Journal of Applied Psychology
2017, Vol. 102, No. 1, 43–66

© 2016 American Psychological Association
0021-9010/17/$12.00 http://dx.doi.org/10.1037/apl0000160

The Effects of Predictor Method Factors on Selection Outcomes: A Modular Approach to Personnel Selection Procedures

Filip Lievens
Ghent University

Paul R. Sackett
University of Minnesota-Twin Cities

Past reviews and meta-analyses typically conceptualized and examined selection procedures as holistic entities. We draw on the product design literature to propose a modular approach as a complementary perspective to conceptualizing selection procedures. A modular approach means that a product is broken down into its key underlying components. Therefore, we start by presenting a modular framework that identifies the important measurement components of selection procedures. Next, we adopt this modular lens for reviewing the available evidence regarding each of these components in terms of affecting validity, subgroup differences, and applicant perceptions, as well as for identifying new research directions. As a complement to the historical focus on holistic selection procedures, we posit that the theoretical contributions of a modular approach include improved insight into the isolated workings of the different components underlying selection procedures and greater theoretical connectivity among different selection procedures and their literatures. We also outline how organizations can put a modular approach into operation to increase the variety in selection procedures and to enhance the flexibility in designing them. Overall, we believe that a modular perspective on selection procedures will provide the impetus for programmatic and theory-driven research on the different measurement components of selection procedures.

Keywords: personnel selection, assessment, predictor method factors, validity, subgroup differences

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The most recent treatment of personnel selection in the *Annual Review of Psychology* was entitled "A Century of Selection" (Ryan & Ployhart, 2014). This title was aptly chosen because there are few domains in industrial and organizational psychology that have generated such a consistent interest among academicians and practitioners. Traditionally, the emphasis in the selection domain has been on selection procedures as a whole. This focus on predictor methods as holistic entities is understandable because this is how these procedures are used in operational selection practice.

This article proposes a complementary approach to conceptualizing selection procedures and reviewing their effectiveness. As will be argued, a modular approach that breaks down selection procedures into their basic underlying measurement components can further advance selection procedure theory, research, and design. Thus, unlike prior narrative reviews and meta-analyses, we do *not* aim to provide a review of selection procedures as a whole. Instead, we review the effects of key measurement components that make up these selection procedures. We examine the effects of these measurement components on construct saturation, criterion-related validity, subgroup differences, and applicant perceptions.

Modularity: Definition, Characteristics, and Benefits

Historically, in product design, two schools of thought can be distinguished (Baldwin & Clark, 2000). One view considers a product as it is, namely as an all-in-one package. The other product design stream proposes a modular approach by breaking a product down into smaller key components (aka "building blocks"). As a general systems concept, modularity refers to the extent to which a system's components can be separated and recombined (Baldwin & Clark, 2000; Christensen, 2001; Gershenson, Prasad, & Zhang, 2003; Schilling, 2000). Popular examples of modular systems are computers, buildings, and cars. For instance, when purchasing a computer one can "mix and match" various components, such as the processor or hard drive. Within each of these components, one can further choose the processor's speed or the hard drive's size. In a similar vein, a selection procedure can be regarded as being composed of a fixed set of smaller relatively independent components that fit together. For example, one might break down a traditional personality inventory into smaller components such as information source (self vs. others), degree of contextualization (generic vs. contextualized), and response format (close-ended vs. open-ended). Depending on the choices made per component, different measures of personality are constructed.

This article was published Online First September 12, 2016.

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University; Paul R. Sackett, Department of Psychology, University of Minnesota-Twin Cities.

We thank Philip Bobko, Bernd Carette, Brian Connelly, Fritz Drasgow, Huy Le, Dan Putka, Ann Marie Ryan, Thomas Rockstuhl, Deborah Rupp, and Neal Schmitt for their valuable comments on a previous version of this paper. Initial ideas for this paper originated from an inspiring morning run together around Lake Calhoun and Lake of the Isles in Minneapolis in September 2012.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@ugent.be

Modularity is often adopted when products have become established on the market and have evolved further in their life cycle (Christensen, 2001; Schilling, 2000). Given the long history and impressive body of research on selection procedures as a whole, we therefore believe a modular approach is timely and has much to offer to selection procedure theory, research, and design. In particular, on the basis of the product design literature (Baldwin & Clark, 2000; Christensen, 2001; Ulrich & Eppinger, 2004; see also Sanchez & Mahoney, 1996; Schilling, 2000), we posit that a modular approach to selection procedures has the following conceptual and practical benefits.

First, a modular approach allows breaking down a large and complex system into smaller more manageable parts. Whereas the functioning of the system as a whole remains typically a black box, a modular approach enables gaining better insight into the workings of the different components. Applied to selection procedures, this means a modular approach might illuminate which components of the procedures contribute to, for instance, more valid predictions, smaller subgroup differences, or favorable applicant perceptions (even if they are designed/intended to assess the same constructs). So, by going beyond selection procedures as holistic entities we can shed light on a lot of "why's" and "when's" in our knowledge about the effectiveness of selection procedures (Bobko & Roth, 2013; Outtz, 1998).

As a second conceptual advantage, a modular approach to selection procedures promotes identifying and exploiting communalities among selection procedures. That is, it may show that the same components underlie superficially different selection procedures and that they produce similar effects across them. In turn, knowledge about a specific component's effects might then be fruitfully used across various selection procedures. So, a modular lens might spur theoretical connectivity and integrative knowledge across selection tools by uncovering deeper level communalities among these selection tools and their literatures.

Third, a modular approach creates a window of opportunity to set up experiments in which one or two components are modified (while holding others constant). Such experimentation with different configurations may serve as a catalyst for innovation and for improving existing selection procedures.

A Modular Conceptualization of Selection Procedures

Applying a modular approach requires identifying the key components of selection procedures.[1] We start with Arthur and Villado's (2008) distinction between predictor constructs and predictor methods because they constitute major building blocks of selection procedures. Predictor constructs denote the psychological attributes captured by a selection procedure. On the predictor construct side, various frameworks have been developed to further subdivide the constructs. For instance, taxonomies for cognitive ability and personality exist, as do classifications for the constructs targeted by interviews (Huffcutt, Conway, Roth, & Stone, 2001), situational judgment tests (SJTs; Christian, Edwards, & Bradley, 2010), and assessment centers (ACs; Arthur, Day, McNelly, & Edens, 2003).

Predictor methods denote the specific techniques by which construct-relevant information is elicited and collected (Arthur & Villado, 2008). Examples are paper-and-pencil tests, interviews, or simulations. Thus, while predictor constructs address *what* is measured, predictor methods address *how* information about candidates is collected. Just like the subdivisions in predictor constructs, it is possible to break down predictor methods into smaller components, which we call "predictor method factors." For instance, Arthur and Villado (2008) mentioned stimulus format, response format, and scoring approach as three such predictor method factors (p. 440).

Predictor method factors can be defined as key underlying dimensions on which predictor methods vary. Or to put it differently, a predictor method reflects an assemblage of predictor method factors. Because predictor method factors are aspects of test design, a first characteristic is that they reflect features under the control of test designers. Consistent with a modular approach, another characteristic is that they can be seen as relatively independent features (although some factors are more likely to co-occur or in some cases are also almost certain not to co-occur). As a third characteristic, predictor method factors cut across different selection procedures. For instance, in both an oral presentation and an interview, candidates provide oral responses, which exemplifies the role of response format across selection tools.

Contrary to the predictor construct taxonomies mentioned previously, conceptual progress with regard to breaking down predictor methods into smaller components (predictor method factors) has been slow. Therefore, we followed a three-step process to identify a relevant set of predictor method factors. First, we reviewed prior frameworks of predictor method factors. As shown in Table 1, Vernon (1962) was the first to list critical underlying method factors of cognitive ability tests. Cattell and Warburton (1967) undertook a similar effort for personality inventories. More recent frameworks sought to determine the underlying measurement components of ACs (Thornton, 1992), interviews (Huffcutt & Arthur, 1994), computer-based tests (Parshall, Davey, & Pashley, 2000), and simulations (Le, 2013). A limitation of these previous frameworks is that they were confined to one selection tool and included a limited set of method factors. They did provide a good start for us to delineate a set of predictor method factors that are relevant across many selection procedures.

In a second step, we used various criteria for limiting the number of predictor method factors. Consistent with our definition of predictor method factors, we included only factors under the assessment designer's direct control. So, we excluded methodological aspects of how a predictor is studied (e.g., range restriction, research design). Although important, these are not part of actual selection tool design.[2] Finally, we excluded technical aspects (e.g., input devices such as a keyboard) or more fine-grained features (e.g., 5- vs. 7-point rating scales).

These two steps produced a preliminary list of predictor method factors. In a final step, eight well-known authorities in personnel selection (four were past Society for Industrial and Organizational

[1] We restrict ourselves to selection procedures where one is directly evaluating candidates' responses, rather than an indirect inference from some aspect of their behavior (e.g., reaction time [RT] in the case of an Implicit Association Test; see Uhlmann et al., 2012). For the same reason, we also do not consider psychophysiological measures (e.g., galvanic skin response in interviews) to be part of our domain of selection procedures.

[2] For the same reason, we excluded potential effects (e.g., fakability, test motivation) of method factors. Though such effects are important (for producing intended/unintended variance), they are not method factors themselves.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Overview of Prior Frameworks of Predictor Method Factors

Label used in this article	Vernon (1962): Cognitive ability tests	Cattell & Warburton (1967): Personality and motivation inventories	Thornton (1992): Assessment center exercises	Huffcutt & Arthur (1994): Interviews	Parshall et al. (2000): Computer-based tests	Le (2013): Simulations
Stimulus format	Presentation type Form of test material	Test item characteristics	Standardization of stimulus material		Item format Media inclusion	
Contextualization					Fidelity	Authenticity
Stimulus presentation consistency			Standardization of stimulus material	Question standardization	Interactivity	Stimuli flexibility
Instructions		Instruction system				
Response format	Response type Speediness		Structure of response mode		Response action	
Response evaluation consistency	Difficulty level	Scoring modality		Scoring standardization	Scoring algorithm Complexity	Scoring Evaluation focus

Psychology [SIOP] presidents) commented on our list of predictor method factors. Resulting from this process, we identified seven predictor method factors: (1) stimulus format, (2) contextualization, (3) stimulus presentation consistency, (4) response format, (5) response evaluation consistency, (6) information source, and (7) instructions (see Table 2 for definitions). We do not assert that this set of method factors is exhaustive; our goal was to identify a parsimonious set of broad components that constitute critical sources of variation across predictor methods. We are open to the possibility that in the future evidence might emerge for other components.

