8-2017

A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups

François S. DE KOCK
Erasmus University of Rotterdam

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

Marise Ph. BORN
Erasmus University of Rotterdam

DOI: https://doi.org/10.1111/ijsa.12176

A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups.

De Kock, François S., Institute of Psychology, Erasmus University, Rotterdam, Rotterdam, The Netherlands

Lievens, Filip, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Gent, Belgium
Born, Marise Ph., Institute of Psychology, Erasmus University, Rotterdam, Rotterdam, The Netherlands

Correspondence
François De Kock, School of Management Studies, University of Cape Town, Private Bag X3, Rondebosch, Cape Town 7701, South Africa.
Email: francois.dekock@uct.ac.za

Published in International Journal of Selection & Assessment. Sep 2017, Vol. 25 Issue 3, p240-252.

https://doi.org/10.1111/ijsa.12176

Submitted version

Abstract

Despite the growing interest in dispositional reasoning as a construct and determinant of good raters ('good judges'), its measurement still requires attention. We address two measurement issues in the present study. First, this study tests a hierarchical model as a more parsimonious account for dispositional reasoning than component- or general-factor models that were examined in earlier studies. So, this provides a more comprehensive test of the different measurement models underlying dispositional reasoning data. Second, we assess the measurement invariance of dispositional reasoning measure scores across two different populations of assessors that are often trained and used in workplace assessments, namely psychology students (N = 161) and managers (N = 160). Results 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 and psychology students through measurement invariance analysis showed relatively similar factor structures underlying dispositional reasoning scores across these groups, but metric invariance could be only partially established.

Introduction

The characteristics of the good raters ('good judges') have intrigued researchers and practitioners for a long time (e.g., see Funder, [17]; Taft, [44]; Vernon, [48]). Recent efforts to explain individual differences in judgment accuracy have shown promise for dispositional reasoning as a key determinant of what makes a good judge. Dispositional reasoning can be defined as a rater's complex knowledge of traits, behaviors, and situations' potential to elicit traits into manifest behaviors (Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, [13]). Research (Christiansen et al., [13]; De Kock, Lievens, & Born, [16]) revealed that interviewers' dispositional reasoning was the strongest predictor of accuracy among a set of individual differences that included demographics, personality, and general cognitive ability. Moreover, it showed discriminant validity with personality traits and convergence with measures of cognitive ability (rs =.43 and.68, in the two studies cited, respectively).

Conceptually, dispositional reasoning has three distinguishable components: trait induction is the ability to know how traits manifest themselves in behavior; trait extrapolation is an understanding of how traits and their behavioral manifestations naturally co-vary; and trait contextualization refers to the ability to identify situations that are relevant for expressing traits (De Kock et al., [16]). Importantly, each of these components is not measured through a self-report questionnaire. Instead, Christiansen et al. measured these components via a multiple choice test in which people, for instance, have to assign adjectives to constructs (Big Five) or determine which situation is the best for observing specific trait-related behavior related to constructs such as complexity or sociability (see also examples in Tett & Guterman, [45]).

Despite the growing interest in dispositional reasoning as a construct and determinant of a good judge, its measurement still requires further attention. The measurement drawbacks of earlier studies are twofold. First, although Christiansen et al. ([13]) conceptualized dispositional reasoning as consisting of three components, their measure 'did not permit reliable subscale scores to be computed for the hypothesized domains' (p. 143). To address this issue, De Kock et al. ([16]) revised the original measure to yield reliable subscale scores and found that a three-factor solution fitted the data reasonably well. However, measures of ability in the same conceptual domain often show both 'positive manifold' (Horn & Cattell, [23]) and an hierarchical nature (see Carroll, [9], for a review), where broad factors at a higher stratum affect narrow factors at lower strata. As dispositional reasoning exhibits characteristics of an ability measure (De Kock et al., [16]) it may also potentially have an hierarchical configuration—including a general factor influencing the three specific components. Therefore, this study tests a hierarchical model as a more parsimonious account for the underlying structure of dispositional reasoning scores than component- or general-factor models that were examined in earlier studies. This provides a more comprehensive test of the different measurement models underlying dispositional reasoning data.

A second measurement issue is that prior dispositional reasoning studies used two different populations of judges, namely either psychology students [1] (Christiansen et al., [13]; Powell & Bourdage, [37]; Powell & Goffin, [38]) or managers (De Kock et al., [16]). From a practice perspective, a focus on either of these two populations makes indeed a lot of sense because both groups constitute the typical pools of assessors that are trained and used in workplace assessments (Krause & Thornton, [28]; Lievens, [29]). Evidence also suggests that combining psychologists and managers produces the greatest predictive validity (Gaugler, Rosenthal, Thornton, & Bentson, [19]). However, only when the measurement structure is invariant between these two populations, dispositional reasoning scores can be compared and merged across these groups of assessors. Therefore, it is important to know whether the dispositional reasoning measure works equally well for both populations.

These two unclear measurement features of dispositional reasoning impede progress not only on the aforementioned conceptual issues, but it has also practical implications for the use of the dispositional measure. For example, assessor training interventions may be tailored to target specific components (induction,

extrapolation, or contextualization) if these components are distinguishable. Moreover, lack of measurement invariance (MI) of dispositional reasoning scores across rater populations might require developing different measures for the respective groups (i.e., managers vs. psychologists).

In short, this study aims to contribute to the small albeit growing literature on dispositional reasoning as a key construct by investigating its dimensionality through a more comprehensive set of confirmatory factor analysis models (hierarchical, component-models, and general-factor models). In addition, we examine the invariance of this measure across two samples (psychology students and managers) that are often trained in workplace assessments.

