment accuracy (Funder, 1995, 2012) that suggest that assessors are an integral link in the chain of factors that lead to eventual accuracy.

We have synthesized the prior literature into a comprehensive 'good judge'-model (see Figure 6.2) of individual differences in judgment accuracy. Further, our model outlines clusters of assessor constructs linked to specific judgment processes (cue detection and cue utilization) important for accurate judgment.

Recommendations for Practice

There are in our view three promising ways to advance practices used for rater training and selection, given the findings of the present study. The first is the identification of judges on the basis of constructs that can predict accuracy. In contrast to earlier scepticism that empirical evidence would be found of individual differences that predict accuracy (Funder, 1999; Guion, 2011), it appears there are indeed assessor constructs that can consistently predict accuracy. These appear to lie mostly in the cognitive domain (e.g., Christiansen et al., 2005; De Kock et al., 2015). Our results suggest that organizations may consider using cognitive ability and dispositional reasoning measures to select raters, as they may predict judgment accuracy in personnel selection and could therefore be considered 'job-relevant'. In addition, dispositional reasoning component scores are job-related and may be useful to screen interviewers.

A second potential way to apply our findings to advance rating practices is to consider ways to develop assessor constructs that predict accuracy. But first we need to know whether these constructs can be developed. The implicit question is whether or not a good judge is born, or made. Dispositional reasoning, for example, broadly adheres to the criteria for an intelligence measure. Early attempts to enhance one of its components (induction) experimentally have been unsuccessful in laboratory conditions (Powell & Goffin, 2009). So, before we can recommend that dispositional reasoning training be used to enhance rating quality, we need empirical evidence to support its efficacy. Another potentially useful training approach may lie in interventions that focus on each component of dispositional reasoning (Christiansen et al., 2005). Currently, most rater training is based on frame of reference training (Roch, et al., 2012; Woehr & Huffcutt, 1994) – this type of training does not typically entail the contextualization and extrapolation components, but generally only the induction component.

Finally, given that the effects we observed for the predictive and incremental validity of judges' personality and chronically accessible traits were generally small and somewhat divided, it would be too early to make recommendations for practice, pending further research corroboration. At the very least, our results provide early indications that it may not 'take one to know one' as a judge of personality in interviews, that is, in our study interviewers were just as effective to infer the personality traits of targets that did not share their own (interviewers') traits. However, we suggest that practitioners consider using the chronic accessibility measure presented here to make interviewers aware of their accessible traits and potential blind spots for inaccessible traits when interpreting others' behavior.

Future Research Directions

Our series of studies have brought a number of potential future research directions to the fore. The respective studies reported on in this dissertation have generated a comprehensive list of research questions.

A few of the more promising questions are outlined next. For example, there are some questions about dispositional reasoning to be addressed in future research work. Our research could not test all the criteria for dispositional reasoning as an intelligence measure and, therefore, studies are needed to determine whether dispositional reasoning increases with age (Mayer, et al., 1999). Furthermore, the nomological network surrounding dispositional reasoning can be explored further. Consider the conceptually related constructs of emotional intelligence and social intelligence (Lievens & Chan, 2010). We call for more studies that would test the relationships between dispositional reasoning and social and emotional intelligence. These studies could also try to disentangle the relative importance of these constructs in predicting accuracy.

Questions also remain about the role of the rating context in assessors' reliance on individual difference constructs. For example, would the components of dispositional reasoning relate differentially to accuracy criteria if the judgment context were altered? In one of our studies, a high-structure interview was used. When standardized rating materials, instructions and criteria are provided (typical in high-structure interviews) judges are encouraged to use normative theories (relating to job-related dimensions) to interpret behavior, rather than personal, idiosyncratic theories (Uggerslev & Sulsky, 2008), implying that implicit personality theories (i.e. extrapolation) would be irrelevant in high-structure rating tasks. Also, when standardized rating materials (e.g. behavioral checklists) specify which interview dimensions are implied by certain behaviors, the role of the induction component could be diminished. In summary, aspects of the rating context may potentially act as boundary conditions for the relative importance of dispositional reasoning components in judgment accuracy. For example, dispositional reasoning may predict judgment accuracy stronger in rating tasks that are highly ambiguous and more complex (DeNisi, et al., 1984; Lance, et al., 2004), than in rating tasks that are simple and relatively unambiguous, because in the former, assessors need to rely heavily on constructs that may facilitate cue detection and cue interpretation. We call for more research that explores how situations may moderate which components of dispositional reasoning influence judgment accuracy.

Accurate interviewers in one of the studies reported on in this dissertation were also able to understand how traits are activated by situations, in line with trait-activation theory (Tett & Guterman, 2000). In doing so, results add to the growing literature base (e.g., Jansen et al., 2013; Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015) that shows the importance of judging situations in personnel selection. In fact, situation assessment may be inextricably linked to dispositional inferences, as

assessors may do both at the same time (Krull & Erickson, 1995). Along these lines, it would make sense to investigate assessors' ability to read and infer the characteristics of situations as potential predictors of judgment accuracy.

Finally, we have argued that the prominence of cognitive factors as predictors of judgment accuracy may stem from the rationale that cognitive factors allow assessors to deal with the inherent complexity of the judgment task in HRM ratings (DeNisi, et al., 1984; Lance, et al., 2004). As such, there is a strong conceptual overlap between the predictor and criterion space (Schmitt & Chan, 1998). Future studies that share our interest in uncovering what makes a good judge in subjective ratings in personnel selection may benefit from taking a cue from these findings. If we are to have any success, we need to stay close to understanding the rating task and judgment processes that facilitate accuracy.

6.3 Conclusion

In closing, the research reported here sought to add to the literature on what makes a good judge in HRM. A review of the empirical literature shows that our understanding of assessor constructs that explain judgment accuracy has come a long way. Assessors that produce accurate ratings of others may share a number of characteristics, and cognitive factors are likely to continue playing the principal role in individual difference research investigating judgment accuracy. The studies reported on in this dissertation shed light on the role and nature of dispositional reasoning in interviewer judgment accuracy. In addition, we presented evidence in support of the view that dispositional reasoning may form part of the cognitive ability domain of constructs. Moreover, results from two of our studies showed that dispositional reasoning has distinguishable subcomponents, including induction, extrapolation and contextualization. Assessor constructs related to personality were also investigated in our systematic review of earlier research, as well as in two of our empirical studies. Our introduction of trait-specific accuracy criteria and chronic trait accessibility as a predictor did not yield promising results. On the whole, our findings corroborated the research base by indicating that personality-related factors may play a supporting role in the assessor's repertoire of constructs utilized in the judgment task. Although we are coming closer to understanding the assessor constructs that shape judgment accuracy, there are aspects of assessor individual differences that have hardly been touched on. Together, studies of these constructs hold potential to further enhance the quality of assessment ratings in HRM.

Nederlandse samenvatting

De personeelsselectie leunt gewoonlijk zwaar op assessoren als beoordelaars van sollicitant-kenmerken. Assessoren zijn bijvoorbeeld interviewers, beoordelaars tijdens assessment center opdrachten, en managers die prestaties van sollicitanten tijdens arbeidsproeven evalueren. Organisaties gebruiken assessoren veelal om een oordeel over de prestaties van sollicitanten tijdens selectieprocedures te verkrijgen. Deze beoordelingen zijn belangrijk voor besluiten in personeelsselectie met betrekking tot het aanbieden van ander werk of promoties. Gezien hun belangrijke rol in de beoordeling van huidig en toekomstig personeel, is het opmerkelijk dat er over de rol van assessoren in personeelsselectie tot op heden nog niet veel duidelijkheid bestaat (zie ook Guion & Gibson, 1988; Hough & Oswald, 2000; Sackett & Lievens, 2008). Hoewel er wel bekend is dat assessoren onderling verschillen in de accuraatheid van hun beoordelingen (Borman, 1977; Van Iddekinge, Sager, Burnfield & Heffner, 2006) weten we nog maar weinig over de redenen voor deze verschillen.

Dit proefschrift bestudeert de assessor. Meer specifiek wordt nagegaan of bepaalde kenmerken van assessoren de variatie in accuraatheid tussen assessoren kunnen verklaren. Anders gezegd, in welke mate kunnen kenmerken van assessoren de verschillen in accuraatheid bij hun beoordelingen tijdens personeelsselectie verklaren?

Voor dit proefschrift is hiertoe gebruik gemaakt van de literatuur over sociale cognitie en van modellen over de accuraatheid van beoordelingen (bijv. Funder, 1999). In vier studies die hier beschreven worden is uitgegaan van het zogeheten 'Realistic Accuracy Model' (RAM; Funder, 1955, 1999, 2012). Dit model stelt dat de zogenaamde 'goede beoordelaar' (*the good judge*) bijdraagt aan een nauwkeurige beoordeling. Een 'goede beoordelaar' herkent bepaalde gedragingen van sollicitanten en gebruikt deze om tot nauwkeurige beoordelingen te komen. De veronderstelling van de in dit proefschrift gepresenteerde onderzoeken is dat bepaalde kenmerken van assessoren deze beoordelingsprocessen kunnen verbeteren.

Een geïntegreerd profiel van de 'goede beoordelaar' dat is gebaseerd op wetenschappelijk onderzoek kan bovendien bijdragen aan het verbeteren van de selectie en opleiding van assessoren. Om op die behoefte in te gaan, worden in dit proefschrift vier studies beschreven. Op basis van eerdere literatuur zijn er verscheidene kenmerken van assessoren geïdentificeerd als mogelijke voorspellers voor de accuraatheid van door hen gegeven beoordelingen van sollicitanten. Deze kenmerken zijn onderzocht in drie empirische studies. Het gaat om de volgende kenmerken: algemene intelligentie, dispositioneel redeneren, de Big Five-persoonlijkheidskenmerken, chronisch toegankelijke persoonlijkheidskenmerken en

demografische factoren. Deze kenmerken zullen hieronder in de bespreking van de desbetreffende studie worden uiteengezet.