A Modular Review of Selection Procedure Effectiveness

In a modular approach, one aims to gain knowledge about each major underlying component and its effects. Therefore, we review personnel selection research in terms of the seven predictor method factors and how they affect construct saturation, criterion-related validity, subgroup differences, and applicant perceptions.[3] Although most of the previously mentioned criteria are well known, construct saturation deserves some more explanation. Generally, scores on a selection procedure contain intended variance, unintended variance, and error variance. The degree to which total score variance in a measure reflects intended construct variance is also referred to as construct saturation (see Lubinski & Dawis, 1992; Roth, Bobko, McFarland, & Buster, 2008). For example, if the choice of particular method factors adds unwanted cognitive load to a measure designed as noncognitive, construct saturation is reduced. Because construct saturation can affect validity and subgroup differences, we view it as a mediator of the relationship between method factors and these outcomes.

To gain knowledge about how each method factor affects these criteria, we relied on two types of studies. One type consisted of primary studies that conducted a comparative evaluation of predictor method factor choices. In the prototypical primary study included in our review, one predictor method factor (e.g., stimulus format; Chan & Schmitt, 1997) was manipulated, with other aspects (i.e., test content, other method factors) being held constant. Second, we relied on the results of moderator studies in meta-analyses. For instance, in the most recent meta-analysis on SJTs (Christian et al., 2010), a moderator analysis examined the effect of stimulus format (textual vs. audiovisual) on criterion-related validity. Such meta-analytic evidence has the advantage of being more cumulative. Yet, this also comes with a price because other potentially important factors were often not controlled for.

Stimulus Format

Definition. We define stimulus format as the modality by which the test stimuli (e.g., information, questions, prompts) are

[3] Where relevant, we also report on the effects on reliability, though we view this as an intermediate outcome that will subsequently affect the primary outcomes of criterion-related validity and subgroup differences (i.e., increasing reliability increases both of these outcomes). Similarly, we discuss construct equivalence, in the context of equivalence between alternate forms, as another intermediate outcome.

Table 2
Predictor Method Factors, Their Definitions, and Categories

Predictor method factor	Definition	Predictor method factor category/choice
Stimulus format	Modality by which test stimuli (information, questions, prompts) are presented to test-takers	- Textual stimuli - Pictorial stimuli - Auditory stimuli - Dynamic audiovisual stimuli - Videoconference/remote interactive stimuli - Face-to-face interactive stimuli
Contextualization	The extent to which a detailed context is provided to test-takers	- Decontextualized - Low contextualization - Medium contextualization - High contextualization
Stimulus presentation consistency	Level of standardization adopted in presenting test stimuli to test-takers	- Free stimuli - Adaptive stimuli - Fixed stimuli
Response format	Modality by which test-takers are required to respond to test stimuli	- Close-ended - Textual constructed - Pictorial constructed - Audio constructed - Audiovisual constructed - Videoconference/remote interaction - Face-to-face interaction
Response evaluation consistency	Level of standardization adopted in terms of evaluating test-takers' responses	- Unconstrained judgment - Calibrated judgment - Automated scoring
Information source	Individual responding to the test stimuli	- Behavior exhibited (or choices made) by the candidate in the assessment context - Self-reports by the candidate about events beyond the assessment context - Reports by others about events outside the assessment context
Instructions	The extent to which directions are made explicit to test-takers about which perspective they should take to respond to the test stimuli	- General instructions - Specific instructions

presented to test-takers. As shown in Table 1, this first predictor method factor was often included in earlier frameworks. Alternate labels used were "presentation type" or "item format."

Prior research. In prior studies, six stimulus format categories can be generally distinguished.[4] The first category consists of *textual stimuli*. Examples are written verbal reasoning items, memos, letters, or e-mail messages (as part of an in-basket exercise). The second category comprises of *pictorial stimuli*. Examples of such stimuli are charts in an in-basket exercise, or facial pictures in an emotional intelligence task. The third category consists of the presentation of *auditory* stimuli. Examples are telephone interview questions, voice overs, voice messages in a Personal Computer (PC) simulation, foreign language samples, or samples of music for testing music listening skills. The fourth stimulus format category consists of formats that present *dynamic audiovisual stimuli*. Here, finer distinctions can be made by differentiating between video scenes, 2D animation (cartoon), 3D animation, or avatar-based formats. Finally, the fifth and sixth stimulus format categories refer to *videoconference* (aka remote, online) and *face-to-face interactions*, respectively. Examples of these categories are videoconference or live stimuli exhibited by interviewers, role-players, or other candidates.

As shown in Table 3, one piece of knowledge about this predictor method factor comes from experiments that modified stimulus formats in the context of the assessment of interpersonal situational judgment. Cognitive load theory (Sweller, 1988) served as the dominant underlying conceptual framework for predicting differences. For example, in an interpersonal SJT, textual stimuli produced scores with a higher unintended cognitive saturation (e.g., because of wording/sentence complexity) than audiovisual stimuli (Lievens & Sackett, 2006). Audiovisual items of interpersonal SJTs had also higher validity (Christian et al., 2010; Lievens & Sackett, 2006), smaller Black–White subgroup differences (Chan & Schmitt, 1997), and more favorable applicant perceptions (Kanning, Grewe, Hollenberg, & Hadouch, 2006; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000) than textual items.

Other knowledge about the effects of stimulus format comes from research comparing videoconference to live face-to-face interactions in employment interviews. This research relied on interpersonal communication and social bandwidth theories (Potosky, 2008; Van Iddekinge, Raymark, Roth, & Payne, 2006) and posited that face-to-face interactions involve more cues and more social presence than videoconference interactions. Research showed that in videoconference interviews candidate ratings and applicant reactions are therefore lower (Chapman, Uggerslev, &

[4] These categories represent broad categories and finer distinctions are possible. One such distinction pertains to the medium for conveying the stimuli (Parshall et al., 2000; Potosky, 2008). For instance, textual and pictorial stimuli might be presented via a paper-and-pencil or computerized medium (PC, tablet, smartphone, etc.).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Overview of Prior Research Findings and Relevant Theories Related to Predictor Method Factors

Predictor method factor	Relevant theories	Validity	Construct saturation	Subgroup differences	Applicant perceptions	Specific vs. common effects
Stimulus format	Cognitive load theory, (Sweller, 1988); media richness theory (Daft & Lengel, 1984; Potosky, 2008); multiple resource theory of attention (Wickens, 1984); social bandwidth theory (Potosky, 2008)	Higher validity for audiovisual vs. textual interpersonal SJTs: meta-analysis (Christian et al., 2010)	More cognitive saturation for textual vs. audiovisual situational judgement tests (SJTs, Lievens & Sackett, 2006)	Smaller Black–White differences for audiovisual vs. textual SJTs (Chan & Schmitt, 1997)	More favorable perceptions of audiovisual vs. textual SJTs (Kanning et al., 2006); Less favorable perceptions of videoconference vs. face-to-face interviews (e.g., Chapman et al., 2003)	Selection procedures studied are SJTs and interviews. Effects are not common but specific for SJTs (interpersonal constructs) and interviews.
Contextualization	Interactionism (Mischel & Shoda, 1995); procedural justice theory (Gilliland, 1993); information processing theory (Hembree, 1992; Kintsch, 1988); cognitive load theory (Sweller, 1988); pragmatic reasoning schema theory (Cheng & Holyoak, 1985)	Large effect of no vs. low on validity of personality ratings: meta-analysis (Schaffer & Postlethwaite, 2012) Higher validity for past-behavior interviews than for situational interviews, controlling for response evaluation consistency (Taylor & Small, 2002)	More error variance for no vs. low in personality scales (Robie et al., 2000; Schmit et al., 1995); Less error variance for no vs. high for cognitive tests (Hattrup et al., 1992) Larger cognitive saturation for situational interviews; larger personality saturation for past-behavior interviews: meta-analysis (Berry et al., 2007)	No effect of no vs. high on Black–White differences in cognitive tests (DeShon et al., 1998; Hattrup et al., 1992)	No effect of no vs. low on perceptions of personality scales (Holtz et al., 2005)	Selection procedures studied are personality inventories, SJTs, interviews, and cognitive tests. Effects seem to be common: Increased validity for the higher contextualized versions in both personality and interviews. Consistent effects on construct equivalence across GMA and personality.

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3 (*continued*)

Predictor method factor	Relevant theories	Validity	Construct saturation	Subgroup differences	Applicant perceptions	Specific vs. common effects
Stimulus presentation consistency	Procedural justice theory (Gilliland, 1993); media richness theory (Daft & Lengel, 1984; Potosky, 2008); trait activation theory (Tett & Burnett, 2003); item generation theory (Irvine & Kyllonen, 2002)	Higher validity for higher levels of structure in interviews: meta-analysis (Huffcutt & Arthur, 1994)	Lower cognitive saturation for higher levels of structure in interviews: meta-analysis (Berry et al., 2007) Lower levels of equivalence for adaptive stimuli in SJTs (Lievens & Sackett, 2007) and ACs (Brummel et al., 2009)	Smaller subgroup differences for higher levels of structure in interviews in concurrent settings: meta-analysis (Huffcutt & Roth, 1998)	Moderate relation between consistency and applicant reactions: meta-analysis (Hausknecht et al., 2004) Lower interactional justice perceptions of structured interviews (Conway & Peneno, 1999); Lower perceptions of fixed vs. adaptive SJTs (Kanning et al., 2006)	Selection procedures studied are interviews, SJTs, and assessment centers (ACs). Most effects are common: Consistent effects on construct equivalence (reliability) and validity across interviews, SJTs and ACs. Consistent effects on applicant perceptions across interviews and SJTs.
Response format	Cognitive load theory (Sweller, 1988); media richness theory (Daft & Lengel, 1984; Potosky, 2008); multiple resource theory of attention (Wickens, 1984); prejudice theories (Koch et al., 2015)	Same validity for MC and written constructed knowledge test scores (Edwards & Arthur, 2007) Higher validity for webcam vs. written constructed interpersonal SJT scores (Funke & Schuler, 1998; Lievens, De Corte, & Westerveld, 2015)	Higher cognitive saturation for written constructed SJT scores; higher Extraversion saturation for webcam SJT scores (Lievens, De Corte, & Westerveld, 2015)	Smaller Black–White differences for written constructed knowledge test (Edwards & Arthur, 2007)	Higher job relatedness perceptions for written constructed knowledge tests (Edwards & Arthur, 2007) Higher media richness perceptions for webcam SJTs (Lievens, De Corte, & Westerveld, 2015)	Selection procedures studied are cognitive tests, SJTs, and ACs. Some effects are common: Consistent effects on construct saturation and applicant perceptions across cognitive tests and SJTs. Other effects are specific: Validity effects are moderated by the construct (cognitive vs. noncognitive).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3 (*continued*)