STUDY BACKGROUND

Dispositional reasoning: conceptualization and research

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., [16]). Dispositional reasoning may allow good judges to process behavioral information toward accurate trait inferences. Research (Christiansen et al., [13]; De Kock et al., [16]) showed that interviewers' dispositional reasoning was the strongest predictor of accuracy among a set of individual differences that included demographics, personality, and general cognitive ability. In both these studies, participants watched videotaped segments of individuals responding to employment interview questions and judged the characteristics of the video interviewees. Accuracy was measured by comparing raters' judgments with those of 'true scores,' which were derived from targets' self-reported personality dimensions (Christiansen et al., [13]), or subject matter expert ratings of interviewees' performance (De Kock et al., [16]). Moreover, dispositional reasoning scores showed discriminant validity with personality traits and convergence with measures of cognitive ability (rs =.43 and.68, in the two studies cited, respectively). Finally, in one of these studies (De Kock et al., [16]) dispositional reasoning showed incremental validity ($\Delta R2 = .09$, $p = .004$; small to medium effect size, Cohen's $f2 = .11$) over general mental ability to predict a key accuracy criterion (Borman's Differential Accuracy scores). As such, these findings speak for the practical use of dispositional reasoning measures to screen and select assessors in organizations. Other research investigated whether it is possible to develop assessors' dispositional reasoning through training. Early attempts (Powell & Bourdage, [37]; Powell & Goffin, [38]) to enhance one of the components of dispositional reasoning—so-called behavior-trait knowledge, also known as 'induction' (De Kock et al., [16])—with training, have been unsuccessful, however.

Competing models of dispositional reasoning

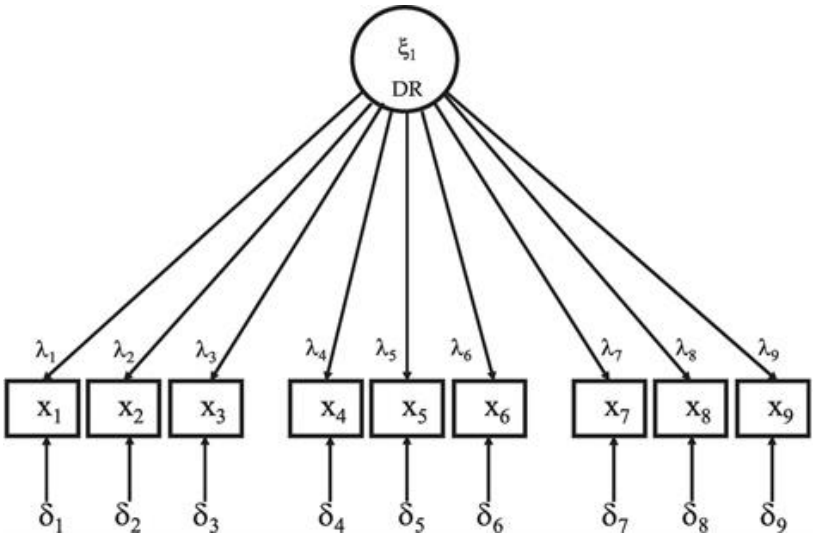
Christiansen et al. ([13]) conceptualized dispositional reasoning as consisting of three components. However, their subscales of the different components were too short to provide reliable subscale scores. So, they assumed a general-factor model. De Kock et al. ([16]) extended the original measure to yield reliable subscale scores and found that a three-factor solution (component-model) fitted the data reasonably well. Apart from testing these models, this study tests for the first time also an hierarchical model as a more parsimonious account for dispositional reasoning than the component- or general-factor models that were examined in earlier studies.

Model 1: General-factor model

In a general-factor model underlying dispositional reasoning scores (see Figure 1), assessors' procedural and declarative knowledge structures that relate to multiple domains—in this case, the areas of knowledge of behaviors, traits, and situations—are encapsulated in a single broad factor. For example, items that measure one component (e.g., trait induction) overlap with items that tap into another (e.g., trait extrapolation), resulting in a broad dispositional reasoning latent variable that causes variance in all items, irrespective of the component that a specific item was designed to measure. Therefore, the model assumes no distinction between separate dispositional reasoning components.

In the broader literature, a well-known example of a general-factor model is Spearman's ([42]) 'g-theory', that is, the view that performance at one type of cognitive task tends to be comparable to performance at other cognitive tasks. General-factor models also exist in other literatures such as general affectivity (Cropanzano, Weiss, Hale, & Reb, [15]).

Figure 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, as demonstration

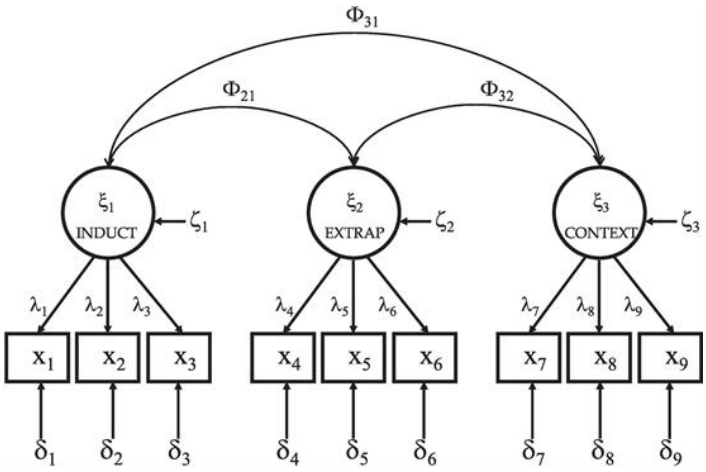


Model 2: Three components (First order)

In a component-model of dispositional reasoning, specific abilities related to understanding traits, behaviors, and situations cluster into three facets. So, in such a model, dispositional reasoning has three distinguishable components: trait induction is the ability to know how traits manifest themselves in behavior; trait extrapolation is an understanding of how traits and their behavioral manifestations naturally co-vary; and trait contextualization refers to the ability to identify situations that are relevant for expressing traits. In a component-model (see Figure 2), items load onto these three separate dimensions, with no cross-loadings allowed.

