An In-Depth Look at Dispositional Reasoning and Interviewer Accuracy
François S. De Kock\textsuperscript{a}, Filip Lievens\textsuperscript{b} & Marise Ph. Born\textsuperscript{a}
\textsuperscript{a} Erasmus University Rotterdam
\textsuperscript{b} Ghent University
Published online: 07 Jul 2015.

To cite this article: François S. De Kock, Filip Lievens & Marise Ph. Born (2015) An In-Depth Look at Dispositional Reasoning and Interviewer Accuracy, Human Performance, 28:3, 199-221, DOI: 10.1080/08959285.2015.1021046
To link to this article: http://dx.doi.org/10.1080/08959285.2015.1021046

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &
An In-Depth Look at Dispositional Reasoning and Interviewer Accuracy

François S. De Kock
Erasmus University Rotterdam

Filip Lievens
Ghent University

Marise Ph. Born
Erasmus University Rotterdam

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 the dispositional reasoning construct. Therefore, we use a componential approach to test if dispositional reasoning adheres to classical criteria for an intelligence. Results from 146 managerial interviewers who observed videotaped interviewees showed that 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 classical criteria for an intelligence.

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, Macan, & Shahani-Denning, 2012) concludes that good judges 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 (e.g., see Letzring, 2008; McLarney-Vesotski, Bernieri, & Rempala, 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

Correspondence should be sent to François S. De Kock, Erasmus University Rotterdam, Private Bag, Rondebosch, Cape Town 7701, South Africa. E-mail: francois.dekock@uct.ac.za

1In this article, we use the terms judges, raters, assessors, and interviewers interchangeably.
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, labeled 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 covary), and *trait contextualization* (the ability to identify situations that are relevant to different traits; see Figure 1).

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 the subcomponents of dispositional reasoning in accuracy is unclear. Prior studies on dispositional reasoning did not consider its underlying components. Given that they could not measure the components of dispositional reasoning reliably, earlier studies collapsed them into a single, broad measure. As a result, we know little about the componential nature of dispositional reasoning and how these components individually 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 of dispositional reasoning adhere to the criteria for classical intelligence measures.

It is important to determine whether dispositional reasoning represents an intelligence, for both practical and theoretical reasons. On one hand, it could expand our insight into the specific rater constructs (see Jones & Born, 2008) that affect accuracy. Practically, 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.

**STUDY BACKGROUND**

**Criteria for an Intelligence**

For it to be considered an intelligence, a specific mental ability must meet several criteria (Carroll, 1993; Flanagan, Genshaft, & Harrison, 1997; Mayer, Caruso, & Salovey, 1999). First, a conceptual 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,

---

2Note that we assess the conceptual criterion, in other words, theoretical interrelatedness of the three components of dispositional reasoning, in our discussion of positive manifold.
FIGURE 1 Understanding the components of dispositional reasoning: trait induction, trait extrapolation and trait contextualization.
DE KOCK, LIEVENS, BORN

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, that is, be empirically distinct. To this correlational criterion, we add the requirement that it should predict related external criteria (e.g., 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 (Nisbett et al., 2012). 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 covary 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 16 narrow abilities was .37, indicating medium to large 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,

H1: The components of dispositional reasoning will show positive manifold internally, that is, 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 1).

\[^{3}\text{We used Cohen’s (1988) guidelines to interpret effect sizes for Pearson correlations, that is, small (.10), medium (.30), and large (.50) effects. However, to interpret effect sizes for incremental validity, we used Cohen’s guidelines for hierarchical regression } f^2 \text{ as small (.02), medium (.15), and large (.35) effects.}\]
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 covary 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: (a) behavioral identification, where behavior is evaluated in terms of relevant categories, and (b) 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, 1994). Ultimately, interviewers who make correct behavior-trait inferences would therefore form a more accurate overall impression of the interviewee.

Considering these arguments, we posit:

\[ H2a: \text{Trait induction will moderately (} r > .30 \text{) relate to interviewer accuracy.} \]

**Trait Extrapolation and Interviewer Accuracy**

Trait extrapolation can be defined as the understanding of how traits and their behavioral manifestations naturally covary. 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 covary. 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 trait covariation is a fundamental premise of implicit personality theory (IPT; Jackson, Chan, & Stricker, 1979; Schneider, 1973). This theory posits that people use naïve, commonsense 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-min 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 the following:

**H2b:** 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 or he 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 10 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,

**H2c:** Trait contextualization will moderately \((r > .30)\) relate to interviewer accuracy.

**Discriminant Validity With Personality**

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, & Pervin, 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 the following:

**H3:** 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,

H4: 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 \( (f^2 > .02)\).