Predictor method factor	Relevant theories	Validity	Construct saturation	Subgroup differences	Applicant perceptions	Specific vs. common effects
Response evaluation consistency	Cognitive continuum theory (Hammond, 2010); rating process models (Ilgen et al., 1993; Lord & Maher, 1990)	Higher validity for higher levels of structure in interviews: meta-analysis (Huffcutt & Arthur, 1994); Higher validity for scoring keys that control for rating tendencies in SJTs (McDaniel, Psotka, Legree, Yost, & Weekley, 2011).	Scoring key affects the type of procedural knowledge measured in SJTs (Motowidlo & Beier, 2010)	Smaller subgroup differences for higher levels of structure in interviews in concurrent settings: meta-analysis (Huffcutt & Roth, 1998) Scoring key that controls for rating tendencies reduces Black-White differences in SJTs (McDaniel et al., 2011).	Moderate relation between consistency perceptions and applicant reactions: meta-analysis (Hausknecht et al., 2004)	Selection procedures studied are interviews, SJTs, and ACs. Most effects are common: Effects on reliability and validity are consistent across interviews, SJTs, and ACs. Effects on subgroup differences and applicant reactions often confound different types of consistency.
Information source	Socioanalytic theory (Hogan & Shelton, 1998); self-other knowledge asymmetry model (Vazire, 2010); realistic accuracy model (Funder, 1999); social relations model (Kenny, 1994)	Other-reports of personality add to self-reports: meta-analyses (Connelly & Ones, 2010; Oh et al., 2011) Use of psychologist and peer assessors leads to higher validities: meta-analysis (Gaugler et al., 1987).		Minor Black–White differences for self-reports of personality inventories (Bobko & Roth, 2013)	Mediocre perceptions for self-reports of personality: meta-analysis (Hausknecht et al., 2004)	Selection procedures studied are personality inventories, interviews, and ACs. Consistent effects across these selection procedures for complementary use of information sources to increase validity.

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3 (*continued*)

Predictor method factor	Relevant theories	Validity	Construct saturation	Subgroup differences	Applicant perceptions	Specific vs. common effects
Instructions	Situational strength theory (Meyer et al., 2010); procedural justice theory (Gilliland, 1993)	Similar validity for knowledge and behavioral tendency SJTs: meta-analysis (McDaniel et al., 2007) Lower validity for transparent constructs in ACs (Ingold et al., 2016)	Larger cognitive saturation for SJTs with knowledge instructions; larger personality saturation for behavioral tendency SJTs: meta-analysis (McDaniel et al., 2007)	Smaller Black-White differences for behavioral tendency instructions: meta-analysis (Whetzel, McDaniel, & Nguyen, 2008)	Higher opportunity to perform perceptions in ACs with transparent constructs (Ingold et al., 2016)	Selection procedures studied are personality inventories, interviews, SJTs, and ACs. Some effects are common: Effects of transparency instructions on validity. Other effects are specific: Effects of knowledge vs. behavioral tendency instructions are specific to SJTs.
Conclusions		Higher validity for - Audiovisual stimuli and constructed responses (interpersonal constructs); - Higher consistency levels; - Different information sources; - Instructions that do not make constructs measured transparent; - Low contextualization (personality).	Less cognitive saturation for - Audiovisual stimuli; - Constructed responses; - Instructions triggering past behavior; - Instructions triggering behavioral tendencies (instead of knowledge).	Smaller subgroup differences for - Audiovisual stimuli; - Constructed responses; - Higher consistency levels; - Instructions triggering past behavior; - Instructions triggering behavioral tendencies (instead of knowledge).	More favorable applicant perceptions for - Audiovisual stimuli; - Face-to-face stimuli; - Adaptive stimuli; - Constructed response formats; - Consistency in response evaluation.	

Webster, 2003; Sears, Zhang, Wiesner, Hackett, & Yuan, 2013; Van Iddekinge et al., 2006).

Conclusion. Prior research related to the stimulus format shows the importance of this predictor method factor as determinant of selection outcomes (criterion-related validity, construct saturation, subgroup differences, and applicant perceptions) for SJTs and interviews. However, given that prior research focused only on two selection procedures (interpersonal SJTs and interviews) and on a comparison of a limited number of stimulus factor choices, it is difficult to draw conclusions about whether the effects of this predictor method component generalize across selection procedures in general. As it seems now, Table 3 suggests that the effects are rather construct and selection procedure specific than common across constructs and selection procedures. For example, audiovisual stimulus formats affect the validity of SJT scores only when they reflect interpersonal constructs.

Contextualization

Definition. We define contextualization as the extent to which test stimuli are embedded in a detailed and realistic context.[5] This method factor resembles the "authenticity" factor in Le's (2013) framework on simulations and the "fidelity" one in Parshall et al.'s (2000) framework on PC-based testing (Table 1).

Prior research. In prior research, different levels of contextualization were adopted.[6] At one extreme, test stimuli can be void of any contextualization to minimize adding unintended variance to test scores. This *decontextualized* category is exemplified by many verbal or numerical reasoning items and personality items.

In *low levels of contextualization*, a situational keyword (aka tag) is added. So far, most selection research on contextualization has focused on the effects of adding such minor levels of contextualization (e.g., "at work" tags) to existing personality items (Table 3). The underlying idea of adding contextual tags is based on interactionism, namely that personality is not a consistent predictor across different situations because people's behavioral tendencies are a function of their individual characteristics as well as their perception of the situation (Jansen et al., 2013; Mischel & Shoda, 1995). It then follows that better prediction for work criteria can be obtained for contextualized "at work" scales than for generic ones. As shown in Table 3, a meta-analysis confirmed that contextualized tags increased mean validities of personality scores from .11 to .24 (Shaffer & Postlethwaite, 2012). There is also evidence that contextualized personality scores have incremental validity over generic ones (Bing, Whanger, Davison, & VanHook, 2004). Moreover, research found that the factor structure of contextualized and generic personality ratings was invariant, but that error variances were smaller in the contextualized form (Robie, Schmit, Ryan, & Zickar, 2000; Schmit, Ryan, Stierwalt, & Powell, 1995). Last, there is scant research on perceptions of contextualized personality scales. Students favored the contextualized variant over the generic one but the difference in perceptions between the two formats did not reach statistical significance (Holtz, Ployhart, & Dominguez, 2005).

A *medium level of contextualization* is characterized by inserting general contextual descriptions. This means that the context is broadly depicted in terms of "who," "when," "where," and "why" (see Johns, 2006). In situational interviews ("What would you do if you encountered the following situation . . .?") and SJTs, such medium levels of contextualization are adopted. In line with interactionism (Campion & Ployhart, 2013), it is assumed that test takers make sense of this general context and that this construal guides responses. In SJTs, research on the effects of medium contextualization levels is scarce. As an exception, Krumm et al. (2015) demonstrated that up to 70% of SJT items could be solved correctly even when the context (item stem) was stripped from the items. This result raises questions about the interactionist assumptions underlying SJTs.

Finally, *high levels of contextualization* are characterized by specifying the contextual information. In other words, whereas in medium contextualization the "who," "when," "where," and "why" are described only in general terms, in high levels of context, more detailed information is given about each of these aspects (e.g., the "who" via a description of the main characters, the "where" via a description of the firm). We find this high level of contextualization in serious games, AC exercises, and in behavioral description interviews in which candidates are asked to describe in detail a past situation encountered. There exists a lot of research on the differences between situational (medium level of contextualization) and behavior description interviews (high level of contextualization). The meta-analysis of Taylor and Small (2002) revealed that past behavior questions demonstrated higher validity for predicting job performance than did situational questions, when response evaluation consistency (rating scale type) was controlled for. Regarding construct saturation, Levashina, Hartwell, Morgeson, and Campion (2014) reviewed the relationship between situational and behavior description interview scores and other constructs and concluded that the two interview types measure different constructs, with situational interviews more strongly related to cognitive ability and job knowledge (Berry, Sackett, & Landers, 2007) and behavior description interviews more strongly related to experience and personality traits such as achievement orientation, extraversion, and oral presentation skills.

There is also some research on adding detailed realistic context to cognitive ability tests (e.g., via business-related graphs and tables). Hattrup, Schmitt, and Landis (1992) found that such contextualized ability tests assessed constructs equivalent to the ones in traditional ability tests, although reliabilities were higher for traditional tests. Adding business-related (Hattrup et al., 1992) or social context (DeShon, Smith, Chan, & Schmitt, 1998) to ability items did not produce the expected decrease in Black–White subgroup differences.

Conclusion. Research on contextualization is predominantly conducted in the personality, interview, and ability domains. Table 3 reveals relatively consistent results across them. When different contextualization levels are compared (e.g., in personality tests or interviews), validity is higher for the more contextualized variant. Research on the equivalence of different contextualization conditions also paints a consistent picture: Error variances are smaller when tags (low contextualization levels) are added to personality

[5] In personnel selection, this context will typically be job-related. Yet, this is not always the case. For example, one might embed math problems in shopping or in train schedules. Our definition adheres to the level of contextualization (and not to the type of context).

[6] Although one might equate the classic distinction between "signs" and "samples" with this method factor, this does not capture the full range of contextualization levels outlined here.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

scales, whereas higher contextualization levels increase error variance in cognitive test scores. Contextualization effects on subgroup differences and applicant reactions are minimal.

Stimulus Presentation Consistency

Definition. We define this predictor method factor as the level of standardization that interviewers/assessors/test administrators adopt in presenting test stimuli to test-takers. In other words, this factor refers to the degree to which procedural variations in presenting test stimuli across test takers are reduced. Earlier frameworks included this predictor method factor using terms such as "standardization of stimulus material" and "question standardization" or their antonyms ("stimuli flexibility" and "interactivity"; Table 1).

Prior research. In general, three broad categories of stimulus presentation consistency can be distinguished in prior research across selection procedures. In the first category, *free stimuli*, there exist virtually no predetermined and standardized guidelines regarding the stimuli to be presented. Unexpected candidate responses and reciprocal interactions between the parties might lead to turns and sidetracks so that new and unforeseen stimuli occur. Examples are interviews or role plays without question standardization.