Componential views of constructs are also encountered in the psychology literature. Examples of componential models can be found for emotional intelligence (Mayer, Caruso, & Salovey, [35]) and for other 'specific' intelligences (Gardner, [18]).

Figure 2: A confirmatory factor analysis of the structure of dispositional reasoning: A three-component (first-order) model (M2). Only nine indicator variables are used in this example, as demonstration

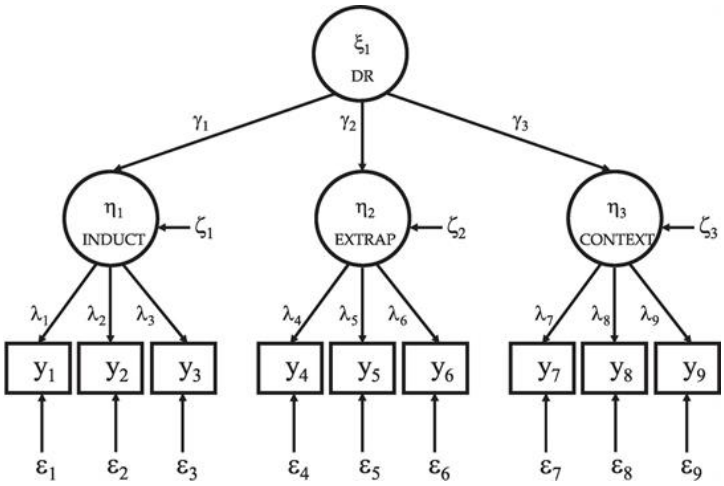


Model 3: Hierarchical model (Second order)

Dispositional reasoning can also be considered a hierarchically ordered construct, with a general factor influencing the three specific components (see Figure 3). An hierarchical structure for dispositional reasoning suggests a broad dispositional reasoning latent construct (i.e., higher-order factor) causing variance in the three specific components (i.e., lower-order factors).

In the broader literature, measures of ability in the same conceptual domain often show an hierarchical nature (see Carroll, [9] , for a review). For instance, in the intelligence literature, the early general (g) versus specific (sn) intelligence debate has given way to a consensus view of the hierarchical nature of abilities where broad factors at a higher stratum affect narrow factors at lower strata.

Figure 3: A confirmatory factor analysis of the structure of dispositional reasoning: Hierarchical (second-order) model (Model M3). Only nine indicator variables are used in this example, as demonstration



Rater groups and MI

As noted above, prior dispositional reasoning studies used two different populations of assessors, namely either psychology students (Christiansen et al., [13]; Powell & Bourdage, [37]; Powell & Goffin, [38]) or managers (De Kock et al., [16]). Both of these groups constitute the pools of assessors that are often trained and used in workplace assessments (Krause & Thornton, [28]; Lievens, [29]). In support of this point, a survey of AC selection and development programs of 144 organizations in 18 countries (Thornton & Krause, [46]) reported that 70% used line managers, whereas external (44%) or internal (22%) psychologists were also a popular choice.

Previous studies also found some rating differences between these two populations. For example, Barr and Hitt ([2]) examined the selection decisions of professional interviewers and students and found significant differences in the number and nature of factors used. In several studies, Lievens ([29] , [30] , [31]) found that psychology students were better able to provide distinct assessment center ratings than managers. Lievens attributed these findings on psychology students' education that had versed them more into the notion of psychological constructs and their behavioral indicators.

Although these prior studies hint that a dispositional reasoning measure might work differently for psychology students and managers, no earlier studies have considered the MI of dispositional reasoning across both of these groups. MI (Millsap, [36]) determines whether 'an assessment instrument is measuring the same constructs in exactly the same way across groups' (Byrne & Stewart, [8] , p. 287). Without invariance between managers and psychology students, between-group comparisons of test scores may be misleading: that is, 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, [11]). Only when the measurement structure is invariant between these two populations, dispositional reasoning scores can be compared across these assessor groups.

METHOD

Participants

Combined sample

For our study, it was important to limit the sample to participants that form part of a broader population of potential assessors. Therefore, a combined sample (N = 321) of managers (49.8%) and psychology students (50.2%) was selected because these are the people who are most likely to be trained as assessors (Krause & Thornton, [28]). The combined sample (54.4% females and 45.6% males) comprised 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. English was the official workplace language of all participants, although the prevalent first languages among these respondents were English (40.8%) and Afrikaans (19%).

Group 1: Psychology students

We recruited 161 students in Industrial-Organizational Psychology from two universities in South Africa. Students were at various levels of academic seniority, although most (59.5%) were postgraduates (i.e., they had finished their Bachelor's degrees and were doing Honors- or Masters-degrees at the time of the study). The rest were Bachelor's students.

Group 2: Managers

Our second group consisted of 160 managerial personnel [2] working in various line and staff functions (e.g., HRM, finance, etc.) within two organizations: a national police training academy and a supervisor training college. All of these respondents were undergoing staff development training when they were assessed.

A comparison of the two samples showed that managers were generally older ($M = 42.3$ years, $SD = 6.7$ years) than psychology students ($M = 22.8$ years, $SD = 3.5$ years), $t(221.02) = 31.142$, $p < .001$. 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%).