Overzicht van Empirische Bevindingen

Voorafgaand aan en als uitgangspunt voor de drie empirische studies is eerst de literatuur met betrekking tot individuele verschillen in de accuraatheid van de beoordeling van anderen systematisch bestudeerd (**Hoofdstuk 2**). Eerdere onderzoeken zijn daartoe beoordeeld om na te gaan wat we weten en wat we nog niet weten over het profiel van de 'goede beoordelaar'. Voor dit doel werden empirische studies over een periode van meer dan 60 jaar (van 1953 tot 2015) geanalyseerd en geëvalueerd om zo verscheidene individuele verschillen te identificeren die de accuraatheid van ander-beoordelingen zouden kunnen verklaren. Het doel van deze review was om een profiel van de 'goede beoordelaar' te construeren en aanwijzingen in kaart te brengen voor toekomstig onderzoek over individuele verschillen die de accuraatheid in ander-beoordelingen zouden kunnen verklaren.

De review toonde aan dat er meer bekend is over de 'goede beoordelaar' dan eerder werd gedacht (zie Funder, 1999; Guion, 1999). De review omvat 126 individuele effecten uit 48 gepubliceerde artikelen en ongepubliceerde proefschriften en theses. Uit de review bleek dat cognitieve factoren een dominante en systematische rol spelen in de accuraatheid van beoordelingen. Zo lieten oudere studies zien dat accurate assessoren niet alleen intelligenter zijn (dat wil zeggen dat ze een groter verstandelijk vermogen hebben) maar recente studies (Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005; Powell & Goffin, 2009) toonden ook aan dat zij waarschijnlijk over een groter vermogen tot dispositioneel redeneren beschikken. Dispositioneel redeneren wordt gedefinieerd als het begrijpen van persoonlijkheidseigenschappen van anderen door het verband tussen deze eigenschappen en gedrag in te zien, de onderlinge verbanden tussen persoonlijkheidskenmerken te begrijpen, en door te hebben in hoeverre specifieke situaties relevant zijn voor het kunnen beoordelen van persoonlijkheidskenmerken (Christiansen et al., 2005). Algemene intelligentie en dispositioneel redeneren bleken behoorlijk goede tot sterke voorspellers te zijn van accuraatheid in de beoordeling van anderen (Cohen, 1988). In contrast hiermee liet de review zien dat accurate assessoren niet noodzakelijkerwijs een specifieke persoonlijkheid hebben. Persoonlijkheidskenmerken van assessoren bleken relatief zwakke ($0 < r < .20$; een klein effect, Cohen, 1988) en inconsequente voorspellers te zijn van de accuraatheid van ander-beoordelingen. De review toonde aan dat geen van de Big-Five persoonlijkheidskenmerken betrouwbare voorspellers waren van accuraatheid.

Hoofdstukken 3, 4 en 5 rapporteren empirisch onderzoek naar specifieke individuele verschilkenmerken die naar voren kwamen uit de review. Alle studies zijn gedaan bij Zuid-Afrikaanse organisaties.

In hoofdstukken 3 en 4 stond dispositioneel redeneren centraal (Christiansen et al., 2005). Volgens het framework van dispositioneel redeneren kan de accuraatheid van de beoordeling door de interviewer afhangen van drie factoren, namelijk *inductie uit persoonlijkheidskenmerken* (het vermogen om te weten hoe persoonlijkheidskenmerken zich in gedrag tonen), *extrapolatie uit persoonlijkheidskenmerken* (het begrijpen hoe persoonlijkheidskenmerken en de gedragsuitingen ervan kunnen co-variëren), en *persoonlijkheidskenmerk contextualisatie* (het vermogen om situaties die relevant zijn voor verschillende persoonlijkheidskenmerken te identificeren).

Ter illustratie: Iemand die een hoog vermogen tot persoonlijkheidskenmerk-inductie heeft, weet bijvoorbeeld dat een kennis die veel praat (i.e., waargenomen gedrag) hoogstwaarschijnlijk ook extravert zal zijn (i.e., onderliggende persoonlijkheidskenmerk; Goldberg, 1992). Een persoon met een hoog extrapolatievermogen is zich bewust van het feit dat mensen die eerlijk zijn, over het algemeen ook betrouwbaar zijn (Goldberg, 1992). Beoordelaars met hoge niveaus van persoonlijkheidskenmerk-contextualisatie begrijpen de mate waarin situaties gunstig kunnen zijn voor het tot uitdrukking brengen van een persoonlijkheidskenmerk. Persoonlijkheidskenmerk-contextualisatie stelt een beoordelaar bijvoorbeeld in staat om in te zien dat extraversie zich eerder manifesteert in een situatie waarin de persoon omringd is door anderen, dan in een situatie waarin de persoon alleen is.

Hoewel het concept dispositioneel redeneren aanvankelijk opgevat werd als opgebouwd uit deze drie theoretisch te onderscheiden componenten, werd het inzicht in de voorgestelde factorstructuur van dispositioneel redeneren belemmerd door een gebrek aan betrouwbare subschaalscores om de componenten te kunnen meten. Als gevolg hiervan was het in eerdere studies niet mogelijk om de verschillende componenten te testen en de individuele componenten empirisch verder te onderzoeken.

Vanwege het ontbreken van een meetinstrument voor het betrouwbaar kunnen meten van componentscores van dispositioneel redeneren is de 'Revised Interpersonal Judgment Inventory' (RIJI) ontwikkeld (voor bijzonderheden over de herziening van het originele meetinstrument, zie De Kock, Lievens, & Born, 2015). **Hoofdstuk 3** behandelt de interne meeteigenschappen van de RIJI. De componentmetingen werden gedaan bij twee steekproeven uit twee verschillende groepen assessoren (managers, $N = 160$; psychologiestudenten, $N = 161$) met het doel om verschillende alternatieve modellen voor de onderliggende factorstructuur van dispositioneel redeneren te toetsen. Een algemeen factor-model (M1), een drie-componenten-model (M2), en een tweede-orde-model (M3) werden met elkaar vergeleken. Het laatste model combineert de drie componenten met een algemene hoger-order factor van dispositioneel redeneren. Doordat het meetinstrument bij verschillende assessoren typen (managers en psychologiestudenten) was afgenomen,

kon de aan dispositioneel redeneren onderliggende factorstructuur tussen beide groepen worden vergelijken.

De resultaten toonden aan dat in beide geteste groepen de RIJI een betrouwbaar en valide meetinstrument was. De factoranalyse onderschreef de drie-componenten-theoretische structuur die door Christiansen et al. (2005) was voorgesteld, maar het model was niet metrisch invariant tussen de groep managers en de groep psychologiestudenten. Het niet kunnen aantonen van metrische invariantie (de factor ladingen bleken niet gelijk tussen de twee groepen) was voldoende reden om geen verdere parameter beperkingen te toetsen, iets wat voor het bepalen van preciezere invariantie niveaus vereist is. Het hiërarchische model van dispositioneel redeneren (met een algemeen latente hoger-orde factor die de drie specifieke componenten van lagere orde – inductie, extrapolatie, en contextualisatie – beïnvloedt) liet in de gecombineerde steekproef een goede fit (passing) zien. Deze fit was niet significant slechter dan de fit van het drie-componenten-model. Hiermee werd ondersteuning gevonden voor het idee van dispositioneel redeneren als een hiërarchisch geordend construct. Het hiërarchische model toonde voorts een aanvaardbare fit in zowel de manager- als de studentensteekproeven, maar het toonde alleen configurale invariantie tussen deze groepen. Als meetinvariantie niet vastgesteld kan worden, dan kunnen (verschillen in) scoregemiddelden tussen de groepen in dispositioneel redeneren niet zomaar vergeleken en geïnterpreteerd worden. Het is dan namelijk niet duidelijk of de verschillen in scores toe te schrijven zijn aan de werkelijke verschillen tussen de twee groepen in dispositioneel redeneren, of aan verschillende psychometrische reacties op de items van de testschaal (Cheung & Rensvold, 2002). Deze bevinding betekent dat er verschillen kunnen bestaan tussen managers en psychologiestudenten in de wijze waarop de componenten van dispositioneel redeneren zich uiten. Het is dan ook van belang om verder onderzoek te doen naar de oorzaak van de verschillen in reacties op de items tussen managers en psychologiestudenten.

Samenvattend, de resultaten tonen relatief aanvaardbare meeteigenschappen van het meetinstrument voor dispositioneel redeneren in de groepen waarin de steekproef is uitgevoerd, namelijk managers en psychologiestudenten, met uitzondering van meetinvariantie.

In **Hoofdstuk 4** werd nagegaan in welke mate dispositioneel redeneren de accuraatheid van beoordelingen van sollicitanten kon voorspellen tijdens een interview. Deze studie week op twee manieren af van eerder onderzoek over dispositioneel redeneren. Ten eerste werd uitgegaan van de componenten-opvatting van dispositioneel redeneren zoals dit in Hoofdstuk 3 beschreven is. Dat wil zeggen dat de rol werd onderzocht van de sub-componenten van persoonlijkheidskenmerk-inductie, persoonlijkheidskenmerk-extrapolatie en persoonlijkheidskenmerk-contextualisatie in het voorspellen van accuraatheid. Hiertoe werd de RIJI afgenomen bij een steekproef (N = 146) van managers op middenniveau bij de

politie. Tegelijkertijd maakten deze respondenten een intelligentie- en een persoonlijkheidstest, en deden ze een taak met betrekking tot beoordelingsaccuraatheid. Deze taak bestond uit het beoordelen van gestructureerde interviews van acht sollicitanten.