The second category is called *adaptive stimuli*, in which there exist predetermined and standardized guidelines about the key stimuli to be presented, whereas the administration of substimuli depends on test-takers' responses to the previous stimuli. So, the path that a candidate takes through the assessment is contingent upon a candidate's prior responses to the key stimuli, thereby creating some level of interactivity. Guidelines given to interviewers to formulate questions around a specific set of main topics in the employment interview constitute one example. Similarly, role-players might receive guidelines to discuss a series of themes in a role-play exercise. Depending on the candidate replies, the conversation wanders within the boundaries of the broad script. Other well-known examples are branched/nested/nonlinear SJT items (where administration of branched items depends on candidates' replies to a previous key item, Kanning et al., 2006) or computer adaptive test (CAT) items (where the number and the difficulty levels of subsequent items are based on candidates' performance on previous items). These examples show that within this adaptive stimuli category, there exists a finer differentiation between adapting the stimuli to be presented through person-based approaches (e.g., interviewers, role-players) versus technology-based approaches (e.g., branching, CAT).

The third and last category pertains to *fixed stimuli* wherein there exist predetermined and standardized guidelines so that all test-takers are presented with the same or comparable stimuli in the same order (no matter how they respond to the stimuli) and progress in the same way. Stimuli do not need to be identical across candidates; we view statistically equated alternate forms as fitting within the rubric of fixed stimuli. Predetermined time limits in the presentation of the stimuli can also be added. Traditional paper-and-pencil ability tests are a prime exemplar of the use of fixed stimuli. Other examples are interviewers asking the same questions in the same order across candidates (highly structured interviews) or role-players using a set of predetermined verbatim and sequenced prompts.

Most of our knowledge accumulated over the years related to this predictor method factor comes from employment interview research (Table 3).[7] A robust finding is that—in line with psychometric theory—higher levels of interview structure (i.e., combination of stimulus presentation consistency and response scoring consistency, see below) reduce error variance (idiosyncratic interviewer biases) and produce higher reliability (Huffcutt, Culbertson, & Weyhrauch, 2013). In addition, increasing structure in the interview has beneficial effects on validity up to a level where validities seem to asymptote (Huffcutt & Arthur, 1994). In terms of construct saturation, Berry et al.'s (2007) meta-analysis found that higher levels of consistency in interviews result in lower interview-cognitive test correlations. This might also explain why higher interview structure is associated with smaller subgroup differences than lower interview structure (Huffcutt & Roth, 1998). It should be noted further that the subgroup differences for ratings in higher structured interviews increase when cognitive constructs are assessed and in applicant samples (instead of in incumbent samples; Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002).

In the last years, the effects of stimulus presentation consistency have also been examined outside the interview domain. Although Item Response Theory (IRT) permits test designers to keep the reliability and construct measurement of cognitive ability and personality scores constant across test takers and administrations, this endeavor is considerably more challenging for approaches that capture a variety of (sometimes poorly understood) constructs as is the case in SJTs. Therefore, principles behind item generation theory (Irvine & Kyllonen, 2002) have been used in SJTs to isolate "radicals" (item characteristics that matter) from "incidentals" (i.e., superficial item characteristics). If SJT items can be developed that differ only in terms of incidental characteristics, it might be possible to develop adaptive SJTs that still capture the same underlying constructs. Yet, even small variations in the situations presented in SJT items significantly lower alternate-form equivalence (Lievens & Sackett, 2007).

Apart from SJTs, AC exercises are another application domain for examining the effects of stimulus presentation consistency. This is needed because similar equivalence problems as with SJTs have been observed even among carefully developed alternate AC exercises (Brummel, Rupp, & Spain, 2009). Therefore, trait activation theory (Tett & Burnett, 2003) has been employed for developing adaptive role-player prompts (Lievens, Schollaert, & Keen, 2015; Lievens, Tett, & Schleicher, 2009). An advantage of this interactionist theory for developing such adaptive stimuli is that it enables identifying slightly different situational cues that still activate the same underlying constructs.

A final piece of knowledge deals with the effects of stimulus presentation consistency on applicant perceptions. Procedural justice theory (Gilliland, 1993) served as main theoretical framework. Meta-analytic research across selection procedures reveals that consistency perceptions and overall procedural justice perceptions are moderately related (Hausknecht, Day, & Thomas, 2004). Yet, there is also a point where too much consistency results in lower

[7] In most prior research, however, the effects of stimulus presentation consistency were confounded with those of response evaluation consistency.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

interactional justice perceptions. For instance, interviewees perceive high structure interviews as "cold" (Conway & Peneno, 1999). SJTs with fixed stimuli are also less favorably perceived than branched SJTs with adaptive stimuli (Kanning et al., 2006).

Conclusion. Given that well-developed technologies (e.g., IRT) exist for ensuring stimulus presentation consistency in General Mental Ability (GMA) tests and personality scales, most past research on this factor focused on comparing low versus high levels of stimulus presentation consistency in interviews. Table 3 shows that the effects of higher consistency levels reducing measurement error and increasing validity are well established in the interview domain and seem to extend to other domains (SJTs and ACs) as well. Consistency is also a key determinant of applicant perceptions of selection procedures. A last conclusion is that the effects of extreme levels of stimulus presentation consistency on validity and applicant perceptions (interactional justice) are marginal or even detrimental.

Response Format

Definition. We define response format as the modality by which test-takers are required to respond to test stimuli (see Edwards & Arthur, 2007). So, this factor does not refer to how these responses are subsequently evaluated (i.e., response evaluation consistency below). As shown in Table 1, this predictor method factor was represented in earlier frameworks as "response type," "structure of response mode," and "response action."

Prior research. Traditionally, *close-ended* response formats (multiple-choice or forced-choice response formats) have been most frequently used in selection. In this response format, the possible response options are predetermined and prompted. Candidates choose, rank, or rate the predetermined response options. The close-ended response options might be text-based, pictorial, auditory, or video-based (see Sireci & Zenisky, 2006, for a list of innovative Multiple Choice [MC] formats).[8]

Over the years, alternatives to close-ended formats have been sought in the form of open-ended (aka constructed) response formats. The same categories apply here as the ones discussed for stimulus format. So, a second response format category comprises *textual* constructed responses in which candidates produce a textual response. Examples are essays, constructed responses to a planning exercise, or sentence completion. In a third category, candidates are required to produce a *pictorial* response. An example is a creativity test in which candidates are asked to draw a picture. Especially in the educational domain, there exists a longstanding research tradition of comparing these constructed responses with close-ended ones. The most recent meta-analysis (Rodriguez, 2003) revealed that close-ended scores had higher reliabilities than their constructed counterparts. Construct equivalence could be established only when the two response formats kept the item stem constant. In case of different item stems, construct equivalence was significantly lower.

Only recently selection researchers have started to experiment with constructed response formats. On the basis of cognitive load theory (Sweller, 1988), it has been argued that constructed formats lead to lower cognitive load and therefore lower subgroup differences. Edwards and Arthur (2007) confirmed that written constructed responses to a knowledge test substantially reduced subgroup differences and yielded more favorable test perceptions among African Americans compared with close-ended ones. Similar criterion-related validity results were found for the two formats of this knowledge test. Conversely, Funke and Schuler (1998) discovered significant criterion-related validity differences between these two formats for an interpersonal SJT. Recently, Arthur et al. (2014) compared integrity SJT scores across three close-ended response formats (rate, rank, and pick the best). Thus, they focused on a finer distinction within the close-ended response format category. The rate response format came out as most favorable: It did not add unintended cognitive load and led to lower subgroup differences.

Because of information technology advancements, constructed formats are no longer limited to textual constructed ones. On the basis of media richness theory (Daft & Lengel, 1984), new categories, such as *audio and audiovisual* response formats, have been proposed. Typical examples of the audio category include candidate answers to telephone interview questions, whereas in *audiovisual* responses, candidates are asked to videotape their performance. Examples here are video resumes (Waung, Hymes, & Beatty, 2014) or webcam SJTs, in which candidates' performance is recorded when reacting to short scenarios. It is then argued that the use of such a response format conforms more to the communal nature of interactions in specific subgroups, which in turn might reduce subgroup differences (see theories about cultural interaction patterns; Helms, 1992). Comparative research related to these recent constructed response formats is scarce. One study discovered that *audiovisual* response format scores had higher extraversion saturation and higher validity than written constructed ones (Lievens, De Corte, & Westerveld, 2015). This initial piece of evidence seems to suggest that such audiovisual response formats generate construct-relevant information for predicting sales, leadership, or interpersonal performance. Finally, the sixth and seventh response format categories refer to *videoconference and face-to-face interactions*, respectively. Examples include videoconference (remote) or live interactions with interviewers, role-players or with a panel during a presentation. These formats are richer than the previous ones because there is two-way communication among candidates and interviewers or role-players (either face-to-face or via videoconference).

Conclusion. In recent years, the search for response formats other than close-ended ones has generated increasing interest. Similar to the research base on stimulus format, cognitive load theory and media richness theory have been used as theoretical frameworks. As shown in Table 3, the research evidence is mostly based on comparisons between close-ended and constructed textual response formats. As a key conclusion, use of these constructed formats seems to result in less cognitive load, more favorable applicant perceptions, and smaller subgroup differences. It is important that these results have been found consistently across various selection procedures (cognitively oriented tests, SJTs). The effects of close-ended versus constructed response formats on the criterion-related validity of test scores seem to depend on the construct. Only for interpersonal constructs is va-

[8] Other finer distinctions are possible in terms of the number of responses or media used (e.g., PC, smartphone). Time limits or requirements to elaborate (e.g., Levashina, Morgeson, & Campion, 2012) might also be included.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

lidity higher for constructed formats. Strikingly, our review revealed similar effects for stimulus format manipulations. Audiovisual stimuli led to more attractive and less cognitively saturated test scores with smaller subgroup differences and higher validity (but again only for interpersonal constructs).

Response Evaluation Consistency

Definition. We define response evaluation consistency as the level of standardization that interviewers/assessors/test administrators adopt in terms of evaluating test-takers' responses. This factor pertains to reducing procedural variations in how test takers' responses to the stimuli are evaluated. Table 1 shows that response evaluation consistency was present in earlier frameworks as "scoring modality," "scoring standardization," "scoring algorithm," or "scoring/evaluation focus."