Procedure

The data collection was completed in multiple sessions within the respective organizations. After introducing the research as part of assessor training to develop self-insight about their dispositional reasoning, we explained participants' rights and requested their informed consent. Next, participants independently completed the research questionnaire, before they were debriefed and thanked for their participation. Following their study participation, assessors each received an individual feedback report summarizing their performance on the measure.

Measures

Dispositional reasoning

To measure the dispositional reasoning components, we used the Revised Interpersonal Judgment Inventory (R-IJI) (De Kock et al., [16])—a revision of the original IJI (Christiansen et al., [13]). The Revised IJI consisted of 64 items that measure three components. Example items for each subscale may be found in Appendix A.

Induction

The induction 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 ([20]) 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 from an earlier study (Tett & Guterman, [45]) 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, for instance ‘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.

Biographical characteristics

To enable normative comparisons, we also requested respondents’ biographical details.

Statistical analysis

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 when it is posited 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 (i.e., between specialized facets of a broader construct) become the input matrix for the higher-order factor analysis. The HCFA attempts to provide a more parsimonious account for the inter-correlations among lower-order factors (Brown, [4]).

Robust maximum likelihood estimation was employed to estimate all models, unless stated otherwise. We used a number of fit indices to evaluate model fit, including SBχ2(Satorra & Bentler, [41]), CFI, RMSEA (and its 90% confidence intervals), and SRMR. As recommended by Byrne and Stewart ([8]), the following minimum cutoffs were applied to infer acceptable model fit: SBχ2 (Satorra & Bentler, [41]) with $p > .05$; CFI > .95; RMSEA < .08; and SRMR < .08. Our analyses were conducted with Lisrel 9.2 (Jöreskog & Sörbom, [25]).

Data preparation for HCFA

Before we conducted the HCFA, 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 full 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, [34]; Wolf, Harrington, Clark, & Miller, [50]). Therefore, we reduced the number of items in the scales to allow for sufficient power and ensure appropriate model identification—issues that were important for the subsequent hierarchical model analyses. Upon inspection of the issues associated with reducing the number of items in the

scales (see Yang, Nay, & Hoyle, [51]) we decided to create four indicator variables for each first-order latent variable by using parcels of items within each scale as manifest variables, using the procedures outlined by Little, Cunningham, Shahar, and Widaman ([32]). Our parceling strategy is explained in Appendix B. Using parcels in CFA has distinct advantages: Not only do they allow retaining measurement information from many items, but in most conditions, less biased parameter estimates result when parcels are used (Hair, Black, Babin, & Anderson, [21]). However, we acknowledge that combining items into parcels may also artificially enhance the reliability estimates of scores from the measure (Hair et al., [21]).

Model specification

The hierarchical CFA model (see Figure 1) hypothesizes for both managers and psychology students the following (in line with Byrne and Stewart, [8]): (a) a dispositional reasoning structure is best represented by a single higher-order factor of dispositional reasoning and three lower-order factors (trait induction, trait extrapolation, and trait contextualization); (b) each observed variable (i.e., parcel) has 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 is explained by the higher-order factor of dispositional reasoning; (d) measurement error terms are uncorrelated; and (e) factor disturbances are uncorrelated.

Model identification

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, [27]). The hierarchical model (M3) that we hypothesized (see Figure 3) 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 must also be identified in itself. As a solution that specifies a single second-order factor over three first-order factors is just-identified (Brown, [4]), the residuals of induction and extrapolation were constrained to be equal (using a procedure outlined by Byrne, [7]) to achieve identification at the higher-order level of the model.

Latent variable scaling

In addition to adequate model identification, it was necessary to scale the second-order factor of dispositional reasoning in the model because it has no observed measures and must be provided a metric (Brown, [4]). We decided to fix the variance of the second-order dispositional reasoning factor to 1.0 because it left all three direct effects of dispositional reasoning on the first-order factors as free parameters.

Higher-order CFA procedure

After completing the data preparation, we followed the general sequence of HCFA proposed by Brown ([4]), which was to: (a) develop a ‘well-behaved’ first-order CFA solution, in other words, one that fits well and is conceptually valid; (b) examine the magnitude and pattern of correlations among factors in the first-order model; and (c) fit the second-order model, based on conceptual and empirical grounds.

Measurement invariance

Finally, we conducted MI analysis (Millsap, [36]) of the best fitting factor model between managers and psychology student samples. To establish the MI of the first-order models of the factor structure underlying our measure of dispositional reasoning, between managers and psychology students, we followed available guidelines for general MI (e.g., Brown, [4]; Millsap, [36]; Raykov, Marcoulides, & Li, [39]; Vandenberg & Lance, [47]), but also specific guidelines to assess invariance of hierarchical models (e.g., Byrne & Stewart, [8]; Chen, Sousa, & West, [10]; Cheung, [11]). Our testing strategy involved hierarchical steps comparing the fit of a series of more constrained models with less constrained models, relying on the Likelihood Ratio (LR) test (Tabachnick & Fidell, [43]) at each step. The LR test involves a comparison of the $\chi 2$-values of the unconstrained and constrained models and statistically significant increase in $\chi 2$ as a result of constraining a specific set of parameters was used as a criterion for rejecting MI.

RESULTS

Descriptive statistics

Table 1 and Figure 4 portray 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 higher on overall dispositional reasoning than managers (M = .45, SD = .14, N = 160), t(287.8) = −22.2, p < .001, two-tailed. The difference of.31 scale points was substantial (scale range: 0%–100%; d = 2.55, large effect size r =.79, Cohen, 1988) and the 95% confidence interval around the difference between the group means was relatively precise (33.7–28.2). As a possible reason, psychology students' education might verse them more into the notion of psychological constructs and their behavioral indicators (Lievens, [29]).