Door middel van de bestudering van dispositioneel redeneren, was het mogelijk om twee dingen te onderzoeken. Ten eerste wilden we onderzoeken of er voldoende bewijs was voor het opvatten van dispositioneel redeneren als een vorm van intelligentie. Om deze vraag te kunnen beantwoorden, onderwierpen we dispositioneel redeneren aan verschillende relevante conceptuele en empirische criteria (Carrol, 1993; Flanagan, Genshaft, & Harrison, 1997; Mayer, Caruso, & Salovey, 1999). Deze reeks tests leverde ondersteuning voor de nomologische plaatsing van dispositioneel redeneren en de drie componenten ervan (inductie, extrapolatie en contextualisatie) ten opzichte van intelligentie en persoonlijkheid.

De resultaten van dit onderzoek lieten zien dat de componenten van dispositioneel redeneren onderling 'intern' voldoende convergeerden en 'extern' voldoende gerelateerd waren aan algemene intelligentie om bewijs te kunnen leveren voor een zogenaamde 'positive manifold' (Horn & Cattell, 1966). 'Positive manifold' houdt in dat metingen van cognitieve vermogens relatief hoog met elkaar correleren. Elke component voorspelde ook ($.14 < r < .33$; ongecorrigeerd) ons 'externe' criterium voor accuraatheid, namelijk de nauwkeurigheid in interviewbeoordeling, en voldeed daarmee aan het correlatie-criterium voor een meetinstrument voor cognitieve vermogens. Voorts vertoonden de componenten incrementele validiteit (klein effect; Cohen, 1988) bovenop algemene intelligentie in de voorspelling van beoordelaarsnauwkeurigheid; de componenten verklaarden variantie in nauwkeurigheid die nog niet door algemene intelligentie was verklaard. Tenslotte leverden de resultaten ook bewijs voor de discriminante validiteit van de componenten met persoonlijkheidsconstructen.

Kortom, de resultaten leverden bewijs voor een nomologisch netwerk waarbij dispositioneel redeneren gezien kan worden als een vorm van intelligentie. Dat wil zeggen dat onze resultaten ondersteuning geven voor het idee dat dispositioneel redeneren een specifiek verstandelijk vermogen is dat door goede beoordelaars aangewend wordt in hun beoordeling om gedragsinformatie in selectie-interviews te gebruiken.

In het hier beschreven onderzoek waren managers met een beter vermogen tot extrapolatie en contextualisatie accuratere beoordelaars van anderen tijdens interviews dan diegenen met onnauwkeurige impliciete persoonlijkheidstheorieën (lage extrapolatie) en een gebrekkig begrip van het activeren van persoonlijkheidskenmerken door specifieke situaties (lage contextualisatie). De inductie-component bleek overigens minder bij te dragen aan het nauwkeurig

beoordelen van anderen in interviews. Replicatie van de gevonden resultaten is echter van belang.

Onze laatste onderzoeksvraag week af van het thema van cognitieve kenmerken dat in Hoofdstuk 3 en 4 aan de orde was gekomen. In **Hoofdstuk 5** ligt de aandacht bij persoonlijkheidskenmerken van de assessoren als voorspellers van hun beoordelaarsaccuraatheid.

Uit ons eerdere overzicht van individuele verschillen als voorspellers van beoordelaarsnauwkeurigheid (Hoofdstuk 2) kon worden geconcludeerd dat persoonlijkheidskenmerken, ondanks hun potentiële theoretische relevantie, minder belangrijk lijken dan cognitieve factoren in het verklaren van de accuraatheid van ander-beoordelingen. Wij beargumenteerden voor deze conclusie dat de te voorspellen criteria te algemeen waren. In eerder onderzoek werd bijvoorbeeld gekeken hoe een interviewer sollicitanten accuraat kon evalueren op een totaal-set van persoonlijkheidskenmerken. De impliciete aanname in deze studies is dat mensen alle persoonlijkheidskenmerken even lastig vinden om anderen op te beoordelen en daarom zou een overkoepelend accuraatheidscriterium vastgesteld dienen te worden op het niveau van het persoonlijkheidsprofiel zonder onderscheid in afzonderlijke persoonlijkheidskenmerken. Onderzoek van Allik, Realo, Mõttus, en Kuppens (2010) demonstreert echter dat persoonlijkheidskenmerken onderling verschillen in hoe accuraat zij beoordeeld kunnen worden. Bijvoorbeeld, in een eerder overzicht van 32 studies (Kenny et al., 1994), was de overeenstemming tussen de beoordeling door verschillende waarnemers van dezelfde proefpersoon hoger voor de beoordeling van extraversie dan voor andere persoonlijkheidskenmerken. In een recentere studie (Borkenau et al., 2015) waarin studenten de persoonlijkheden van anderen moesten beoordelen op basis van hun essays, waren de beoordelingen van 'openheid voor nieuwe ervaringen' (openness to experience) het accuraatst. Bij elkaar genomen, lijken er voldoende argumenten te zijn voor het gebruik van accuraatheidscriteria voor specifieke persoonlijkheidskenmerken/ per persoonlijkheidskenmerk. Op deze wijze kan gesproken worden over metingen van accuraatheid per persoonlijkheidskenmerk, zoals extraversie-accuraatheid.

In Hoofdstuk 5 werden dan ook demogelijke effecten van persoonlijkheidskenmerken op beoordelaarsaccuraatheid voor elk kenmerk apart onderzocht. We berekenden accuraatheidsmetingen voor elk van de Big-Five persoonlijkheidskenmerken afzonderlijk waarop anderen beoordeeld moesten worden.

Eerdere bevindingen suggereerden dat de zelfbeoordeling van beoordelaars een belangrijke basis vormt van waaruit beoordelaars hun indrukken van anderen vormen (Alicke, Dunning, & Krueger, 2005). De kennis over bepaalde persoonlijkheidskenmerken van anderen zou immers kunnen voortkomen uit de persoonlijkheidskenmerken van de interviewer zelf. Anders gesteld, misschien zijn

assessoren beter in het beoordelen van persoonlijkheidskenmerken waarmee ze zelf zeer goed op de hoogte zijn. Bijvoorbeeld, een consciëntieuze assessor zal wellicht beter zijn in de beoordeling van consciëntieusheid bij anderen. Het 'Realistic Accuracy Model' (RAM; Funder, 1999) ondersteunt dit idee: De ontdekking van bepaalde cues bij anderen en het gebruik hiervan zouden makkelijker zijn voor persoonlijkheidskenmerken waarmee de interviewers bekend zijn omdat deze persoonlijkheidskenmerken deel van hun eigen persoonlijkheidsprofiel uitmaken.

Onder verwijzing naar de construct-toegankelijkheid theorie van Higgins (2012), formuleerden we de hypothese dat de deskundigheid in beoordeling van persoonlijkheidskenmerken voortkomt uit de mate waarin de betreffende persoonlijkheidskenmerken toegankelijk zijn voor de interviewer. Bijvoorbeeld, een persoon die 'toegang' heeft tot het construct consciëntieusheid zal geneigd zijn om anderen meer te beschrijven in woorden gerelateerd aan consciëntieusheid dan in woorden die andere persoonlijkheidskenmerken beschrijven. Om deze hypothese te toetsen werden twee studies uitgevoerd waarin een zelf-rapportage persoonlijkheidsvragenlijst werd afgenomen (Studie 1: studenten, N = 183; Studie 2: managers, N = 223; Deze steekproeven waren onafhankelijk van de steekproeven in de andere hoofdstukken). Tevens beantwoorden de managers (Studie 2) een open vraag over construct-toegankelijkheid (Higgins, 1982), waaruit hun meest toegankelijk persoonlijkheidskenmerk kon worden afgeleid. Voor zover wij weten, zijn er geen eerdere studies gedaan naar de verhouding tussen de voor interviewers chronisch-toegankelijke persoonlijkheidskenmerken, hun eigen persoonlijkheidskenmerken, en hun beoordelaarsaccuraatheid met betrekking tot deze persoonlijkheidskenmerken.

De resultaten van deze laatste studie lieten zien dat een hoge eigen score op een persoonlijkheidskenmerk (bijv. 'ik ben een zeer hartelijke persoon', een aanduiding van extraversie; Goldberg, 1992), deze persoon niet deskundiger maakte in het beoordelen van anderen op datzelfde kenmerk. Als voorbeeld, extraverte participanten in onze studies waren in vergelijking met introverte participanten niet beter in het beoordelen van bepaalde extraversie-gerelateerde gedragscues van anderen. Onze resultaten repliceerden daarmee eerdere studies (bijv. Borman, 1979; Borman & Hallam, 1991; Christiansen, et al., 2005; Hjelle, 1969; Lippa & Dietz, 2000; Powell & Goffin, 2009; Vogt & Colvin, 2003) die vonden dat persoonlijkheidskenmerken van de interviewers geen voorspellende waarde hadden voor de accuraatheid in de beoordeling van anderen op diezelfde persoonlijkheidskenmerken.

Wel bleek dat interviewers die meer geneigd waren om anderen te beschrijven in termen van extraversie en 'openheid voor nieuwe ervaringen', ook accurater waren in het beoordelen van deze persoonlijkheidskenmerken in anderen. Dit resultaat geeft daarmee gedeeltelijk steun aan onze hypothese dat chronische toegankelijkheid tot persoonlijkheidskenmerken de beoordelingsaccuraatheid voor deze

persoonlijkheidskenmerken kan voorspellen, namelijk voor extraversie en 'openheid voor nieuwe ervaringen'.

Conclusie
Het onderzoek dat in dit proefschrift wordt beschreven, laat zien dat assessorkenmerken de accuraatheid van assessoren in personeelsselectie kunnen voorspellen. Op deze manier draagt dit proefschrift bij aan het beantwoorden van de vraag: 'wat kenmerkt de goede beoordelaar'?