METHOD

Participants

We recruited participants (police managers) undergoing a 7-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 5 years) comparable work experience as managers. There were 146 managers in the sample (24.6% female, 75.4% male). In terms of race, the sample comprised 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%), whereas 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 postsecondary 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, Conway, Roth, & Stone, 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 preceding item, 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 from 1 (poor response) to 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 qualified industrial psychologists (with at least a master’s degree in IO Psychology) and professors in IO Psychology—to rate the videotaped 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, $\text{ICC}_{tot}(2, k) = .86$, and 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 at the beginning of the testing session. The Wonderlic Personnel Test is widely used to measure general cognitive ability (Wonderlic Personnel Test, Inc., 2002). It is a 50-item timed test (12 min), 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 generally vary between .82 and .95 (Wonderlic Personnel Test, Inc., 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.

**Revision of the measure.** 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 that 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 nonuniversity 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 cutoffs 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

---

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

5 Because our pilot sample was small, we applied liberal cutoffs in the item analysis that would flag only those items that clearly fell short of our requirements for further consideration.
DISPOSITIONAL REASONING AND INTERVIEWER ACCURACY

(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, that is, 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, 2006; 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 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, from very inaccurate to very accurate (Goldberg, 2005). In our sample, the mean interitem correlations for each scale were comparable to those in earlier published studies (Donnellan et al., 2006).

RESULTS

Descriptive Statistics

Table 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 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
<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>—</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. General mental ability</td>
<td>12.46</td>
<td>4.40</td>
<td>.34**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Dispositional reasoning</td>
<td>46.37</td>
<td>13.38</td>
<td>.30**</td>
<td>.68**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4. Trait induction</td>
<td>36.21</td>
<td>22.38</td>
<td>.32**</td>
<td>.52**</td>
<td>.79**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5. Trait extrapolation</td>
<td>51.22</td>
<td>19.24</td>
<td>.18*</td>
<td>.47**</td>
<td>.72**</td>
<td>.48**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6. Trait contextualization</td>
<td>56.57</td>
<td>23.05</td>
<td>.19</td>
<td>.61**</td>
<td>.79**</td>
<td>.51**</td>
<td>.41**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7. Extraversion</td>
<td>11.96</td>
<td>4.58</td>
<td>—20*</td>
<td>—26**</td>
<td>—31**</td>
<td>—25**</td>
<td>—18</td>
<td>—27**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8. Agreeableness</td>
<td>13.03</td>
<td>4.38</td>
<td>.16</td>
<td>.28**</td>
<td>.25**</td>
<td>.16</td>
<td>.23*</td>
<td>.18</td>
<td>—11</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9. Conscientiousness</td>
<td>14.13</td>
<td>4.03</td>
<td>.04</td>
<td>.16</td>
<td>.20*</td>
<td>.11</td>
<td>.16</td>
<td>.10</td>
<td>—19*</td>
<td>.40**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>11. Intellectance/Imagination</td>
<td>11.93</td>
<td>3.83</td>
<td>—01</td>
<td>.11</td>
<td>.07</td>
<td>.05</td>
<td>.06</td>
<td>.05</td>
<td>.03</td>
<td>.40**</td>
<td>.54**</td>
<td>.20*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>12. Interviewer accuracyb</td>
<td>.99</td>
<td>.65</td>
<td>.16</td>
<td>.20*</td>
<td>.34**</td>
<td>.14</td>
<td>.33**</td>
<td>.26**</td>
<td>—12</td>
<td>.15</td>
<td>.13</td>
<td>.07</td>
<td>.04</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>13. Elevation accuracyc</td>
<td>.64</td>
<td>.46</td>
<td>—10</td>
<td>—16</td>
<td>—22*</td>
<td>—21*</td>
<td>—26**</td>
<td>—20*</td>
<td>.17</td>
<td>—09</td>
<td>—16</td>
<td>—12</td>
<td>—03</td>
<td>—28**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>14. Differential elevationc</td>
<td>.66</td>
<td>.29</td>
<td>—13</td>
<td>—12</td>
<td>—08</td>
<td>.02</td>
<td>.00</td>
<td>—11</td>
<td>.00</td>
<td>.02</td>
<td>.01</td>
<td>.01</td>
<td>.12</td>
<td>—23**</td>
<td>—10</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>15. Stereotype accuracyc</td>
<td>.24</td>
<td>.21</td>
<td>—12</td>
<td>—03</td>
<td>—14</td>
<td>—08</td>
<td>—15</td>
<td>—09</td>
<td>.05</td>
<td>—02</td>
<td>.03</td>
<td>.09</td>
<td>.10</td>
<td>—21*</td>
<td>.07</td>
<td>—05</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>16. Differential accuracyc</td>
<td>.46</td>
<td>.22</td>
<td>—04</td>
<td>—12</td>
<td>—18*</td>
<td>—11</td>
<td>—04</td>
<td>—19*</td>
<td>.02</td>
<td>—05</td>
<td>—06</td>
<td>.02</td>
<td>.08</td>
<td>—11</td>
<td>—03</td>
<td>—03</td>
<td>.10</td>
<td>—</td>
</tr>
</tbody>
</table>