Prior research. The issue of response evaluation consistency has received a lot of attention in nearly all selection procedures, with the majority of research conducted in interviews, SJTs, and ACs. Researchers relied on two broad sets of theoretical frameworks, namely judgment and decision-making models (e.g., Hammond, 2010; Lord & Maher, 1990) and performance rating models (e.g., Ilgen, Barnes-Farrell, & McKellin, 1993).

Generally, three categories of response evaluation consistency were studied in past research.[9] In the first category that we label *unconstrained judgment,* one (e.g., interviewer, assessor) evaluates candidates without having preestablished answers or evaluative standards. Global interviewer judgments of interviewees exemplify this category.

The second category that we refer to as *calibrated judgment* implies that interviewers or assessors are trained to use preestablished answers and/or evaluative standards when evaluating candidates, as is often the case in scoring interview answers, essays, role-plays, ACs, and work samples (e.g., Melchers, Lienhardt, Von Aarburg, & Kleinmann, 2011; Woehr & Arthur, 2003). To ensure calibrated judgments, over the years, a plethora of rating aids (e.g., checklists, scoring rubrics) and interviewer/assessor training interventions (e.g., frame-of-reference training) have been proposed. Space constraints preclude detailed discussion of their effectiveness.

The category highest in terms of response evaluation standardization consists of *automated scoring.* Here no interpretative leaps are required because an a priori scoring key (determined via empirical keying, theoretical keying, expert keying or a combination of those) is applied for evaluating candidates. Automated scoring is typically done via computer algorithms, which might vary from simple (dichotomous) to complex (e.g., polytomous or partial credit scoring systems where answers are scored on a number of weighted criteria, Parshall et al., 2000). Automated scoring applies not only to ability tests, biodata, personality scales or SJTs, but also to essays and simulations (see Clauser, Kane, & Swanson, 2002). Again, the literature about the effectiveness of different scoring approaches (e.g., Bergman, Drasgow, Donovan, Henning, & Juraska, 2006) is too voluminous to discuss here. Given the Big Data movement, automated scoring algorithms are likely to expand (Oswald & Putka, in press).

Given that response evaluation consistency has received substantial research attention in selection and related literatures, common results across a variety of selection procedures can be identified (Table 3). As one common thread, calibrated judgment approaches seem to be effective in reducing the interpretative leaps required from interviewers/assessors and minimizing unintended variance in the form of rater idiosyncrasies and rating effects. In turn, this seems to lead to increases in reliability and criterion-related validity. Another key result is that the type of automated scoring key affects the validity of test scores (e.g., Bergman et al., 2006). There is also recent research suggesting effects of the scoring key on construct saturation. For instance, Motowidlo and Beier (2010) manipulated the SJT scoring key (experts vs. novices) on the basis of their theory about knowledge determinants underlying SJTs. These different scoring techniques affected the constructs measured because the SJT measured either job-specific or general domain knowledge. Finally, evidence is suggestive regarding the effects of response evaluation consistency on reducing subgroup differences. Huffcutt and Roth's (1998) meta-analysis showed smaller subgroup differences in structured interviews than in unstructured ones. Regarding applicant perceptions, Hausknecht et al.'s (2004) meta-analysis reported a positive correlation between applicant perceptions of consistency and reactions to selection tools.

Conclusion. Similar to stimulus presentation consistency, response evaluation consistency appears to have common effects across a range of selection procedures, with higher levels leading to less error variance, higher validity, smaller subgroup differences, and favorable applicant perceptions. One caveat is in order, though. Response evaluation consistency effects can often not be distinguished from stimulus presentation consistency effects. For instance, structured interview studies typically encompass both components.

Information Source

Definition. Information source refers to the individual responding to the test stimuli.

Prior research. Three main information categories can be distinguished in prior research. The first category, *behavior exhibited by the candidate or choices made by the candidate in the assessment context*, denotes that the test-taker him-/herself responds to the test stimuli (e.g., completes ability test or SJT items, participates in assessment center exercises). Candidate behavior is subsequently evaluated, either with an objective scoring system or a judgmental process (e.g., an assessor rating). Here differing types of judges can be used, and the effects of differing judges (e.g., assessment center ratings by managers vs. psychologists) on the outcomes of interest can be examined. The second category, *self-reports by the candidate about events beyond the assessment context*, refers to candidates' reports of behaviors, attitudes, values, beliefs, or intentions not bounded by the immediate assessment context (e.g., self-report personality measures, interest measures, life history items, inquiries about plans and intentions). The third category, *reports by others about events outside the assessment context*, parallels the second, except that someone other than

[9] Finer distinctions are again possible. For instance, each category can vary from holistic (globally evaluating performance) to analytic (evaluating each response, Klein et al., 1998; Le, 2013). These various levels apply to the evaluation of individual items/responses and integration of responses to form a total score for a predictor.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the candidate provides the information. These persons should be well acquainted with the focal person and motivated to share job-related information about him/her. Examples are coworkers, supervisors, or direct subordinates. In a selection context, friends or relatives are typically not used.

In the first category, the use of different types of evaluators has been examined. In the domain of assessment centers, Gaugler, Rosenthal, Thornton, and Bentson (1987) report higher validity for psychologists and peer assessors, relative to managers. The other categories (self vs. other-reports) have been predominantly investigated in the personality field. Conceptually, this body of research is based on cumulative knowledge models that posit that each information source adds information over the other one. For example, the self-other knowledge asymmetry model (Vazire, 2010) stipulates that the self has more difficulties with constructs high in evaluativeness (e.g., intellect), whereas constructs low in observability (e.g., emotional stability) are more difficult to assess by others. Similarly, socioanalytic theory assumes that self-ratings reflect one's identity, while other-ratings represent one's reputation (Hogan & Shelton, 1998). According to these models, each of these two information sources balance out their respective drawbacks (self-reports: leniency and impression management; other reports: friendship biases). Generally, the evidence confirmed that adding other-ratings to self-ratings substantially increases the validity of personality for predicting job performance (see meta-analyses, Connelly & Ones, 2010; Oh, Wang, & Mount, 2011, and primary studies, e.g., Kluemper, McLarty, & Bing, 2015; Zimmerman, Triana, & Barrick, 2010). Similar results with other-ratings were found for emotional intelligence (e.g., Elfenbein, Barsade, & Eisenkraft, 2015) and SJTs (MacCann, Wang, Matthews, & Roberts, 2010), though those studies were not done in a selection context.

A given construct can potentially be addressed via different information sources. There is an emerging literature on using employment interviews to assess personality, and Levashina et al. (2014) speculated that interviewer ratings of personality are superior to self-reports. Although research has examined convergence between interviewer and self-reports of personality (Levashina et al., 2014, pp. 262–263), comparative criterion-related validity has not been reported.

Conclusion. Research on this factor has increased in recent years. Table 3 shows consistent evidence across various selection procedures, with significantly higher validities when different sources are combined. This result supports cumulative knowledge frameworks underlying the use of different information sources.

Instructions

Definition. Instructions denote the extent to which directions are made explicit to test-takers about which perspective to take to respond to test stimuli. Only one prior framework (Cattell & Warburton, 1967) included this factor and labeled it "instruction system."

Prior research. On the basis of prior research and situational strength theory (Meyer, Dalal, & Hermida, 2010), we make a distinction between general (weaker) and specific (stronger) instructions. In some cases, candidates receive *general instructions* on how to respond to test stimuli. These instructions do not specify a perspective to candidates on how to respond to test stimuli (e.g., "rate yourself on the following statements" or "answer each of the following interview questions"). In other cases, more *specific instructions* are provided which add a specific perspective for responding to test stimuli.

There exist various ways to make instructions more specific. One example is using a time-bound frame (instead of an unspecified time frame) when probing past behavior. According to the behavioral consistency principle the past behavior-future behavior relationship should be stronger when focusing on the recent past (e.g., asking a firefighter candidate "have you run a 10K in the last year?" reveals more about current physical fitness than "have you ever run a 10K?"). As another example, in ability tests, one might mention that there is a penalty for guessing (instead of right number scoring), thereby changing how test-takers might approach the test stimuli (Rowley & Traub, 1977). In personality measures, a faking warning that stipulates that faking detection mechanisms are in place based on candidates' responses to a PC-administered measure has also been found to affect how candidates approach the test stimuli compared with general instructions that do not specify such faking detection mechanisms (e.g., Landers, Sackett, & Tuzinski, 2011).

The purpose of specific instructions is to reduce construct-irrelevant variance, and thus specific instructions are often preferred. However, in line with situational strength theory, one should not make specific instructions too strong. This is confirmed by research on transparency (i.e., specific instructions that reveal the constructs measured to candidates in a selection procedure). Such transparency instructions seem to be a mixed blessing (Ingold, Kleinmann, König, & Melchers, 2016; Kleinmann et al., 2011; Smith-Jentsch, 2007). In most studies, they enhance perceptions of opportunities to perform, performance, and construct measurement; yet, they also lower validity because they make the situation stronger and suggest to candidates what they should do, rather than allow them to choose what to do (Smith-Jentsch, 2007). This explanation of transparency removing construct-relevant variance fits well with evidence on the validity of candidates' spontaneous inferences about constructs measured in interviews and ACs (i.e., ability to identify criteria; Jansen et al., 2013).

Similar to other predictor method factors, we note that finer distinctions can be made. For example, in SJTs, the distinction between two more general instructions, namely behavioral tendency instructions ("what would you do?") and knowledge-based instructions ("what should you do?") has been widely researched. Meta-analytic research shows that SJT behavioral tendency instructions exhibit higher personality saturation and lower subgroup differences, while knowledge-based instructions show higher cognitive saturation and higher subgroup differences (McDaniel, Hartman, Whetzel, & Grubb, 2007). Criterion-related validity was unaffected.

Conclusion. Research on instructions spans a variety of selection procedures. As shown in Table 3, there is evidence across selection procedures that instructions are a powerful way of influencing candidates' construal of test stimuli and performance. For example, different instruction sets in SJTs affect construct saturation and subgroup differences. In addition, AC and interview research shows that overly strong instructions (i.e., transparency instructions) influence test performance and reduce validity, whereas this is not the case when candidates infer the constructs to be assessed themselves.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Summary: Are There Consistent Effects Across Selection Procedures?