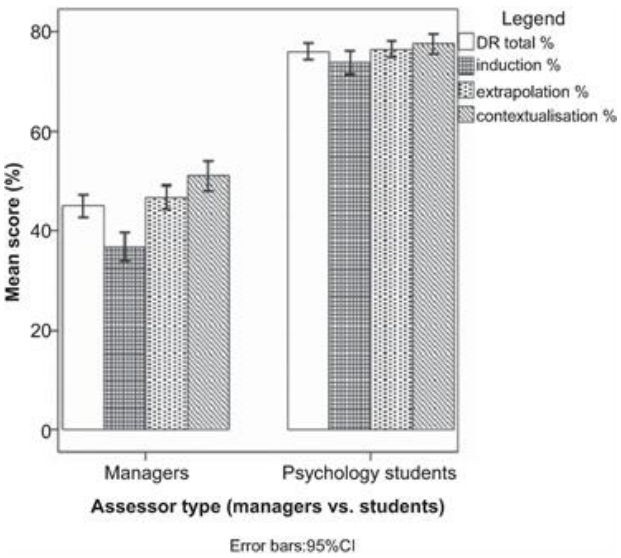
Table 1: Descriptive statistics and intercorrelations for the managers' and psychology-students' samples[a]

Variables	Descriptives	1	2	3	4
	M (*SD*)	Correlations for managers (n = 160)			
1. Induction	.37 (.18)	(α = .74)			
2. Extrapolation	.47 (.16)	.56**	(α = .66)		
3. Contextualization	.51 (.19)	.47**	.44**	(α = .75)	
4. Total DR	.45 (.14)	.83**	.80**	.80**	(α = .85)
		Correlations for psychology-students (n = 161)			
1. Induction	.74 (.16)	(α = .69)			
2. Extrapolation	.77 (.10)	.53**	(α = .44[a])		
3. Contextualization	.77 (.13)	.53**	.36**	(α = .59)	
4. Total DR	.76 (.10)	.87**	.74**	.79**	(α = .79)

Note. Total N = 321. DR = Dispositional reasoning total scores. [a]Lower reliabilities for students are likely due to substantial lower dispersion for this sample. For example, psychology students had lower mean item variance (.14 vs .21) and scale variance (5.46 vs 13.09) for extrapolation, as compared to the manager sample. When the two samples were combined, the alphas were higher: full measure (.93), induction (.86), extrapolation (.82), and contextualization (.82).
*p < .05; **p < .01 (two-tailed).

For the sake of brevity, the mean differences across the two subsamples for component scores are not reported; however, they were all statistically significant, $p < .001$. Table 1 also reports the intercorrelation (uncorrected for unreliability) between the dispositional reasoning component scores for the two subsamples.

Figure 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%5no correct answers and 100%5all items correct



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 2 for the combined sample. Table 3 reports the results separately for managers and psychology students. The general-factor model (M1, see Figure 3) of dispositional reasoning was assessed by a first-order confirmatory factor analysis based on data from the combined sample. The fifteen item parcels serve as indicators of the general dispositional reasoning factor. The general-factor model (M1) was tested and the fit was acceptable, $\chi2(90, N = 321) = 191.50$, $p < .001$, Satorra–Bentler $\chi2\ (90, N = 321) = 180.99$, $p < .001$, 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.

Table 2: Fit indices for factor structure models of dispositional reasoning measure in combined sample[a]

Model	χ^2	S-Bχ^2	*df*	S-Bχ^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)

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); 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: 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)

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); 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.
**$p < .01$.

Three-component model (M2)

Next, we evaluated a three-component factor model, with trait induction, trait extrapolation, and trait induction as separate components (see Figure 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, $\chi2(87, N = 321) = 117.60$, $p = .016$, Satorra–Bentler $\chi2$ (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) and consistent with our expectation, all factors were significantly interrelated (range of zs = 6.76–10.48). Factor intercorrelations (among the various subdimensions of the dispositional reasoning components, M2) were generally large ($.84 < \varphi < .95$). So, 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 (second-order) factor model of dispositional reasoning—this model proposes a general component, influencing the three specific components of induction, extrapolation, and contextualization—was tested and support was found because the model showed good fit, $\chi2$ (87, $N = 321) = 117.60$, $p = .016$, Satorra–Bentler $\chi2$ (87, $N = 321) = 113.29$, $p < .05$, Robust CFI = .99, TLI = .98, RMSEA = .03, 90% CI: [0.015; 0.048]. 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, [4]). Looking at our results (the completely standardized estimates from the solution), each of the first-order factors loads strongly on the second-order dispositional reasoning factor: induction ($\gamma = .98$) and extrapolation ($\gamma = .96$) loaded more strongly than contextualization ($\gamma = .88$). As such, dispositional reasoning as a higher-order factor accounted for substantial proportions of variance in the individual components: induction 96% (1 − .04), extrapolation 91.5% (1 − .085), and contextualization 77.1% (1 − .229).

Table 4: Tests of invariance of dispositional reasoning in managers and psychology-students

Model	χ^2	*df*	χ^2_{diff}	Δdf	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 $\leq .05$); SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker–Lewis index; NFI = normed fit index; PFI = parsimonious fit index.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Model comparison

We compared the baseline model (general-factor model, M1) with the comparison models. A chi-square difference test (Bryant & Satorra, [5], [6]) indicated that the nested model (M2) showed significantly poorer fit compared to the baseline (M1) general-factor model, Satorra–Bentler χ2diff(3, N = 321) = 45.033, p < .001. Therefore, the three-component model of dispositional reasoning fitted significantly better than a general-factor model.