Uit een overzicht van de empirische literatuur blijkt dat ons begrip over kenmerken van assessoren die relevant zijn voor hun beoordelingsaccuraatheid, al grote vorderingen gemaakt heeft. Een belangrijke bevinding is dat assessoren die accurate beoordelingen over anderen geven een aantal kenmerken met elkaar gemeen lijken te hebben. Vooral cognitieve kenmerken lijken van belang.

De empirische studies in dit proefschrift geven onder andere verdere duidelijkheid over de aard en de rol van dispositioneel redeneren bij de beoordelingsaccuraatheid van de interviewer. Tevens werd bewijs geleverd dat dispositioneel redeneren behoort tot het domein van cognitieve vermogens. Bovendien bleek dat dispositioneel redeneren onderscheidende sub-componenten heeft, namelijk inductie, extrapolatie en contextualisatie.

De persoonlijkheidskenmerken van assessoren werden onderzocht in de systematische review van eerder onderzoek en in twee van de drie empirische studies. De introductie van persoonlijkheidsspecifieke criteria en chronische persoonlijkheidstoegankelijkheid als voorspellers, leverde geen veelbelovende resultaten op. Meer onderzoek is nodig om deze bevindingen te repliceren.

De resultaten geven een aantal praktische richtlijnen voor assessorenselectie en -opleiding. Het is belangrijk om assessoren te screenen op dispositioneel redeneren; de scores op de 'Revised Interpersonal Judgment Inventory' (meetinstrument voor de componenten van dispositioneel redeneren) kunnen beoordelingsaccuraatheid voorspellen. Er is verder onderzoek nodig om te bepalen of dispositioneel redeneren ontwikkeld kan worden en indien dat zo is, wat de beste methode daarvoor zou zijn.

Tot slot, met betrekking tot de persoonlijkheidskenmerken van assessoren, werden de resultaten van eerder onderzoek bevestigd: Persoonlijkheidsgerelateerde factoren lijken geen directe rol te spelen in de accuraatheid van beoordelingen van assessoren.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219-245. http://dx.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.* Ph.D., Louisiana State University and Agricultural & Mechanical College, Ann Arbor. Retrieved from http://search.proquest.com/docview/303606504?accountid=14500 ProQuest Dissertations & Theses A&I database.

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

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*, 351-371. http://dx.doi.org/10.1177/1094428114547952

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. http://dx.doi.org/10.1037/0022-3514.72.3.558

Alicke, M. D., Dunning, D., & Krueger, J. I. (Eds.). (2005). *The self in social judgment*. New York: Psychology Press.

Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., & Metspalu, A. (2010). Variance determines self-observer agreement on the Big Five personality traits. *Journal of Research in Personality, 44*, 421-426. http://dx.doi.org/10.1016/j.jrp.2010.04.005

Allik, J., Realo, A., Mõttus, R., & Kuppens, P. (2010). Generalizability of self–other agreement from one personality trait to another. *Personality and Individual Differences, 48*, 128-132. http://dx.doi.org/10.1016/j.paid.2009.09.008

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., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In P. Z. Mark (Ed.), *Advances in Experimental Social Psychology* (Vol. 32, pp. 201-271): Academic Press.

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. http://dx.doi.org/10.1037/0022-3514.69.3.518

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-Sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology, 16*, 4-13. http://dx.doi.org/10.1207/s15327663jcp1601_2

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

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality & Social Psychology, 64*, 431-441. http://dx.doi.org/10.1037/0022-3514.64.3.431

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

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290. http://dx.doi.org/10.1037/h0055756

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 128-138. http://dx.doi.org/10.1027/1614-2241/a000014

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836-874. http://dx.doi.org/10.1037/0021-9010.77.6.836

Bandalos, D. L., & Finney, S. J. (2001). Item parcelling issues in structural equation modeling. In G. A. Marcoulides & R. E. Shumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology, 50*, 869-878. http://dx.doi.org/10.1037/0022-3514.50.5.869

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology, 22*, 293-311. http://dx.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. http://dx.doi.org/10.1037/0022-3514.49.5.1129

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Barrick, M. R., Patton, G. K., & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology, 53*, 925-951. http://dx.doi.org/10.1111/j.1744-6570.2000.tb02424.x

Beauregard, K. S., & Dunning, D. (1998). Turning up the contrast: Self-enhancement motives prompt egocentric contrast effects in social judgments. *Journal of Personality and Social Psychology, 74*, 606-621. http://dx.doi.org/10.1037/0022-3514.74.3.606

Becker, B. E., & Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology, 71*, 662-671. http://dx.doi.org/10.1037/0021-9010.71.4.662

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. http://dx.doi.org/10.1037/0021-9010.67.2.151

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477. http://dx.doi.org/10.1037/0033-2909.105.3.467

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence, 20*, 309-328. http://dx.doi.org/10.1016/0160-2896(95)90013-6

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

Biesanz, J. C. (2010). The Social Accuracy Model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*, 853-885. http://dx.doi.org/10.1080/00273171.2010.519262

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494. http://dx.doi.org/10.1037/0021-9010.74.3.478

Blackman, M. C. (2002). Personality judgment and the utility of the unstructured employment interview. *Basic and Applied Social Psychology, 24*, 241-250. http://dx.doi.org/10.1207/153248302760179156

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology, 34*, 164-181. http://dx.doi.org/10.1006/jesp.1997.1347

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214-234. http://dx.doi.org/10.1207/s15327957pspr1003_2

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*, 477-492. http://dx.doi.org/10.1037/0033-2909.134.4.477

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. http://dx.doi.org/10.1016/j.jrp.2009.03.007

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*, 546-553. http://dx.doi.org/10.1037/0022-3514.65.3.546

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology, 86*, 599-614. http://dx.doi.org/10.1037/0022-3514.86.4.599

Borkenau, P., Mosch, A., Tandler, N., & Wolf, A. (2015). Accuracy of judgments of personality based on textual information on major life domains. [Advance online publication.]. *Journal of Personality*. http://dx.doi.org/10.1111/jopy.12153

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252. http://dx.doi.org/10.1016/0030-5073(77)90004-6

Borman, W. C. (1979). Individual difference correlates of rating accuracy using behavior scales. *Applied Psychological Measurement, 3*, 103-115. http://dx.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. http://dx.doi.org/10.1016/0749-5978(87)90018-5

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965-973. http://dx.doi.org/10.1037//0021-9010.86.5.965

Borman, W. C., Eaton, N. K., Bryan, J. D., & Rosse, R. L. (1983). Validity of army recruiter behavioral assessment: Does the assessor make a difference? *Journal of Applied Psychology, 68*, 415-419. http://dx.doi.org/10.1037/0021-9010.68.3.415

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. http://dx.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.* 8817333 Ph.D., Stevens Institute of Technology, Ann Arbor. ProQuest Dissertations & Theses A&I database.

Brehmer, B. (1988). Chapter 1 The development of social judgment theory. In B. Brehmer & C. R. B. Joyce (Eds.), *Advances in Psychology* (Vol. 54, pp. 13-40): North-Holland.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137-154. http://dx.doi.org/10.1016/0001-6918(94)90048-5

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Brtek, M. D., & Motowidlo, S. J. (2002). Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology, 87*, 185-191. http://dx.doi.org/10.1037//0021-9010.87.1.185

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *The handbook of social psychology* (Vol. 2, pp. 634-654). Reading, MA: Addison-Wesley.

Brunswik, E. (1956). *Perception and the representative design of experiments*. Berkeley: University of California Press.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 372-398. http://dx.doi.org/10.1080/10705511.2012.687671

Bryant, F. B., & Satorra, A. (2013). EXCEL macro file for conducting scaled difference chi-square tests via LISREL 8, LISREL 9, EQS, and Mplus. Available from the authors.

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 287-321. http://dx.doi.org/10.1207/s15328007sem1302_7

Byron, K. (2008). Differential effects of male and female managers' non-verbal emotional skills on employees' ratings. *Journal of Managerial Psychology, 23*, 118-134. http://dx.doi.org/10.1108/02683940810850772

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105. http://dx.doi.org/10.1037/h0046016

Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 3-52). New York: Academic Press.

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. http://dx.doi.org/10.1111/j.2044-8325.1987.tb00253.x

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. http://dx.doi.org/10.1016/j.jrp.2007.01.004

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In N. Helmuth (Ed.), *The scientific study of general intelligence* (pp. 5-21). Oxford: Pergamon.

Cattell, R. B. (1965). A Biometrics invited paper. Factor analysis: An introduction to essentials II: The role of factor analysis in research. *Biometrics, 21*, 405-435. http://dx.doi.org/10.2307/2528100

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. http://dx.doi.org/10.1016/j.jrp.2010.11.001

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal, 12*, 471-492. http://dx.doi.org/10.1207/s15328007sem1203_7

Cheung, G. W. (2008). Testing equivalence in the structure, means, and variances of higher-order constructs with structural equation modeling. *Organizational Research Methods, 11*, 593-613. http://dx.doi.org/10.1177/1094428106298973

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Chongming Yang, Nay, S., & Hoyle, R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement, 34*, 122-142. http://dx.doi.org/10.1177/0146621609338592

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. http://dx.doi.org/10.1207/s15327043hup1802_2

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290. http://dx.doi.org/10.1037/1040-3590.6.4.284

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

Colvin, C. R., & Bundick, M. J. (2001). In search of the good judge of personality: Some methodological and theoretical concerns. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 47-67). Mahwah, NJ.: Lawrence Erlbaum Associates.