In the prior sections, we presented a modular approach to selection procedures and reviewed the available research accordingly. This modular review brings together for the first time various selection procedure literatures that often evolved relatively independently from each other. As noted before, a key assumption underlying a modular approach is that the same components underlie different selection procedures and that they produce similar effects across them. Accordingly, deeper level similarities across different selection procedures and their literatures might be identified. In Table 3 (last column), we therefore summarized whether the effects of given predictor method factor choices are common (across selection procedures) or specific (per selection procedure and/or construct).

Generally, our review showed that there is evidence of consistent effects across selection procedures for the majority of predictor method factors (stimulus presentation consistency, response evaluation consistency, instructions, and to some extent also contextualization and information source). Conversely, the effects of stimulus and response format manipulations seem to be specific to selection procedures and constructs. That said, the sometimes fragmented research also suggests that we still have some leaps to take in the direction of a truly modular approach to selection procedures. More future evidence for common effects is important: It might promote a more integrative body of knowledge, theoretical connectivity, and cross-fertilization among the different selection procedure literatures because knowledge about a component might be used across various procedures.

Examining Common Selection Procedures Through a Modularity Lens

To this point, our discussion has focused on the predictor method factors, with various selection procedures used to illustrate the factors. We now shift our focus to an explicit examination of five common selection procedures (cognitive tests, personality inventories, interviews, SJTs, and ACs) using our seven-factor framework. Per procedure, we identify possible method factor choices and, where available, review research on the effects of method factor choices on criterion-related validity. The outcome of criterion-related validity is used simply to illustrate the effects of method factor choices; a similar examination could be done for other outcomes. In addition to illustrating effects of method factor choices on validity, this examination also sheds light on the modularity of the various selection procedures. Selection procedures for which a broad range of method factor configurations are possible are more modular than procedures where fewer method factor choices are viable.

Table 4 lists the common selection procedures and breaks them down by predictor method factor. Each of these common selection procedures can be seen as a historically derived constellation of particular method factor choices. In addition, we historically assigned constructs to such specific constellations. The cells in Table 4 present different constellations of each of these common selection procedures when predictor method choices other than the traditional ones are made (e.g., item stems set in a business context in ability tests). In Table 4, we also indicate whether research examined the effects of variation in each method factor, and highlight in bold when these different configurations mattered (i.e., improved validity).

Generally, most cells of Table 4 are filled, which signals that different configurations of the selection procedures exist. So, just like other modular systems, it seems possible to modify common selection procedures by "mixing and matching" their components. Yet, there are also differences in modularity between these selection procedures. Interviews, SJTs, and ACs can be situated on the higher end of the modularity continuum because all predictor choices have been manipulated (i.e., there are no blank cells) and several of them had substantial effects on criterion-related validity. Conversely, cognitive ability tests are situated at the lower end of the continuum because some predictor method choices do not make sense and the effects have generally been small. In light of their good predictive validity record, it is understandable that cognitive ability tests score lower on modularity; there is simply little need to experiment with different approaches.[10] In recent years, personality inventories have become increasingly modular because of calls for increasing their criterion-related validity, with changes in information source (other-reports) and contextualization ("at work" tags) producing substantial effects.

Scientific and Theoretical Utility of a Modular Approach in Selection

Generally, a modular approach instills a different mindset among selection researchers because it shifts the attention from thinking in terms of selection procedures as all-in-one packages to conceptualizing them in terms of their underlying components. Such a modular focus is of great scientific and theoretical utility for several reasons. Below we detail these reasons, moving from more descriptive to more prescriptive ones.

First, a modular focus has scientific merits in guiding an improved *description and documentation* of which predictor method factor/facet choices were operationalized in a given selection procedure, illustrating that a modular approach is also useful for knowledge accumulation even when nothing is manipulated. In case predictor method factors were manipulated, a modular approach also requires describing which ones where held constant. Such a careful description of the different method factor choices is important for subsequent meta-analyses. Many moderators in selection meta-analyses were typically study features (e.g., concurrent vs. predictive). A modular approach should lead to fine-grained meta-analyses at the level of predictor method factors or at the level of the interaction between predictor method factors and constructs.

Second, a modular approach has value in better *explaining* divergent findings across studies. Suppose two independent research teams examine the effects of stimulus format. One discovers the auditory format outperforms the textual one, whereas the other finds no differences. Yet, suppose one team used contextualized items, whereas the other relied on decontextualized items. Thus, insight in which factors were manipulated/held constant helps explaining divergent results.

[10] The situation is different for the large subgroup differences of cognitive ability tests. As shown in Table 3, some predictor method factors affect subgroup differences in cognitive ability tests.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Examples of Effects of Predictor Method Factor Choices on Criterion-Related Validity for Five Common Selection Procedures

Variable	Cognitive ability test	Personality inventory	Employment interview	Situational judgment test (SJT)	Assessment center
Stimulus format	Textual vs. pictorial tests	—	Live vs. telephone vs. videoconference interviews	**Textual (.27; k = 15; N = 8,182) vs. video-based interpersonal SJTs (.47; k = 2; N = 437) (Christian et al., 2010)**	Case analyses (.19; k = 11; N = 2,479) vs. group discussions (.17; k = 24; N = 5,009) (Hoffman et al., 2015)
Contextualization	Generic items vs. items set in business context	**Generic (.08; k = 72; N = 11,876) vs. contextualized Extraversion scale (.25; k = 18; N = 2,692) (Shaffer & Postlethwaite, 2012)**	**Past behavior with BARS (.63; k = 11; N = 1,119) vs. situational with BARS (.47; k = 29; N = 2,142) (Taylor & Small, 2002)**	Generic (.35; k = 71; N = 6,747) vs. detailed SJT item stems (.33; k = 10; N = 2,218) (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001)	Generic vs. customized AC exercises
Stimulus presentation consistency	Traditional vs. IRT based tests	Traditional vs. IRT based personality scales	**Low (.20; k = 15; N = 7,308) vs. high question standardization (.57; k = 27; N = 4,358) (Huffcutt & Arthur, 1994)**	Linear vs. branched SJTs	Different role-player prompt formats
Response format	Close-ended (.12; N = 220) vs. constructed response knowledge tests (.24; N = 235) (Edwards & Arthur, 2007)	**Different close-ended formats: Rating (.08; N = 60) vs. forced choice Conscientiousness scale (.46; N = 60) (Christiansen et al., 2005)**	Open-ended vs. MC interview questions	**Written (.08; N = 75) vs. audiovisual constructed SJT for predicting training scores (.30; N = 75) (Lievens et al., 2015)**	Open-ended vs. MC in-baskets
Response evaluation consistency	Different automated scoring formats (e.g., dichotomous, partial credit)	Different automated scoring formats (e.g., dimensional, typological)	**Different rating aids: Low (.20; k = 15; N = 7,308) vs. high scoring standardization (.57; k = 27; N = 4,358) (Huffcutt & Arthur, 1994)**	**Different automated scoring formats: raw consensus (.06; N = 702) vs. standardized consensus (.34; N = 702) (McDaniel et al., 2011)**	**Lower validities for control (.21; N = 40) vs. frame-of-reference training (.31; N = 40) (Schleicher, Day, Mayes, & Riggio, 2002)**
Information source	—	**Self- (.09; k = 37; N = 5,809) vs. other-report of Extraversion (.24; k = 14; N = 1,735) (Oh et al., 2011)**	Different interviewers (psychologists, managers)	Self- (.29; N = 324) vs. other-reports for predicting grade point average (GPA; .24; N = 324) (MacCann et al., 2010)	**Higher validities for psychologists than for managers as assessors (r = .26 between categorical moderator and validity) (Gaugler et al., 1987)**
Instructions	Information about presence or absence of penalty for guessing (Rowley & Traub, 1977)	Information about presence or absence of (real-time) faking warnings (Landers et al., 2011)	Transparency (.22; N = 110) vs. nontransparency instructions for predicting performance in simulations (.24; N = 159) (Klehe, König, Kleinmann, Richter, & Melchers, 2008)	Knowledge (.26; k = 96; N = 22,050) vs. behavioral tendency instructions (.26; k = 22; N = 2,706) (McDaniel et al., 2007)	**Transparency (.08; N = 87) vs. nontransparency instructions for predicting task performance (.24; N = 89) (Ingold et al., 2016)**
Conclusions: - Modularity level	Moderate	Moderate to high	High	High	High
- Impact of predictor method factor choices	Small effects of most manipulations	Moderate effects of some manipulations	Moderate effects of some manipulations; others unknown	Moderate effects of some manipulations	Moderate effects of some manipulations: others unknown

Note. Unless stated, job performance served as criterion. Dashes indicate that no manipulations have been conducted. Bold means that different predictor method factor choices significantly affected validity. BARS = Behaviorally-Anchored Rating Scales.

Third, by decomposing selection procedures into seven distinct components a modular focus opens up a plethora of opportunities to manipulate specific factors. So, a modular focus spurs more *experimentation and innovation* in the selection domain. For instance, the bulk of subgroup differences research focused on the stimulus format factor (see review of Schmitt & Quinn, 2010). Therefore, Bobko and Roth (2013) recently advocated that we should parse out selection procedures by other relevant method factors. Widening the scope of factors for reducing subgroup differences is exactly what a modular approach does. We anticipate most progress when researchers use theory to manipulate a limited set of facets of one or two predictor method factors, while holding others constant. Such experiments can be conducted in both lab and field settings. Although it will often be more feasible to do this in lab than in operational settings, the studies mentioned in Table 3 show that field experiments have been conducted. It is also possible to run field experiments in settings where experimental predictors are added to an operational test battery for research purposes. Moreover, as recently argued by Oswald and Putka (in press), Big Data and their "regular flow of data and re-occurring analytic cycles" shows considerable prospects for "a replicable empirical basis" for testing effects of predictor method factors.

Fourth, a modular approach leads to improved testing of and insight into *cause-effect relationships* because the effects of a given method factor are typically isolated from other confounding method factors. For example, Chan and Schmitt's (1997) influential study in which a video SJT was transcribed and test takers randomly assigned to video versus written conditions documented clearly the large causal effect of the unwanted cognitive load of the written version on subgroup differences.