We also compared a model in which the correlations between dispositional reasoning were freely estimated; and a nested comparison model in which the correlations were constrained to be unity.[3] We used the raw data as input for the analysis and found relatively poor fit of the nested model, χ2 (90, N = 321) = 189.843, p < .01, RMSEA = .06 (90% CI:.04;.07). A chi-square difference test indicated that the nested model (specifying the relationship between dispositional reasoning facets as perfectly correlated) showed significantly poorer fit, compared to the baseline model, χ2diff(3, N = 321) = 72.303, p < .001. Therefore, the evidence suggests that the components are empirically distinct from one another.

The goodness-of-fit of the hierarchical model (M3) is the same as the three-component first-order model (M2) in which factors are allowed to co-vary freely. According to Brown ([4]), this is because a solution that specifies a single second-order factor over three first-order factors is just-identified (Brown, [4]) and, therefore, it is not appropriate to statistically compare M3 with M2. Only when the higher-order model is over-identified, can the nested χ2 be used to determine whether the specification in M3 produces a significant degradation in fit relative to the first-order solution.

However, apart from the higher-order solution not resulting in a decrease in model fit, it also provides a more parsimonious account for the correlations among the first-order factors. So, a higher-order model with dispositional reasoning as a general factor in turn influencing induction, extrapolation, and contextualization,

explains variance in test scores better than a general-factor model. 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.

Measurement invariance

To compare the factor structure of dispositional reasoning between managers and psychology students, we conducted MI analyses (see Table 4). In line with the suggestions of Brown ([4]), a baseline model was first established in each group, followed by tests of equivalence across groups at each of several increasingly stringent levels of invariance.

Table 4: Tests of invariance of dispositional reasoning in managers and psychology-students

Model	χ^2	*df*	χ^2_{diff}	Δdf	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 $\leq .05$); SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker–Lewis index; NFI = normed fit index; PFI = parsimonious fit index.
* $p < .05$. ** $p < .01$. *** $p < .001$.

First-order (M2) invariance

Preliminary analyses

It is preferable to conduct multiple-groups CFA with relatively balanced sample sizes, as was the case in the present study (managers: N = 160; students: N = 161). The Robust ML estimator was used in estimation of all models and, therefore, all analyses are based on the Satorra–Bentler scaled statistic (SBχ2; Satorra & Bentler, [41]). To evaluate all models we relied on SBχ2, as well as on CFI, the root mean square error of approximation (RMSEA), and SRMR, in line with the recommendations of Byrne and Stewart ([8]). The evaluation criteria we apply for each fit index are outlined in Byrne and Stewart ([8]): Values that adhere to the following cutoffs indicate significant reduction in fit when comparing two nested models: (a) if corrected ΔSBχ2/Δdf shows statistical significance; (b) ΔCFI >.01; and (c) the root mean square error of approximation (RMSEA) > .08. The first item parcel within each subscale was used as a marker indicator to define the metric of the latent variable.

Testing for baseline models

As the estimation of baseline models involves no between-group constraints, the data were analyzed 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 4, overall fit statistics for the three-factor solution are consistent with good model fit in both managers and psychology students. 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. 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 analyzed simultaneously in a file combining data for both groups to obtain estimates. Given that the baseline models are now fitted simultaneously in a multigroup evaluation, the criterion for configural invariance is that goodness-of-fit should indicate a well-fitting model. So, we conducted the simultaneous analysis of equal form. As shown in Table 4, this solution provides an acceptable fit to the data. This solution (i.e., configural model) serves as the baseline model for subsequent tests of MI 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 goodness-of-fit is adequate and if there is minimal degradation in fit from the configural model. The analysis evaluates whether factor loadings (unstandardized) of the dispositional reasoning 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, $\chi 2\text{diff}(12) = 39.60$, $p < .001$. As this value is statistically significant, it follows that the constraints of equal factor loadings in the restricted model do not hold (Byrne & Stewart, [8]), suggesting that the two models are not equivalent across the manager and psychology student groups. As 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, [4]). This means that a unit change in the underlying latent variable is not associated with statistically equivalent change in the observed measures (item parcels [4]) in both groups.

A closer look at the factor loadings revealed that the mean factor loading for managers was.57 (SD = .12) and for psychology students.48 (SD = .13). Of these, 80% were invariant (within 1.96 SD). The three loadings that were not invariant (> 1.96 SD) were equally spread across components. 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 4, however. Overall, from these results we conclude that only partial MI (Byrne, Shavelson, & Muthén, 1989) between managers and psychology students exists for our measure.

DISCUSSION

This study contributes to the small albeit growing literature on dispositional reasoning as a key construct by investigating its dimensionality through a more comprehensive set of confirmatory factor analysis models (hierarchical, component models, and general-factor models). In addition, we test the invariance of this measure across two samples (psychology students and managers) that are often trained in workplace assessments.

Results supported an hierarchically configured model for dispositional reasoning, with a general factor at a higher stratum driving three specific facets (trait induction, trait extrapolation, and trait contextualization) at a lower stratum. Moreover, the hierarchical model showed acceptable fit within both our psychology students and manager samples. So, we found evidence for a relatively common factor structure for dispositional reasoning in both samples. However, we also observed some lack of metric invariance for the dispositional reasoning measure between managers and students, in other words, the factor loadings were overall not equivalent between managers and students.