Connelly, B., & Hulsheger, U. (2012). A narrower scope or a clearer lens for personality? Examining sources of observers' advantages over self-reports for predicting performance. *Journal of Personality, 80*, 603-631. http://dx.doi.org/10.1111/j.1467-6494.2011.00744.x

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. http://dx.doi.org/10.1037/a0021212

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection & Assessment, 15*, 110-117. http://dx.doi.org/10.1111/j.1468-2389.2007.00371.x

Cooper, A. J., Smillie, L. D., & Corr, P. J. (2010). A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality and Individual Differences, 48*, 688-691. http://dx.doi.org/10.1016/j.paid.2010.01.004

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.

Craven, C. L. (1988). *Rater accuracy training: An examination of students rating their instructor.* 8826002 Ph.D., DePaul University, Ann Arbor. Retrieved from http://search.proquest.com/docview/303641628?accountid=14500 ProQuest Dissertations & Theses A&I database.

Critcher, C. R., & Dunning, D. (2009). Egocentric pattern projection: How implicit personality theories recapitulate the geography of the self. *Journal of Personality and Social Psychology, 97*, 1-16. http://dx.doi.org/10.1037/a0015670

Critcher, C. R., Dunning, D., & Rom, S. C. (2015). Causal trait theories: A new form of person knowledge that explains egocentric pattern projection. *Journal of Personality and Social Psychology, 108*, 400-416. http://dx.doi.org/10.1037/pspa0000019

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177-193. http://dx.doi.org/10.1037/h0044919

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473. http://dx.doi.org/10.1037/h0057173

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. http://dx.doi.org/10.1037/h0040957

Cropanzano, R., Weiss, H. M., Hale, J. M. S., & Reb, J. (2003). The structure of affect: Reconsidering the relationship between negative and positive affectivity. *Journal of Management, 29*, 831-857. http://dx.doi.org/10.1016/s0149-2063_03_00081-3

Davis, M. E. (1999). *Influence of assessor individual differences on rating errors and rating accuracy in assessment centers.* 9952674 Ph.D., The University of Nebraska - Lincoln, Ann Arbor. 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.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674. http://dx.doi.org/10.1126/science.2648573

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. http://dx.doi.org/10.1080/08959285.2015.1021046

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. http://dx.doi.org/10.1002/per.795

DeGroot, T., & Motowidlo, S. J. (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology, 84*, 986-993. http://dx.doi.org/10.1037/0021-9010.84.6.986

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*, 360-396. http://dx.doi.org/10.1016/0030-5073(84)90029-1

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

Digman, J. M. (1990). Personality structure: Emergence of the Five-Factor model. *Annual Review Of Psychology, 41*, 417-440. http://dx.doi.org/10.1146/annurev.ps.41.020190.002221

Dipboye, R. L., Gaugler, B., & Hayes, T. (1990). *Individual differences among interviewers in the incremental validity of their judgments*. Paper presented at the Meeting of the Society for Industrial and Organizational Psychology, Miami, Fl.

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.

Dipboye, R. L., & Macan, T. M. (1988). A process view of the selection/recruitment interview. In R. S. Schuler, S. A. Youngblood & V. L. Huber (Eds.), *Readings in personnel and human resource management* (pp. 253-269). St Paul, MN: West Publishing Co.

Donnellan, M. B., Baird, B. M., Lucas, R. E., & Oswald, F. L. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*, 192-203. http://dx.doi.org/10.1037/1040-3590.18.2.192

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*, 9-15. http://dx.doi.org/10.1037/0021-9010.71.1.9

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology, 63*, 341-355. http://dx.doi.org/10.1037/0022-3514.63.3.341

Dunning, D., & McElwee, R. O. B. (1995). Idiosyncratic trait definitions: Implications for self-description and social judgment. *Journal of Personality and Social Psychology, 68*, 936-946. http://dx.doi.org/10.1037/0022-3514.68.5.936

Ebbesen, E. B., & Allen, R. B. (1979). Cognitive processes in implicit personality trait inferences. *Journal of Personality and Social Psychology, 37*, 471-488. http://dx.doi.org/10.1037/0022-3514.37.4.471

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological review, 85*, 395-416. http://dx.doi.org/10.1037/0033-295x.85.5.395

Eysenck, H. J. (1970). *The structure of human personality*. London: Methuen.

Farr, J. L., & Tippins, N. T. (Eds.). (2010). *Handbook of employee selection*. New York, NY: Routledge.

Fiske, S. T., & Macrae, C. N. (Eds.). (2012). *The SAGE handbook of social cognition*. Thousand Oaks, CA: SAGE Publications.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture*. New York: McGraw-Hill.

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Flanagan, D. P., Genshaft, J. L., & Harrison, P. L. (Eds.). (1997). *An integration and synthesis of contemporary theories, tests, and issues in the field of intellectual assessment*. New York: Guilford.

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. http://dx.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. http://dx.doi.org/1989-29364-001

Fletcher, G. J. O., Rosanowski, J., Rhodes, G., & Lange, C. (1992). Accuracy and speed of causal processing: Experts versus novices in social judgment. *Journal of Experimental Social Psychology, 28*, 320-338. http://dx.doi.org/10.1016/0022-1031(92)90049-p

Friedman, J. N. W., Oltmanns, T. F., & Turkheimer, E. (2007). Interpersonal perception and personality disorders: Utilization of a thin slice approach. *Journal of Research in Personality, 41*, 667-688. http://dx.doi.org/10.1016/j.jrp.2006.07.004

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*, 75-90. http://dx.doi.org/0033-2909/87/S00.75

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670. http://dx.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. http://dx.doi.org/10.1177/0963721412445309

Funder, D. C., & Colvin, C. R. (1997). Congruence of self and others' judgments of personality. In R. Hogan, J. John, J. Johnson & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 617-647). San Diego: Academic Press.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*, 409-418. http://dx.doi.org/10.1037/0022-3514.52.2.409

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44*, 107-112. http://dx.doi.org/10.1037/0022-3514.44.1.107

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61*, 457-476. http://dx.doi.org/10.1111/j.1467-6494.1993.tb00778.x

Gage, N. L., & Cronbach, L. (1955). Conceptual and methodological problems in interpersonal perception. *Psychological Review, 62*, 411-422. http://dx.doi.org/10.1037/h0047205

Gangestad, S. W., Simpson, J. A., DiGeronimo, K., & Biek, M. (1992). Differential accuracy in person perception across traits: Examination of a functional hypothesis. *Journal of Personality and Social Psychology, 62*, 688-698. http://dx.doi.org/10.1037/0022-3514.62.4.688

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Book/Harper Collins.

Gatewood, R. D., Feild, H. S., & Barrick, M. (2010). *Human resource selection*. Mason, OH: Cengage Learning.

George, E. (2006). *Interviewer accuracy across levels of structure in the employment interview.* 3226127 Ph.D., Colorado State University, Ann Arbor. 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?* Master of Psychology, Maastricht University, Maastricht.

Gibson, J. E. M. (2006). *Interpersonal perception: Don't worry, be happy.* NR27667 Ph.D., University of Victoria (Canada), Ann Arbor. Retrieved from http://search.proquest.com/docview/304983390?accountid=14500 ProQuest Dissertations & Theses A&I database.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189-211). New York: Guilford.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229. http://dx.doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*, 26-42. http://dx.doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (2005). *A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*. Paper presented at the Presidential symposium at the sixth Annual Meeting of the Association for Research in Personality, New Orleans.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96. http://dx.doi.org/10.1016/j.jrp.2005.08.007

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. http://dx.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. http://dx.doi.org/10.1111/j.1744-6570.1992.tb00852.x

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30. http://dx.doi.org/10.1037//1040-3590.12.1.19

Guion, R. M. (2011). *Assessment, measurement and prediction for personnel decisions* (2nd ed.). New York, NY: Taylor & Francis Group.

Guion, R. M., & Gibson, W. M. (1988). Personnel selection and placement. *Annual Review of Psychology, 39*, 349-374. http://dx.doi.org/10.1146/annurev.ps.39.020188.002025

Guion, R. M., & Highhouse, S. (2011). *Essentials of personnel assessment and selection*. Malwah, NJ.: Routledge.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hall, J. A., Andrzejewski, S., & Yopchick, J. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33*, 149-180. http://dx.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*. http://dx.doi.org/10.1111/jopy.12170

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion, 7*, 438-446. http://dx.doi.org/10.1037/1528-3542.7.2.438

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. http://dx.doi.org/10.1177/0146167207309192

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. *Journal of Personality and Social Psychology, 51*, 37-54. http://dx.doi.org/10.1037/0022-3514.51.1.37

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737-756. http://dx.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.* 9130321 Ph.D., City University of New York, Ann Arbor. 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. http://dx.doi.org/10.1016/0749-5978(91)90024-N

Hayes, A. F. (2014). The PROCESS macro for SPSS and SAS (Version 2.13). Retrieved from http://www.processmacro.org/download.html

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken, NJ, US: John Wiley & Sons Inc.

Heneman, H. G., Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology, 60*, 748-753. http://dx.doi.org/10.1037/0021-9010.60.6.748

Heneman, R. L., Moore, M. L., & Wexley, K. N. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research, 15*, 431-448. http://dx.doi.org/10.1016/0148-2963(87)90011-7

Higgins, E. T. (2012). Accessibility theory. In P. A. M. Van Lange, A. W. Kruglanski & E. T. Higgins (Eds.), *Handbook of theories of social psychology (Vol 1).* (pp. 75-96). Thousand Oaks, CA: Sage Publications Ltd.