Fifth, a modular approach searches for *more generalizable patterns* (i.e., common effects across selection procedures) that go beyond particular selection procedures, thereby promoting theoretical connectivity among different procedures. In turn, being better able to generalize across studies on selection procedures on the basis of their modular components leads to evidence-based prescriptive advice of both theoretical and practical utility. Rousseau (2006) calls such generalizable knowledge "Big E evidence" (vs. device-specific "small e evidence," p. 260). To illustrate this key benefit, let us come back to the earlier example of research on subgroup differences. Table 3 (last row) summarizes which method factor choices affect subgroup differences across selection procedures, namely (1) video-based stimulus formats (2) constructed response formats, (3) higher levels of response evaluation consistency, and (4) instructions that focus on behavior (instead of knowledge). Across the various selection procedures examined, it also becomes clear that reductions in cognitive saturation and rater idiosyncrasies explain why these factors reduce subgroup differences. In other words, in this specific domain, a modular review of evidence per predictor method factor leads to a more comprehensive picture of the components that affect subgroup differences and their underlying mechanisms.

Sixth, a modular approach has great theoretical utility for gaining insight in new selection trends (such as gamified assessment or scraping of social media content) because it enables *unpacking new unknown trends into known components.* By relying on the same seven-method factor framework to unpack these trends, one avoids being blind-sided by the novelty of such trends. In turn, this unpacking spurs hypothesis formulation regarding these new trends, which can lead to an agenda for future research on them. By extension, one of the potential contributions of a modular approach is that it sheds light onto selection procedures (which are comprised of various modular components) that have not been conceived or tested.

Because there is currently some debate about the reliability and validity of scraping social media content to make inferences about candidates' standing on Knowledge, Skills, Abilities, and Other Characteristics (KSAOs) in employment settings, we use this as an example to illustrate the benefits of how a modular approach unpacks new trends, and helps proposing new research questions and advancing knowledge. To start with, a key premise of a modular approach is that social media are not seen as an all-in-one technology (i.e., a black box) but rather as a collection of predictor method factors (see also McFarland & Ployhart, 2015). Social media content differs among others in terms of stimulus format (e.g., posting of texts, voice messages, pictures), information source (self-reports vs. endorsements and comments posted by others), stimulus presentation consistency (fixed sets of questions as in LinkedIn vs. free stimuli in Facebook), response evaluation consistency (extraction of social media content by recruiters vs. by machine-learning algorithms), and instructions (social media platforms as weak vs. strong situations). So, a modular focus encourages researchers to go beyond a specific social media format and use theory and prior research (Table 3) for testing hypotheses about which facets improve the reliability and validity of the inferences made via scraping content on social media platforms.

Although the previous information shows the scientific and theoretical benefits of a modular approach, a caveat is also in order. A modular focus should not prevent us from continuing to examine predictor methods as a whole. Thus, we warn against using this paper as a "hammer" for criticizing "holistic" selection procedure research. Both lenses are needed.

Practical Utility of a Modular Approach in Selection

A modular approach is useful for organizations when they face challenges with existing selection procedures and consider designing/using alternative ones. In such situations, a modular approach has merits in terms of (a) showing there exist a variety of alternative selection procedures and (b) providing flexibility to redesign existing selection procedures. The objective of a modular approach to selection procedure design consists of assembling predictor method factors to meet a set of desirable requirements (e.g., validity, subgroup differences, and/or applicant perceptions) for assessing a construct given a set of constraints (e.g., cost, time).

To put a modular approach into practice, we suggest adopting the following steps. We illustrate these steps with a recent study in which a modular approach was used to modify actual selection procedures and evaluate its effects (Lievens, De Corte, & Westerveld, 2015). First, it is critical to articulate the selection challenge that the organization is facing. The issues faced will often pertain to dissatisfaction with given selection procedures, search for alternative options, and/or optimization of multiple criteria (e.g., selection procedures with smaller subgroup differences, while still having equal validity and acceptable costs). In our real-world example, the organization had been using role-plays for years. However, the goal was to develop a more contemporary and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

less costly alternative that would also produce a better gender balance in the workforce. To this end, the organization had designed a written interpersonal SJT. Now, the general idea was to develop a hybrid between SJTs and ACs.

Second, we suggest breaking down the existing selection procedure into its underlying components. In our example (Lievens, De Corte, & Westerveld, 2015), the existing SJT was decomposed as follows: stimulus format (textual), contextualization (medium level), stimulus presentation consistency (fixed), instructions (knowledge-based), response format (close-ended written), response evaluation consistency (automated scoring), and information source (choices made by candidate).

Third, we suggest relying on theory and empirical research about each of these method factors (Table 3) to formulate hypotheses about which modifications in the underlying components have the greatest probability to reach a better solution. This step illustrates the enhanced design flexibility that flows from the ability to modify the separate components identified by a modular approach. In the product design literature, this benefit is referred to as the removal of problematic components (aka "exclusion") and the addition of improved components (aka "augmentation").

In the example, decomposing the SJT in its constituting components offered the organization various redesign strategies from which they made evidence-based choices on the basis of knowledge like that presented in Table 3. Specifically, the organization increased the levels of the stimulus and response format factors. The expectation was that changing the SJT stimulus format to an audiovisual one would increase the validity and make it more modern and attractive to applicants. As the organization was concerned that applicants do not show actual behavior in an SJT, it also changed the SJT's response format from close- to open-ended, which was expected to further increase validity on the basis of media richness theory.

In a fourth step, modifications are made to the selection procedures in line with the hypotheses posited. In the example, the organization converted the written SJT into an audiovisual one. Regarding response format, the organization set up a field experiment in which two response formats were pitted against each other: written constructed versus audiovisual constructed (webcam). That is, applicants wrote their answer in half of the video scenes and enacted their answer to a webcam in the other half. In both conditions, trained assessors rated the responses via checklists. These new selection devices were hybrids between AC exercises and SJTs. It is important that the administration costs of these two hybrid selection procedures were lower than that of prior role-plays.

In a fifth step, the effects of these modifications are evaluated in terms of the outcomes of interest. If the modifications do not lead to desired effects, further changes can be made and evaluated. In the example, criterion-related validity results favored the webcam format for measuring interpersonal constructs. However, this open-ended response format did not lead to a decrease in gender differences. The organization assumed this was because assessors saw the candidates in the webcam format (which might have triggered gender stereotypes; see Koch, D'Mello, & Sackett, 2015). Therefore, the organization made a modification related to response evaluation consistency: assessors were no longer pro-

Stimulus Presentation Consistency	Textual	Pictorial	Auditory	Audiovisual	Face-to-face
Fixed	Verbal reasoning items Self-report personality items	Graphical reasoning items Picture-based knowledge test items Facial recognition items	Fixed set of prerecorded telephone-based interview questions Tone Deaf Test Items	Linear multimedia SJT items	Fixed set of predetermined interview questions in same sequence Role-plays with predetermined verbatim prompts
Adaptive	CAT verbal reasoning items CAT self-report personality items	CAT graphical reasoning items CAT facial recognition items	Interactive Voice Recognition (IVR) questions CAT Tone Deaf Test Items	Branched multimedia SJT items CAT multimedia items	Interview questions about predetermined topics Role-plays with generic prompts (discretion for follow up contingent upon candidate responses)
Free	--	--	--	--	Interviews without question standardization Role-plays without any form of standardization
	Stimulus Format				

Figure 1. Example of Predictor Methodology Map. *Note*. Given that we do not want to suggest creating new predictors with characteristics that are undesirable, we placed dashes in those cells. For example, in the cell "textual-free," it is in principle possible to present each test-taker with different written open-ended questions and administer different written follow-up questions to each test-taker. However, this is typically not done for reasons of standardization.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5
Research Questions Per Predictor Method Factor Related to Trends in Selection Research and Practice

Trends in selection research and practice	Research questions	Practical applications
Stimulus format: 3D animated and avatar-based formats have made rapid inroads in practice as alternatives for classic audiovisual formats (e.g., Fetzer & Tuzinksi, 2014). The same is true for videoconference interactive formats as alternative to live face-to-face interactive formats. Systematic comparisons in terms of key outcomes are still scarce.	• To what extent do 3D-animated and avatar-based formats have added value above audiovisual formats in terms of validity and subgroup differences? • To what extent do pictorial and auditory formats have added value above textual formats in terms of validity and subgroup differences? Do construct saturation and applicant perceptions explain the potential effects? • How can cumulative contextual knowledge models (e.g., Gesn & Ickes, 1999) be used for understanding how stimulus formats add information above each other? • How do remote and face-to-face interactive formats compare in terms of adding construct-relevant and construct-irrelevant variance? • How does stimulus format interact with response format? Is a match in terms of media richness required for creating added value?	3D animated situational judgement tests (SJTs); avatar-based SJTs; videoconference interviews; remote (online) assessment center exercises.
Contextualization: The use of higher levels of contextualization in personality inventories substantially increases their validity. So far, other contextualization levels have remained virtually unexplored, even though such research might provide a fertile ground for connecting different selection literatures to one another (e.g., personality inventories, SJTs, and ACs).	• Which levels of contextualization decrease cognitive load (by activating real-world knowledge) and in turn affect subgroup differences? To what extent do group membership and candidate background moderate this effect on the basis of the concreteness fading principle (Fyfe et al., 2014)? • Do higher levels of contextualization always lead to better prediction than lower levels? What are boundary conditions? • To what extent do applicant perceptions of contextualization levels impact on test performance and interact with subgroup membership (see Ryan, 2001)? • What is the relative importance of contextualization and stimulus format for creating added value?	Decisions about the level of contextualization for a variety of traditional selection procedures (e.g., SJTs, assessment centers [ACs]) and more recent ones (e.g., serious games).
Stimulus presentation consistency: There exists growing interest in more adaptive assessment (e.g., Fetzer & Tuzinksi, 2014). Although CAT permits test designers to keep the reliability and construct measurement of ability tests and personality scales constant across test takers and administrations, research on using adaptive stimuli in other selection tools is still in its infancy.	• Can we use item generation theory and trait activation theory in research on adaptive stimuli? • How can construct equivalence for scores based on adaptive stimuli in SJTs be obtained? What is the validity of scores based on adaptive stimuli in SJTs? • How can construct equivalence for ratings based on alternate AC exercises be achieved? • To what extent do scores based on adaptive stimuli exhibit subgroup differences? • Can trade-offs be found between increasing applicant perceptions through the use of adaptive stimuli while still ensuring construct equivalence?	Adaptive assessment formats: Branched SJTs; AC exercises; adaptive simulations; serious games.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5 (*continued*)