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

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).
DISPOSITIONAL REASONING AND INTERVIEWER ACCURACY

Tests of Hypotheses

H1 proposed that the dispositional reasoning components would show positive manifold internally (i.e., they would moderately relate to each other; \( r > .30 \)) as well as externally (i.e., they would moderately relate to a measure of general mental ability; \( r > .30 \)). The correlations (see Table 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 (CIs) generally contained the hypothesized effect ranges, \( \min \ r = .41, \ 95\% \ CI \ [.27, .54], \ max \ r = .61 \ [.50, .70] \). Therefore, H1 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, \( \chi^2(3, N = 142) = 86.078, p < .01, \ root \ mean \ square \ error \ of \ approximation = .44, \ 90\% \ CI \ [.36, .52] \). 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 H1 focused on the positive manifold (“related-set”) criterion for a classical intelligence measure, H2 proposed that trait induction (H2a), trait extrapolation (H2b), and trait contextualization (H2c) would predict interview judgment accuracy, all with moderate \( (r > .30) \) effects. As shown in Table 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, H2 was partially supported as one effect size was weaker than hypothesized.

H3 stated that the three components of dispositional reasoning would show discriminant validity with Big Five personality measures. Table 1 shows that only three (of 15) bivariate relationships between these facets showed small to medium effects, whereas 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], whereas 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 H3.

H4 posited that individual differences in dispositional reasoning components would increment the validity of general mental ability to predict accuracy. Table 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 ($\Delta R^2 = .09, p = .004$) 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 H4.

In addition to the analyses that tested our hypotheses, we also conducted relative weights analysis (Tonidandel & LeBreton, 2011, 2014) to examine which of the three components of dispositional reasoning was most important in determining interview accuracy. As shown in Table 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%).

### 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.

### Table 2: Results of Hierarchical Regression Analyses of Interviewer Rating Accuracy$^a$ (Borman’s Differential Accuracy) on General Mental Ability and Dispositional Reasoning Components

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Step 1</th>
<th></th>
<th>Step 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\text{RW}_{raw}$ [95% CI]</td>
<td>$\text{RW}^b$%</td>
</tr>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General mental ability</td>
<td>.22$^*$</td>
<td>.00</td>
<td>0.01 [0.003, 0.046]</td>
<td>9.96%</td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Induction</td>
<td>-.09</td>
<td>0.01</td>
<td>[0.000, .011]</td>
<td>4.74%</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>.31$^{**}$</td>
<td>0.08*</td>
<td>[0.025, .164]</td>
<td>58.54%</td>
</tr>
<tr>
<td>Contextualization</td>
<td>.18</td>
<td>0.04</td>
<td>[0.004, .113]</td>
<td>26.76%</td>
</tr>
<tr>
<td>Total $R^2$</td>
<td>.05$^*$</td>
<td>.14**</td>
<td></td>
<td>100.00%</td>
</tr>
<tr>
<td>$\Delta R^2$</td>
<td>.05$^*$</td>
<td>.09**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. N = 142. RW = relative weight; CI = confidence interval; Induction = judges’ ability to infer traits from behavior cues; Extrapolation = understanding of how traits naturally covary; Contextualization = to know how situations affect trait expression.

$^a$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 $R^2$. CI around the raw weights were calculated using the bias corrected accelerated method for generating the bootstrapped CIs.