Follow-up analyses showed that only three (20%) observed variables showed substantial differences in factor loadings between the two groups. It is possible that our invariance tests were conservative in the sense that a minority of observed variables, with large (> 1.96 SD) differences in factor loadings between groups, led to failure of the overall test for metric invariance. We also considered the location of the noninvariant items within the component measures: The 'offending' observed variables were not located within particular dispositional reasoning components, but rather, they were evenly spread. Moreover, the item content of the non-invariant observed variables did not reveal any clear pattern that may have provided a theoretical explanation for the differences in factor loadings between managers in psychology students. Regarding the overall strength of factor loadings between the two groups, the mean factor loading (across items) for managers (.57) was higher than psychology students (.48), which may have contributed to the failure in the invariance test. The overall lower factor loading of the student group may have resulted, in part, from the relatively lower dispersion in their item responses, that is, students showed lower variability than managers and they did better overall on the measure.

Descriptive statistics showed that psychology students outperformed managers on the measure of dispositional reasoning by a substantial margin. As noted, earlier studies revealed also other differences between managers and psychology students. For example, prior studies reported that psychology students were better able to provide distinct assessment center ratings than managers (Lievens, [29], [30], [31]) and differed from managers in the number and nature of factors they used for selection decisions (Barr & Hitt, [2]). One interpretation is that—as compared to managers—psychology students may have better developed schemas that relate to understanding traits, behaviors, and situations, by virtue of their education and professional training. However, it is important to qualify these explanations because metric equivalence is required to make meaningful between group comparisons of the respective scores. Without metric equivalence, mean differences in scores between these groups cannot be unambiguously interpreted, because it is unclear whether score differences are due to actual differences in this ability (i.e., the schema-based explanation mentioned above), or to different psychometric responses to the scale items (Cheung & Rensvold, [12]).

This study has several limitations. First, by grouping assessors into two relatively coarse categories (managers vs. psychology students) it may obscure other important individual differences within these groups, such as gender and ethnicity. More research is needed to see how stable are the reported factor solutions for dispositional reasoning between gender and ethnic groups. Second, the modest sample sizes that we used prohibited fitting our models using item-level data. Given the potential limitations of item parceling as a strategy (Little et al., [32]) we also fitted the measurement models first at the item level in the combined sample. In addition, we tested the effect of different parceling strategies on the study's final results—the choice of parceling strategy did not change the substantive conclusions. Third, we did not include psychologists in our study, although psychologists are also an

important group of assessors in practice (Krause & Thornton, [28]). Future studies should investigate the measurement properties of our dispositional reasoning measure in a sample of psychologists.

In terms of future research, we see the following avenues. First, studies should consider the measurement of dispositional reasoning across different cultures. Our measure is based on the Big Five personality framework. Although this framework is relatively universal, personality traits may be expressed in unique ways across cultures (Church, [14]; Heine & Buchtel, [22]). Moreover, people from different cultures may have idiosyncratic interpretations of the same observed behavior and how it clusters into constructs (Willmann, Feldt, & Amelang, [49]). As such, cultural groups may score differently on a common set of items that tap into knowledge and understanding of trait concepts. They may have different psychometric responses to the scale items (Cheung & Rensvold, [12]). So, we recommend that future studies consider MI and mean differences between different cultural groups.

Another issue for future studies is to further evaluate the discriminant validity of our dispositional reasoning measure, to show that it is distinct from general mental ability and other abilities (spatial, analytical, problem-solving, etc.) and personality (attention to detail, empathy, emotional intelligence, etc.) that are often used in 'good judge' studies.

Finally, a fruitful avenue is to consider whether or not dispositional reasoning is independent of the trait or content being assessed. Dispositional reasoning may be understood broadly as the ability to reason about traits and dispositions. Our measure (as with the measure of Christiansen et al., [13]) was 'cast in the mold' of the Big Five personality framework. This typology was a good place to start because it is an overarching framework that is generally accepted. However, in principle we could develop a test that measures people's knowledge about any dispositions, just like with tests of general mental ability different stimulus material can be used in different sets of items. Therefore, measures can be developed also for other referent constructs (e.g., interview dimensions, see Huffcutt, Conway, Roth, & Stone, [24]).

Our findings suggest some implications for practice. As noted, measures of dispositional reasoning may be useful for both groups because they represent the pools of assessors that are often trained in workplace assessments (Krause & Thornton, [28]; Lievens, [29]). In our analyses, an hierarchical model with three components showed the best fit, suggesting that organizations may develop assessor training interventions to target specific components (induction, extrapolation, or contextualization) and they might report both an overall dispositional reasoning score, as well as subscores for the three components. Moreover, lack of MI suggests that some adjustments to the dispositional reasoning measure might be needed according to the respective group (i.e., managers vs. psychology students).

APPENDIX A: EXAMPLE ITEMS FROM THE DISPOSITIONAL REASONING TEST

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	Conscientiousness
Sloppy					X
Irritable	X				

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 organization?’. Then, respondents had to select the most appropriate answer from three options (correct answer in bold):

A. You are busy with a task and people continuously interrupt you
B. On your way home you drive past a broken down vehicle
C. Over the last 2 years, you have been employed at a job that entails working by yourself. Your boss offers you a chance to do essentially the same thing, but in a group of coworkers

APPENDIX B: PARCELING STRATEGY: DIMENSIONALITY CONSIDERATIONS

An appropriate parceling strategy should be identified given the dimensionality of the factor structure underlying a set of item scores. Exploratory factor analysis of our item-level data (using Principal Axis Factoring, with Oblimin rotation, considered appropriate for our data, as suggested by Tabachnick & Fidell, [43]) 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 artifacts. 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, [33]). As a result, techniques that are based on correlations, such as factor analysis, may cause artifacts in the form of spurious 'difficulty factors' with little if any psychological meaning (Bernstein & Teng, [3]; Reise, Waller, & Comrey, [40]). 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, [1]) argue that parceling should be reserved for conditions of unidimensionality, Little and colleagues ([32]) suggest two specific strategies for parceling 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, [26]). Finally, a random item assignment strategy may be used. We decided to utilize random item assignment as a parceling 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 parceling strategies—the choice of parceling strategy had no substantive effect on the final results.