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. http://dx.doi.org/10.1037/0022-3514.43.1.35

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

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*, 554-566. http://dx.doi.org/10.1177/1094428107300396

Hjelle, L. A. (1969). Personality characteristics associated with interpersonal perception accuracy. *Journal of Counseling Psychology, 16*, 579-581. http://dx.doi.org/10.1037/h0028439

Hofstee, W. K. B., de Raad, B., & Goldberg, L. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 65*, 563-576. http://dx.doi.org/10.1037/0022-3514.63.1.146

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253-270. http://dx.doi.org/10.1037/h0023816

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future-remembering the past. *Annual Review of Psychology, 51*, 631-664. http://dx.doi.org/10.1146/annurev.psych.51.1.631

Hoyt, W. T. (2010). Interrater reliability and agreement. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 193-210). New York, NY: Taylor & Francis.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897-913. http://dx.doi.org/10.1037/0021-9010.86.5.897

Huo, Y. P., Huang, H. J., & Napier, N. K. (2002). Divergence or convergence: A cross-national comparison of personnel selection practices. *Human Resource Management, 41*, 31-44. http://dx.doi.org/10.1002/hrm.10018

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59*, 730-742. http://dx.doi.org/10.1037/0022-3514.59.4.730

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10*, 311-321. http://dx.doi.org/10.5465/amr.1985.4278227

Jackson, D. N. (1972). A model for inferential accuracy. *Canadian Psychologist/Psychologie Canadienne, 13*, 185-195. http://dx.doi.org/10.1037/h0082183

Jackson, D. N. (1994). *Jackson personality inventory-revised manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N., Chan, D. W., & Stricker, L. J. (1979). Implicit personality theory: Is it illusory? *Journal of Personality, 47*, 1-10. http://dx.doi.org/10.1111/1467-6494.ep7594161

Janovics, J. E. (2003). *Knowing thyself: The influence of dispositional intelligence on self-rating accuracy.* Ph.D., Central Michigan University, Ann Arbor. Retrieved from http://search.proquest.com/docview/305221786?accountid=14500 ProQuest Dissertations & Theses A&I database.

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology, 98*, 326-341. http://dx.doi.org/10.1037/a0031257

John, O. P. (1990). The 'Big Five' factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research.* (pp. 66-100). New York, NY, US: Guilford Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O.

P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

John, O. P., Robins, R. W., & Pervin, L. A. (2008). *Handbook of personality: Theory and research* (3rd ed.). New York, NY: Guilford Press.

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. http://dx.doi.org/10.1111/j.1468-2389.2008.00429.x

Jöreskog, K., & Sörbom, D. (2015). LISREL (Version 9.2). Skokie, IL: Scientific Software International, Inc.

Jussim, L. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 37, pp. 1-93). San Diego, CA: Elsevier Academic Press.

Kasten, R., & Weintraub, Z. (1999). Rating errors and rating accuracy: A field experiment. *Human Performance, 12*, 137-153. http://dx.doi.org/10.1207/s15327043hup1202_3

Kelly, G. A. (1955). *The psychology of personality constructs*. New York, NY: Norton.

Kenny, D. A. (2004a). Interpersonal perception. In C. Spielberger (Ed.), *Encyclopedia of Applied Psychology* (Vol. 1, pp. 411-413). New York, NY: Elsevier.

Kenny, D. A. (2004b). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8*, 265-280. http://dx.doi.org/10.1207/s15327957pspr0803_3

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102*, 390-402. http://dx.doi.org/10.1037/0033-2909.102.3.390

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the big five. *Psychological Bulletin, 116*, 245-258. http://dx.doi.org/10.1037/0033-2909.116.2.245

Kenny, D. A., & Winquist, L. (2001). The measurement of interpersonal sensitivity: Consideration of design, components, and unit of analysis. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 269-310). Mahwah, NJ: Lawrence Erlbaum Associates.

Kihlstrom, J. F. (2013). The person-situation interaction. In D. Carlston (Ed.), *Oxford handbook of social cognition* (pp. 786-805). Oxford: Oxford University Press.

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. http://dx.doi.org/10.1037/0021-9010.75.5.477

Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgment: Consequences for the validity of student ratings of instruction. *Research in Higher Education, 36*, 177-195. http://dx.doi.org/10.1007/BF02207787

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757-765. http://dx.doi.org/10.1177/0013164494054003022

Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology, 63*, 739-753. http://dx.doi.org/10.1037/0022-3514.63.5.739

Klimoski, R. J., & Donahue, L. M. (2001). Person perception in organizations: An overview of the field. In M. London (Ed.), *How People Evaluate Others in Organizations*. Mahwah, NJ: Lawrence Erlbaum.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111-125). New York, NY: Guilford Publications.

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. http://dx.doi.org/10.1111/1468-2389.00217

Krause, D. E., & Thornton, G. C. (2009). A cross-cultural look at assessment center practices: Survey results from western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. http://dx.doi.org/10.1111/j.1464-0597.2008.00371.x

Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology, 18*, 1-35. http://dx.doi.org/10.1080/10463280701284645

Kruglanski, A. W. (1989). The psychology of being" right": The problem of accuracy in social perception and cognition. *Psychological Bulletin, 106*, 395-409. http://dx.doi.org/10.1037/0033-2909.106.3.395

Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological Inquiry, 1*, 181-197. http://dx.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. http://dx.doi.org/10.1002/ejsp.2420130102

Krull, D. S., & Erickson, D. J. (1995). Judging situations: On the effortful process of taking dispositional information into account. *Social Cognition, 13*, 417-438. http://dx.doi.org/10.1521/soco.1995.13.4.417

Krzystofiak, F., Cardy, R. L., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology, 73*, 515-521. http://dx.doi.org/10.1037/0021-9010.73.3.515

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. http://dx.doi.org/10.1037/0021-9010.89.1.22

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107. http://dx.doi.org/10.1037//0033-2909.87.1.72.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852. http://dx.doi.org/10.1177/1094428106296642

Lee, J., Wong, C. T., Day, J. D., Maxwell, S. E., & Thorpe, P. (2000). Social and academic intelligences: a multitrait–multimethod study of their crystallized and fluid characteristics. *Personality and Individual Differences, 29*, 539-553. http://dx.doi.org/10.1016/s0191-8869(99)00213-5

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42*, 914-932. 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. http://dx.doi.org/10.1027/1864-9335/a000007

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. http://dx.doi.org/10.1037/0022-3514.91.1.111

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: narrative and quantitative review of the research literature. *Personnel Psychology, 67*, 241-293. http://dx.doi.org/10.1111/peps.12052

Levesque, M. J., & Kenny, D. A. (1993). Accuracy of behavioral predictions at zero acquaintance: A social relations analysis. *Journal of Personality and Social Psychology, 65*, 1178-1187. http://dx.doi.org/10.1037/0022-3514.65.6.1178

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.* 3037975 Ph.D., University of Missouri - Saint Louis, Ann Arbor. 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. http://dx.doi.org/10.1037/0021-9010.86.2.255

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., De Fruyt, F., & Van Dam, K. (2001). Assessors' use of personality traits in descriptions of assessment centre candidates: A five-factor model perspective. *Journal of Occupational and Organizational Psychology, 74*, 623-636. http://dx.doi.org/10.1348/096317901167550

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. http://dx.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* (Vol. 28, pp. 99-152). Bingley: Emerald Group Publishing Limited.

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. http://dx.doi.org/10.1023/A:1006610805385

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 151-173. http://dx.doi.org/10.1207/s15328007sem0902_1

London, M. (Ed.). (2001). *How people evaluate others in organizations*. Mahwah, N.J.: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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. http://dx.doi.org/10.1037/0021-9010.70.1.66

Maass, A., Colombo, A., Colombo, A., & Sherman, S. J. (2001). Inferring traits from behaviors versus behaviors from traits: The induction–deduction asymmetry. *Journal of Personality and Social Psychology, 81*, 391-404. http://dx.doi.org/10.1037/0022-3514.81.3.391

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. http://dx.doi.org/10.1037/1082-989x.1.2.130

Markus, H., Smith, J., & Moreland, R. L. (1985). Role of the self-concept in the perception of others. *Journal of Personality and Social Psychology, 49*, 1494-1512. http://dx.doi.org/10.1037/0022-3514.49.6.1494

Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27*, 267-298. http://dx.doi.org/10.1016/S0160-2896(99)00016-1

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516. http://dx.doi.org/10.1037/0003-066x.52.5.509

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *WJ-R technical manual*. Itasca, IL: Riverside Publishing.

McLarney-Vesotski, A., Bernieri, F., & Rempala, D. (2011). An experimental examination of the "good judge". *Journal of Research in Personality, 45*, 398-400. http://dx.doi.org/10.1016/j.jrp.2011.04.005

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

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. http://dx.doi.org/10.1111/j.1468-2389.2010.00516.x

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64*, 53-87. http://dx.doi.org/10.1111/j.1744-6570.2010.01202.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*, 110-110

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. http://dx.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. http://dx.doi.org/10.1111/j.1559-1816.2003.tb02777.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

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. http://dx.doi.org/10.1037/0021-9010.87.2.293

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252-283. http://dx.doi.org/10.1037/h0035002

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in

personality structure. *Psychological Review, 102*, 246-268. http://dx.doi.org/10.1037/0033-295X.102.2.246

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*, 749-761. http://dx.doi.org/10.1037/0021-9010.91.4.749

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality, 52*, 47-54. http://dx.doi.org/10.1016/j.jrp.2014.07.005

Murphy, K. R. (1991). Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy. *Organizational Behavior and Human Decision Processes, 50*, 45-50. http://dx.doi.org/10.1016/0749-5978(91)90033-P

Murphy, K. R. (2012). Individual differences. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 31-47). New York, NY: Oxford University Press, USA.

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. http://dx.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. http://dx.doi.org/10.1037/0021-9010.74.4.619.