$^*p < .05. \quad ^{**}p < .01.
### DISCUSSION

**Main Conclusions**

In personnel selection, there has been a long-standing 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 its components—induction, extrapolation, and contextuality—are important predictors of rating accuracy.

**TABLE 3**

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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Step 1</th>
<th>Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\beta)</td>
<td>(\beta)</td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General mental ability</td>
<td>−.12</td>
<td>−.01</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Induction</td>
<td>−.06</td>
<td></td>
</tr>
<tr>
<td>Extrapolation</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Contextualization</td>
<td>−.18</td>
<td></td>
</tr>
<tr>
<td>Total (R^2)</td>
<td>.01</td>
<td>.04</td>
</tr>
<tr>
<td>(\Delta R^2)</td>
<td>.01</td>
<td>.02</td>
</tr>
</tbody>
</table>

*Note.* \(N = 142\). \(\text{RW} = \) relative weight; CI = confidence interval; 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.

\(^a\)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 \(R^2\). CIs around the raw weights were calculated using the bias corrected accelerated method for generating the bootstrapped CIs.

\(*p < .05. \ **p < .01.\)

(−.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 3) than in the correlational accuracy measure \((R^2 = .14;\) see Table 2). Also, the dispositional reasoning components did not show incremental validity \((\Delta R^2 = .02)\) over general mental ability when the differential accuracy measure was used.
contextualization—would act as classical intelligence measures in predicting interview judgment accuracy. We tested the components against strict conceptual and correlational criteria for a classical 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 an intelligence. That is, they broadly adhered to the correlational criteria we tested. For example, the dispositional reasoning components converged with one another and with general mental ability (i.e., positive manifold). Also, as in previous studies (e.g., Christiansen et al., 2005), broad dispositional reasoning 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: 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 are 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 covary (extrapolation), and how situations affect trait expression (contextualization), than knowing which traits are signaled by behavioral cues (induction). The relative weights analysis confirmed that, among 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 operationalizations 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), or 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 nonstudent samples. Yet it would be interesting to know if 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 videotaped 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. In addition, 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 are 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). Although 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 is an intelligence raises questions about the malleability of the dispositional reasoning components. If it is an intelligence, how does it develop, if at all? Classical 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 as an intelligence. 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, & Wolfe, 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, standardization 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 interviewees 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 if the dispositional reasoning-accuracy link exists with nonmanagerial judges, such as psychologists, trained assessment center 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, Woehr, Mishra, & Kieszczynska, 2012; Woehr & Huffcutt, 1994). However, FOR does not typically entail the contextualization and extrapolation components, generally only the induction component.

CONCLUSION

This study endorsed a componential view on dispositional reasoning and examined its role in interviewer judgment accuracy. Dispositional reasoning is best thought of as an intelligence, as our results showed it broadly adhered to strict criteria for an intelligence measure. Moreover, it has distinguishable subcomponents that constitute important pieces of the puzzle in understanding what makes the “good judge.” No earlier studies have evaluated the role of dispositional reasoning at this level of granularity. We conclude from our findings that dispositional reasoning components are not “just $g$” measures. Rather, they represent specific intelligences in the social-cognitive domain that allow better use of behavior information in interviews. Compared to poor judges, good judges have better developed abilities to extrapolate and contextualize trait-related information. Although the jury is still out on the role of dispositional reasoning components in other judgment contexts, for example, assessment centers and performance rating contexts, we hope this research will trigger further research on related issues.
REFERENCES


Goldberg, L. R. (2005). A scientific collaboration for the development of advanced measures of personality traits and other individual differences. In L. R. Goldberg (Chair), *The international personality item pool and the future of public-domain personality measures*. Presidential symposium at the sixth annual meeting of the association for research in personality, New Orleans, January.


---

**APPENDIX**

Example Items From the Revised Interpersonal Judgment Inventory

**Trait induction**

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

<table>
<thead>
<tr>
<th>Trait</th>
<th>Emotional Stability</th>
<th>Extraversion</th>
<th>Openness</th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sloppy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Irritable</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Trait extrapolation**

For example, one item depicted “John” as “John’s coworkers all describe him as efficient, thorough, and persistent. MOST likely John also:”. Next, respondents had to choose the best answer from the following 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 fantasizing 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 e-mails to accumulate and frustrate your coworkers in the process.
- B. **You notice that the time has just turned 1 pm (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 nonstop since the morning.