Footnotes

1 Psychology students represent an important group of assessors in our study, given that they are normally trained as psychologist assessors.

2 Some manager respondents (n = 146) were also included in another study investigating the criterion-related validity of dispositional reasoning scores (De Kock et al., 16).

3 We thank an anonymous reviewer for this suggestion.

4 As pointed out by an anonymous reviewer, the fundamental meaning of our invariance tests for factor loadings would have been clearer if we had used individual test items, rather than item parcels.

REFERENCES

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.

Barr, S. H., & Hitt, M. A. (1986). A comparison of selection decision models in manager versus student samples. Personnel Psychology, 39, 599-617. https://doi.org/10.1111/j.1744-6570.1986.tb00955.x

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. https://doi.org/10.1037/0033-2909.105.3.467

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

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

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

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

8 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. https://doi.org/10.1207/s15328007sem1302_7

9 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.

10 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. https://doi.org/10.1207/s15328007sem1203_7

11 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. https://doi.org/10.1177/1094428106298973

12 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. https://doi.org/10.1207/S15328007SEM0902_5

13 Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. Human Performance, 18, 123-149. https://doi.org/10.1207/s15327043hup1802_2

14 Church, A. T. (2000). Culture and personality: Toward an integrated cultural trait psychology. Journal of Personality, 68, 651-703. https://doi.org/10.1111/1467-6494.00112

15 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. https://doi.org/10.1016/s0149-2063_03_00081-3

16 De Kock, F. S., Lievens, F., & Born, M. P. (2015). An in-depth look at dispositional reasoning and interviewer accuracy. Human Performance, 28, 1-23. https://doi.org/10.1080/08959285.2015.1021046

17 Funder, D. C. (2012). Accurate personality judgment. Current Directions in Psychological Science, 21, 177-182. https://doi.org/10.1177/0963721412445309

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

19 Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III., & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72, 493-511.

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

21 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.

22 Heine, S. J., & Buchtel, E. E. (2009). Personality: The universal and the culturally specific. Annual Review of Psychology, 60, 369-394. https://doi.org/10.1146/annurev.psych.60.110707.163655

23 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. https://doi.org/10.1037/h0023816

24 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. https://doi.org/10.1037/0021-9010.86.5.897

25 Jöreskog, K., & Sörbom, D. (2015). LISREL (Version 9.2). Skokie, IL: Scientific Software International, Inc.

26 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. https://doi.org/10.1177/0013164494054003022

27 Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). New York: Guilford Press.

28 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. https://doi.org/10.1111/j.1464-0597.2008.00371.x

29 Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. Journal of Applied Psychology, 86, 255-264. https://doi.org/10.1037/0021-9010.86.2.255

30 Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. Journal of Organizational Behavior, 22, 203-221. https://doi.org/10.1002/job.65

31 Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. Journal of Applied Psychology, 87, 675-686. https://doi.org/10.1037/0021-9010.87.4.675

32 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. https://doi.org/10.1207/s15328007sem0902_1

33 Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

34 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. https://doi.org/10.1037/1082-989x.1.2.130

35 Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. Intelligence, 27, 267-298. https://doi.org/10.1016/S0160-2896(99)00016-1

36 Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York, NY: Routledge.

37 Powell, D. M., & Bourdage, J. S. (2016). The detection of personality traits in employment interviews: Can "good judges" be trained? Personality and Individual Differences, 94, 194-199. https://doi.org/10.1016/j.paid.2016.01.009

38 Powell, D. M., & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. Human Performance, 22, 450-465. https://doi.org/10.1080/08959280903248450

39 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. https://doi.org/10.1177/0013164412441607

40 Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. Psychological Assessment, 12, 287-297. https://doi.org/10.1037/1040-3590.12.3.287

41 Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In Proceedings of the American Statistical Association (Vol. 1, pp. 308-313). Alexandria, VA: American Statistical Association.

42 Spearman, C. (1904). "General Intelligence," objectively determined and measured. American Journal of Psychology, 15, 201-292. https://doi.org/10.2307/1412107

43 Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Boston, MA: Allyn & Bacon.

44 Taft, R. (1955). The ability to judge people. Psychological Bulletin, 52, 1-23. https://doi.org/10.1037/h0044999

45 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. https://doi.org/10.1006/jrpe.2000.2292

46 Thornton, G. C., & Krause, D. E. (2009). Selection versus development assessment centers: An international survey of design, execution, and evaluation. The International Journal of Human Resource Management, 20, 478-498. https://doi.org/10.1080/09585190802673536

47 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. https://doi.org/10.1177/109442810031002

48 Vernon, P. E. (1933). Some characteristics of the good judge of personality. Journal of Social Psychology, 4, 42-57. https://doi.org/10.1080/00224545.1933.9921556

49 Willmann, E., Feldt, K., & Amelang, M. (1997). Prototypical behaviour patterns of social intelligence: An intercultural comparison between Chinese and German subjects. International Journal of Psychology, 32, 329-346. https://doi.org/10.1080/002075997400692

50 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, 76, 913-934. https://doi.org/10.1177/0013164413495237

51 Yang, C., 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. https://doi.org/10.1177/0146621609338592