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. http://dx.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. http://dx.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. http://dx.doi.org/10.1016/j.intell.2010.10.001

Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. *Academy of Management Review, 10*, 109-115. http://dx.doi.org/10.5465/AMR.1985.4277361

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101. http://dx.doi.org/10.1037/0003-066X.51.2.77

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Boyle, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior, 32*, 788-818. 10.1002/job.714

Olsson, N. (2000). A comparison of correlation, calibration, and diagnosticity as measures of the confidence–accuracy relationship in witness identification. *Journal of Applied Psychology, 85*, 504-511. http://dx.doi.org/10.1037/0021-9010.85.4.504

Ostroff, C., & Ilgen, D. (1992). Cognitive categories of raters and rating accuracy. *Journal of Business and Psychology, 7*, 3-26. http://dx.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. http://dx.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.* NR03880 Ph.D., University of Calgary (Canada), Ann Arbor. Retrieved from http://search.proquest.com/docview/305029729?accountid=14500 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.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56*, 823-833. http://dx.doi.org/10.1037/0022-3514.56.5.823

Porter, S., Campbell, M. A., Stapleton, J., & Birt, A. R. (2002). The influence of judge, target, and stimulus characteristics on the accuracy of detecting deceit. *Canadian Journal of Behavioural Science/Revue canadienne des Sciences du comportement, 34*, 172-185. http://dx.doi.org/10.1037/h0087170

Powell, D. M. (2008). *Assessing personality in the employment interview: The impact of rater training and individual differences on rating accuracy.* Ph.D., The University of Western Ontario (Canada), Ann Arbor.

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. http://dx.doi.org/10.1080/08959280903248450

Preacher, K., & Hayes, A. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731. http://dx.doi.org/10.3758/BF03206553

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879-891. http://dx.doi.org/10.3758/BRM.40.3.879

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. C. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology, 49*, 85-102. http://dx.doi.org/10.1111/j.1744-6570.1996.tb01792.x

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater

reliability. *Journal of Applied Psychology, 93*, 959-981. http://dx.doi.org/10.1037/0021-9010.93.5.959

Qualtrics Development Company. (2015). Qualtrics (Version August). Provo, Utah, USA: Qualtrics. Retrieved from http://www.qualtrics.com

Rand, T. M., & Wexley, K. N. (1975). Demonstration of the effect, "similar to me", in simulated employment interviews. *Psychological Reports, 36*, 535-544. http://dx.doi.org/10.2466/pr0.1975.36.2.535

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472-494). New York, NY: Guilford.

Raykov, T., Marcoulides, G. A., & Li, C. H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement, 72*, 954-974. http://dx.doi.org/10.1177/0013164412441607

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287-297. http://dx.doi.org/10.1037/1040-3590.12.3.287

Revelle, W., Wilt, J., & Condon, D. M. (2013). Individual differences and differential psychology. In T. Chamorro-Premuzic, A. Furnham & S. von Stumm (Eds.), *The Wiley-Blackwell handbook of individual differences* (pp. 1-38): Wiley-Blackwell.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*, 32-47. http://dx.doi.org/10.1207/s15327957pspr0901_3

Robinson, O. C. (2009). On the social malleability of traits: variability and consistency in Big 5 trait expression across three interpersonal contexts. *Journal of Individual Differences, 30*, 201-208. http://dx.doi.org/10.1027/1614-0001.30.4.201

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. http://dx.doi.org/10.1111/j.2044-8325.2011.02045.x

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464-480. http://dx.doi.org/10.1037/a0038098

Rosenbaum, A. L. (1992). *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. Retrieved from http://search.proquest.com/docview/304018100?accountid=14500 ProQuest Dissertations & Theses A&I database.

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. http://dx.doi.org/10.1037/0021-9010.66.4.442

Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359-392. http://dx.doi.org/10.1111/j.1744-6570.1999.tb00165.x

Ryan, A. M., & Sackett, P. R. (1989). Exploratory study of individual assessment practices: Interrater reliability and judgments of assessor effectiveness. *Journal of Applied Psychology, 74*, 568-579. http://dx.doi.org/10.1037/0021-9010.74.4.568

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428. http://dx.doi.org/10.1037/0033-2909.88.2.413

Sackett, P. R. (1979). *The interviewer as hypothesis tester: The effects of impressions of an applicant on subsequent interviewer behavior.* 8001820 Ph.D., The Ohio State University, Ann Arbor. ProQuest Dissertations & Theses A&I database.

Sackett, P. R. (1982). The interviewer as hypothesis tester: The effects of impressions of an applicant on interviewer questioning strategy. *Personnel Psychology, 35*, 789-804. http://dx.doi.org/10.1111/j.1744-6570.1982.tb02222.x

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. http://dx.doi.org/10.1016/0030-5073(79)90051-5

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59*, 419-450. http://dx.doi.org/10.1146/annurev.psych.59.103006.093716

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology, 67*, 10-17. http://dx.doi.org/10.1037/0021-9010.67.1.10

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103-108. http://dx.doi.org/10.1111/j.2044-8325.1997.tb00634.x

Sait, Z. (2014). *Judge-target trait similarity and accuracy: Does it 'take one to know one'?* M.Com (Org Psych), University of Cape Town, Cape Town.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work & organizational psychology - volume 1: Personnel psychology* (Vol. 1, pp. 165-200). London: SAGE Publications Ltd.

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. http://dx.doi.org/10.1006/obhd.1993.1023

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. http://dx.doi.org/10.1037/0021-9010.81.1.3

Saucier, G., & Goldberg, L. R. (1996). Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality, 10*, 61-77. http://dx.doi.org/10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D

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. http://dx.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. http://dx.doi.org/10.1037/0033-2909.124.2.262

Schmitt, N., & Chan, D. (1998). *Personnel selection: a theoretical approach*: Sage Publications.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222. http://dx.doi.org/10.1016/j.hrmr.2008.03.003

Schneider, D. E., & Bayroff, A. G. (1953). The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology, 37*, 278-280. http://dx.doi.org/10.1037/h0062458

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79*, 294-309. http://dx.doi.org/10.1037/h0034496

Schneier, C. E. (1977). Operational utility and psychometric characteristics of Behavioral Expectation Scales: A cognitive reinterpretation. *Journal of Applied Psychology, 62*, 541-548. http://dx.doi.org/10.1037/0021-9010.62.5.541

Shafer, A. B. (1999). Relation of the big five and factor V subcomponents to social intelligence. *European Journal of Personality, 13*, 225-240. http://dx.doi.org/10.1002/(SICI)1099-0984(199905/06)13:3<225::AID-PER337>3.0.CO;2-V

Sheppard, L. D., Goffin, R. D., Lewis, R. J., & Olson, J. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology, 10*, 24-33. http://dx.doi.org/10.1027/1866-5888/a000030

Shoda, Y., Mischel, W., & Wright, J. C. (1989). Intuitive interactionism in person perception: Effects of situation-behavior relations on dispositional judgments. *Journal of Personality and Social Psychology, 56*, 41-53. http://dx.doi.org/10.1037/0022-3514.56.1.41

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Smith, E. R., & Branscombe, N. R. (1988). Category accessibility as implicit memory. *Journal of Experimental Social Psychology, 24*, 490-504. http://dx.doi.org/10.1016/0022-1031(88)90048-0

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. http://dx.doi.org/10.1016/0749-5978(87)90022-7

Spearman, C. (1904). " General Intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-292. http://dx.doi.org/10.2307/1412107

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. http://dx.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. http://dx.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. http://dx.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. http://dx.doi.org/10.1037/0033-295x.96.1.58

Stanovich, K. E. (1992). *How to think straight about psychology* (3rd ed.). New York, NY: Harper Collins Publishers.

Sternberg, R. J. (Ed.). (2000). *Handbook of intelligence*. New York, NY: Cambridge University Press.

Strupeck, S. A. (2004). *Assessment center ratings in a social context: The effect of accountability on assessor ratings.* 3150401 Ph.D., The University of Tulsa, Ann Arbor. 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. http://dx.doi.org/10.1037/0021-9010.73.3.497

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin, 52*, 1-23. http://dx.doi.org/10.1037/h0044999

Taft, R. (1966). Accuracy of empathic judgments of acquaintances and strangers. *Journal of Personality and Social Psychology, 3*, 600-604. http://dx.doi.org/10.1037/h0023288

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*, 397-423. http://dx.doi.org/10.1006/jrpe.2000.2292

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29. http://dx.doi.org/10.1037/h0071663

Tonidandel, S., & LeBreton, J. (2011). Relative Importance Analysis: A useful supplement to regression analysis. *Journal of Business & Psychology, 26*, 1-9. http://dx.doi.org/10.1007/s10869-010-9204-3

Townsend, R. J., Bacigalupi, S. C., & Blackman, M. C. (2007). The accuracy of lay integrity assessments in simulated employment interviews. *Journal of Research in Personality, 41*, 540-557. http://dx.doi.org/10.1016/j.jrp.2006.06.010

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*, 239-257. http://dx.doi.org/10.1037/0033-295x.93.3.239

Trope, Y., Cohen, O., & Alfieri, T. (1991). Behavior identification as a mediator of dispositional inference. *Journal of Personality and Social Psychology, 61*, 873-883. http://dx.doi.org/10.1037/0022-3514.61.6.873

Trope, Y., & Higgins, E. T. (1993). The what, when, and how of dispositional inference: New answers and new questions. *Personality and Social Psychology Bulletin, 19*, 493-500. http://dx.doi.org/10.1177/0146167293195002

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. http://dx.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. http://dx.doi.org/10.1037/0021-9010.93.3.711

Uleman, J. S., & Bargh, J. A. (Eds.). (1989). *Unintended thought*. New York, NY: Guilford Press.

Van Dam, K. (2003). Trait perception in the employment interview: A Five–Factor model perspective. *International Journal of Selection and Assessment, 11*, 43-55. http://dx.doi.org/10.1111/1468-2389.00225

Van der Kloot, W. A., & Kroonenberg, P. M. (1982). Group and individual implicit theories of personality: An application of three-mode principal component analysis. *Multivariate Behavioral Research, 17*, 471-491. http://dx.doi.org/10.1207/s15327906mbr1704_3

Van Iddekinge, C. H., Sager, C. E., Burnfield, J. L., & Heffner, T. S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment, 14*, 193-205. http://dx.doi.org/10.1111/j.1468-2389.2006.00352.x

Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of Vocational Behavior, 65*, 71-95. http://dx.doi.org/10.1016/s0001-8791(03)00076-9

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. http://dx.doi.org/10.1177/109442810031002

Vernon, P. E. (1933). Some characteristics of the good judge of personality. *Journal of Social Psychology, 4*, 42-57. http://dx.doi.org/10.1080/00224545.1933.9921556

Vogt, D. S., & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of Personality, 71*, 267-295. http://dx.doi.org/10.1111/1467-6494.7102005

Vroom, V. H. (1964). *Work and motivation*. New York: John Wiley & Sons, Inc.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: a theory of rating. *Personnel Psychology, 35*, 521-551. http://dx.doi.org/10.1111/j.1744-6570.1982.tb02208.x

Wiggins, N. H., & Blackburn, M. C. (1976). Implicit theories of personality: An individual differences approach. *Multivariate Behavioral Research, 11*, 267-285. http://dx.doi.org/10.1207/s15327906mbr1103_1

Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior, 13*, 357-367. http://dx.doi.org/10.1002/job.4030130404

Woehr, D. J., & Arthur Jr, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. [Article]. *Journal of Management, 29*, 231-258. http://dx.doi.org/10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. http://dx.doi.org/10.1111/j.2044-8325.1994.tb00562.x

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. http://dx.doi.org/10.1016/s0191-8869(98)00068-3

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*. http://dx.doi.org/10.1177/0013164413495237

Wonderlic, E. F. (1998). *Manual of the Wonderlic Personnel Test & scholastic level exam*. Libertyville, Ill: Wonderlic Personnel Test, Inc.

Wonderlic Personnel Test, I. (2002). *Wonderlic Personnel Test & scholastic level exam user's manual*. Libertyville, IL: Wonderlic Personnel Test, Inc.

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. http://dx.doi.org/10.1111/j.1748-8583.2008.00067.x

Wyer, R. S., & Srull, T. K. (Eds.). (2014). *Handbook of social cognition: Basic processes* (2nd ed. Vol. 1). New York, NY: Psychology Press.

Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry, 22*, 159-182. http://dx.doi.org/10.1080/1047840x.2011.551743

Zalesny, M. D., & Highhouse, S. (1992). Accuracy in performance evaluations. *Organizational Behavior and Human Decision Processes, 51*, 22-50. http://dx.doi.org/10.1016/0749-5978(92)90003-p

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance, 18*, 269-294. http://dx.doi.org/10.1016/0030-5073(77)90031-9

Zedeck, S., Tziner, A., & Middlestadt, S. E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology, 36*, 355-370. http://dx.doi.org/10.1111/j.1744-6570.1983.tb01443.x

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. http://dx.doi.org/10.1080/08959285.2010.501049

Curriculum Vitae

François de Kock was born on January 15th 1976 in Uitenhage, South Africa. He completed his secondary education in 1993 at the Point High School in Mossel Bay. After completing voluntary military service and officer training at the South African Army Gymnasium in 1995, he graduated in 1998 from the Military Academy in Saldanha, which is also the Faculty of Military Science of Stellenbosch University. Shortly after receiving an officer posting in the Army, he continued with his postgraduate studies part-time, completing his Honours Degree in 2001. He received an appointment as junior lecturer at the Military Academy and completed his Masters Degree in 2004 at Stellenbosch University. Following a brief study exchange to Ghent University (Belgium) in 2004, François took up an appointment as industrial-organizational psychologist at the Military Psychological Institute in Pretoria, South Africa, where he found an interest in personnel selection research and development. He conducted various projects for the military, not only in South Africa, but also in various African countries (in diplomatic and United Nations peacekeeping missions). In January 2007, François was appointed as lecturer at the Department of Industrial Psychology, Stellenbosch University. In September 2008, he started a PhD project at the Institute of Psychology at the Erasmus University Rotterdam, studying individual differences in interviewer judgment accuracy. He was offered a position as senior lecturer at the School of Management, University of Cape Town, South Africa, in December 2012. He lectures in the area of psychometrics, psychological assessment, and research methods. His research and consulting interests fall in the broad area of personnel selection, with a special interest in assessor-related topics.

Dankwoord

"I always knew that deep down in every human heart, there was mercy and generosity."
— Nelson Mandela, Long Walk to Freedom

Sonder die vrygewigheid van vele sou hierdie proefskrif telkens gesneuwel het! Dit is my voorreg om elkeen te bedank wat bygedra het op welke wyse, klein óf groot.

Spesiaal wil ek my promotors, Marise Born en Filip Lievens, bedank. Jul uiters bekwame leiding het my deurgaans die moed gegee om my vlerke te sprei. Wat 'n voorreg om hierdie reis met julle te kon meemaak! Desondanks besige skedules het julle tyd gevind vir oorlegpleging, skryfbegeleiding, en motivering. Dikwels het dit gepaard gegaan met reis weg van jul geliefdes. Dankie ook aan Arnold Bakker, Reinout de Vries, en Henk van der Molen vir jul inskiklikheid om te dien in die leeskommissie. Ek was inderdaad deel van 'n prima span.

Dankie aan ons navorsingsvennote by verskeie instansies. Die hulp van Kolonel Charlotte Kotzé, Bevelvoerder van die Polisie Akademie van die Suid-Afrikaanse Polisiediens (SAPD), asook Prof Deon Meiring van die Universiteit Pretoria, het die aanvanklike navorsingstudies moontlik gemaak. Lede van die Suid-Afrikaanse Militêre Gesondheidsdiens het belangrike bydraes gemaak tot die voorafgaande ontwikkeling van die navorsingsmateriaal. In besonder, Generaal Warren Burgess en personeel by die Militêre Psigologiese Instituut (MPI), Luitenant-kolonel Natascha Bruwer, Majoor Yasmeen Ibrahim, en bedryfsielkundiges van die Instituut. Sonder finansiële ondersteuning sou die navorsing nie kon begin óf tot voltooiing gebring word nie en daarom bedank ons graag die volgende organisasies: Andrew W. Mellon Foundation (VSA), National Research Fund (NRF), Universiteit van Kaapstad Navorsingsfonds, en Departement Bedryfsielkunde van die Universiteit van Stellenbosch. Verskeie deskundiges het sonder huiwering gehelp met advies, navorsingsinstrumente, of programkode, onder andere Neil Christiansen, Walter Borman, Barbary Byrne, Deborah Powell, Lorne Sulsky, Bill Balzer en Brenton Wiernik. Baie dankie vir die vrymoedige gees waarin dit gedoen is.

Die respondente in ons navorsing verdien spesiale vermelding. Hieronder tel studente van die Universiteit Kaapstad, Universiteit Pretoria, asook offisier-leerders van die SAPD Akademie. Bestuurders en personeel van verskeie organisasies (bv. Sanlam, en ander wat anonimiteit verkies) het ook bygedra. Ons wense is dat die bedryf in gelyke mate sal baat vind by die navorsingsbevindinge en vrae ontsluit vanuit ons navorsing.

Dankie aan my kollegas by die Erasmus Universiteit Rotterdam (EUR), Universiteit Kaapstad, en Universiteit Stellenbosch, vir jul professionele ondersteuning. As

'buitenpromovendus' van EUR was ek diep getref deur kollegas se begrip en deernis met my unieke behoeftes. Die mildelike geduld van my departementshoofde (Johan Malan, Suki Goodman, en Anton Schlechter), mede-dosente, en studente, was soms nodig met onvoorsiene vertragings in my response op boodskappe, nasienwerk en administrasie. Verder wil ek my mentors bedank vir die steierwerk wat bygedra het tot sukses met hierdie projek: Elize Kotze, Gielie van Dyk, Johan Malan, Callie Theron, Gert Roodt, en Gert Huysamen.

Dankie aan my paranimfe wat, ten spyte van hul eie werkslading, dikwels tyd gemaak het om 'n luisterende oor, of helpende hand, te verleen. Anton, jou opbouende woorde en begrip het my gereeld aangemoedig om nog 'n tree vorentoe te gee. En Gera, ten spyte van die afstand (tussen Kaapstad en Rotterdam) het jy dikwels uit jou pad gegaan om sake vir my te beredder. Dit is lekker om julle langs my te hê met die uitroep van "Hora Est!".

Vriende het ons gesin deur dik en dun bygestaan: Jaco, François, Evert, Ester, en die Rossles, Swarts, Adamse, Liversages en Schoemans. Dit het ongelooflik baie beteken vir ons. My broers, Deon, Wilhelm, en Charl, dankie vir jul broederliefde. En my skoonouers, Dan, Lynette en Pieter, en stiefpa, Joe, dankie vir jul deurlopende belangstelling en ondersteuning. My moeder, Marilen, jy is van kindsbeen my grootste 'cheerleader'; dankie dat Ma nooit opgegee het nie. In my hart wens ek ook dat Pappa (Willem) hierdie groot oomblik kon meemaak. En laastens, aan my vrou Jammies (wat die spreekwoordelike sewe sakke sout moes trotseer): woorde kan nie beskryf wat jy vir my beteken nie. Aan jou, en ook die ligstraaltjies in ons lewe, Dieter en Jana: Pappa is nou klaar met sy 'boek'! Julle is vir my kosbaar.

Dankie aan elkeen. Laastens, dankie aan my Hemelse Vader vir hierdie voorreg.

François

Kaapstad, 2015

François
de Kock

INDIVIDUAL DIFFERENCES IN JUDGMENT ACCURACY IN PERSONNEL SELECTION:

What Makes the 'Good Judge'

studioLIN.nl