Trends in selection research and practice	Research questions	Practical applications
Response format: There exists voluminous research on close-ended formats. Due to the advent of information technology, various constructed formats (especially audiovisual and videoconference interaction) are increasing in popularity, albeit with research lagging behind.	• To what extent do new response formats (audiovisual and videoconference interactions) have added value above close ended formats in terms of validity and subgroup differences? • What is the construct saturation of these new formats and to what extent does this drive subgroup differences? • To what extent do applicant perceptions mediate the relationship between the Response Format × Subgroup Membership interaction and test performance? • What are the underlying theoretical mechanisms behind the effects of these new response formats? Is cognitive load still the dominant explanation? What is the role of divergent thinking, cultural interaction patterns (Helms, 1992), and stereotypes (Koch et al., 2015)? • To what extent is a match between stimulus and response format categories required for ensuring validity and applicant perceptions? Which format (stimulus or response format) is most important for creating added value?	Remote assessment; webcam assessment; webcam SJTs; video resumes.
Response evaluation consistency: There is a growing trend to use Big Data analytics and automated techniques for scoring complex constructed responses. Automated scoring is then regarded as a supplement (or even as alternative) to calibrated judgment, although systematic research is needed.	• To what extent does automated scoring of constructed responses (e.g., written, audio, audiovisual) converge with calibrated judgment? • How does response evaluation consistency interact with response format choice? How to ensure response evaluation consistency for constructed response formats? • What is the relative importance and interplay of stimulus presentation consistency and response evaluation consistency as drivers of subgroup differences and applicant perceptions?	Automated scoring technologies of constructed responses (written essays, scraping of social media content, simulations, work samples); text analytics; social sensing (aka social signal processing).
Information source: Because of response distortion in self-reports, there is increasing research attention to the use of other-reports in the personality domain. So far, research has mainly focused on their validity for predicting job performance.	• What is the validity of other-report ratings for criteria other than job performance (e.g., withdrawal, turnover, subgroup differences, applicant reactions)? • What are the respective validities of using ratings of supervisors, peers, and subordinates? • What are the different constructs underlying self and other ratings? Which theories (e.g., Leising et al., 2013; Vazire, 2010) can help in further identifying their respective "bright" and "blind" spots? • Do applicants perceive other-reports of personality more favorably than self-reports?	Written reference checks; structured telephone-based reference checks; peer assessment in developmental AC exercises.

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5 (*continued*)

Trends in selection research and practice	Research questions	Practical applications
Instructions: There have been calls to make the constructs to be assessed transparent to candidates. So far, research has shown this is a mixed blessing. Many instruments also ask to provide information on past behavior, but a specified vs. unspecified time period for such past behavior questions has not been studied.	• Does the lower validity for transparency instructions depend on the type of construct (personality-like vs. ability-like)? • How do transparency instructions affect construct saturation/unintended variance? • What are the effects of transparency instructions on subgroup differences? • Does specifying a time frame in past behavior queries affect criterion-related validity, subgroup differences, and construct saturation?	Transparent ACs; transparent interviews. Biodata, references, and interviews: Time frame specification in past behavior.

Note. CAT = Computer Adaptive Testing.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

vided with the actual webcam responses but with transcripts of those responses.

The steps and example above illustrate how a modular approach unpacks a selection procedure into separate components and permits flexibly adjusting these components (i.e., stimulus format, response format, and response evaluation consistency) until multiple criteria (i.e., validity, applicant perceptions, and subgroup differences) are satisfied.

Finally, note that regarding step three in this process, a predictor methodology map might serve as a handy tool to visualize the many options that result from crossing facets of two or more predictor method factors that are hypothesized to be relevant for solving the problem. Figure 1 shows an example of a two-dimensional predictor methodology map with 15 cells by crossing stimulus format (five categories) with stimulus presentation consistency (three categories). For ease of presentation, we left out the remote interaction category. Each cell represents a potential predictor method. A striking conclusion is that popular predictor methods represent only the tip of the iceberg because the map reveals various hybrid selection procedures which organizations might experiment with.

Modularity and Selection: Future Research

First, we welcome endeavors that enlarge the seven-factor framework that we used for conceptualizing and reviewing selection procedures. Because selection occurs in many diverse ways we might not have captured all relevant aspects. We reiterate that the facets included are not necessarily mutually exclusive categories for any one procedure and that we are not covering every variation. For example, we left out physiological responses (e.g., eye movements) or response/latency times (Uhlmann et al., 2013). If such measures were to become mainstream in selection, they could be incorporated in future reviews. In a similar vein, we repeat that the level of granularity is deliberately broad. We decided on broad facets because this keeps everything operationally feasible and increases the applicability of the facets across a variety of selection procedures.

Second, this article focused on validity, construct saturation, subgroup differences, and applicant perceptions as selection outcomes. Future studies should extend our approach to understand how method factor choices (e.g., variations in stimulus and response format) make selection tools more susceptible to faking, retesting, and coaching. Prior research mainly compared selection procedures globally in terms of their susceptibility to these effects.

Third, our review reveals a series of future research recommendations regarding specific predictor method factors, as summarized in Table 5. Regarding stimulus presentation consistency, we should examine the effects of adaptive stimuli (e.g., branched SJTs, adaptive simulations) on validity and subgroup differences. On the basis of relevant theories (e.g., item generation theory, trait activation theory) trade-offs need to be found between adaptive formats and ensuring construct equivalence. In terms of response evaluation consistency, a key future challenge deals with establishing convergence between automated and judgmental/calibrated scoring. This applies to automated scoring of texts (e.g., social media content) and nonverbal behaviors (aka "social sensing"; see Schmid Mast, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015).

Regarding contextualization, a critical omission in prior research was the lack of attention to its effects on cognitive saturation. One perspective (Hembree, 1992) posits that adding a realistic context provides cues for recall, thereby activating real-world knowledge and experiences and thus more efficient information processing. According to another perspective, adding context increases cognitive saturation because the contextual information complicates the formation of meaningful representations (Kintsch, 1988). Intriguingly, research in educational psychology shows support for both perspectives because adding context to mathematics and science problems can be both good and bad (e.g., Fyfe, McNeil, Son, & Goldstone, 2014) because there is consensus for a "concreteness fading principle." This principle goes beyond the abstract (generic) versus concrete (contextualized) debate and refers to the fact that material can be presented more generically as people's experience and knowledge increase. So, we recommend that selection researchers factor in candidates' background (e.g., education, experience) when examining contextualization effects on cognitive saturation and subgroup differences.

Related to information source, recent theorizing (e.g., Leising, Ostrovski, & Zimmermann, 2013; Vazire, 2010) should guide future studies in investigating why and when other-reports increase prediction. Hence, the "bright" and "blind" spots of different information sources might be identified. We also need to go beyond job performance as criterion by including withdrawal, turnover, subgroup differences, and applicant reactions.

Regarding instructions (weak vs. strong), future studies should scrutinize whether their effects depend on the construct (personality-like vs. ability-like). To assess "personality-like" constructs, providing cues might make the situation stronger and reduce individual differences in behavior, whereas for "ability-like" constructs cues might ensure that relevant behaviors are displayed, and thus enhance measurement accuracy.

Our review also showed that we have primarily knowledge about the isolated ("local") impact of each method factor. The next step consists of understanding joint effects between components. As attested by the product design literature, some components interact (aka "functional dependencies"), whereas others work independently (Schilling, 2000; Ulrich & Eppinger, 2004). So far, knowledge about interactive effects of predictor method factors is scarce (for exceptions, see Funke & Schuler, 1998; Kanning et al., 2006). Practically, research on joint effects is important for determining when factors work synergistically or antagonistically and for trade-offs between method factor choices. Our review revealed that such synergetic effects might be expected by aligning stimulus with response format because the results of these two factors markedly converged. According to media richness theory (Daft & Lengel, 1984; Potosky, 2008), there should also be a match between stimulus and response format. Thus, our review calls for integrative research on both formats and highlights that investments in higher-level stimuli (e.g., avatars) should be accompanied by similar response format levels. This is especially relevant for interpersonal constructs.

Finally, it might be useful to apply a modular approach to criterion measurement because many method factors seem relevant in criterion measurement. Examples include response evaluation consistency (e.g., rater aids, training), information source (e.g., multisource feedback), response format (e.g., rating forms, narrative formats), instructions (e.g., performance appraisal purpose), contextualization (e.g., generic vs. frame-of-reference rating scales, Hoffman et al., 2012) or stimulus evaluation consistency (e.g., fixed vs. adaptive rating scales, Borman et al., 2001). An intriguing albeit untested hypothesis is for validity to increase when there is a match between criterion and predictor method factor choices.

References

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–153. http://dx.doi.org/10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99,* 535–545. http://dx.doi.org/10.1037/a0035788

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435–442. http://dx.doi.org/10.1037/0021-9010.93.2.435

Baldwin, C. Y., & Clark, K. B. (2000). *Design Rules.: Vol. 1. The Power of Modularity*. Cambridge, MA: MIT Press.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223–235. http://dx.doi.org/10.1111/j.1468-2389.2006.00345.x

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60,* 837–874. http://dx.doi.org/10.1111/j.1744-6570.2007.00093.x

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89,* 150–157. http://dx.doi.org/10.1037/0021-9010.89.1.150

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology, 66,* 91–126. http://dx.doi.org/10.1111/peps.12007

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

Brummel, B. J., Rupp, D. E., & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62,* 137–170. http://dx.doi.org/10.1111/j.1744-6570.2008.01132.x

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). New York, NY: Routledge.

Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests: A theoretical introduction and practical compendium*. Champaign, IL: University of Illinois Press.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159. http://dx.doi.org/10.1037/0021-9010.82.1.143

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology, 88,* 944–953. http://dx.doi.org/10.1037/0021-9010.88.5.944

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416. http://dx.doi.org/10.1016/0010-0285(85)90014-3

Christensen, C. M. (2001). The past and future of competitive advantage. *MIT Sloan Management Review, 42,* 105–109.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgement tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83–117. http://dx.doi.org/10.1111/j.1744-6570.2009.01163.x

Christiansen, N. D., Burns, G., & Montgomery, G. E. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance, 18,* 267–307. http://dx.doi.org/10.1207/s15327043hup1803_4

Clauser, B. E., Kane, M. T., & Swanson, D. (2002). Validity issues for performance-based tests scored with computer automated scoring systems. *Applied Measurement in Education, 15,* 413–432. http://dx.doi.org/10.1207/S15324818AME1504_05

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122. http://dx.doi.org/10.1037/a0021212

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13,* 485–506. http://dx.doi.org/10.1023/A:1022914803347

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organization design. *Research in Organizational Behavior, 6,* 191–233.

DeShon, R. P., Smith, M. R., Chan, D., & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83,* 438–451. http://dx.doi.org/10.1037/0021-9010.83.3.438

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92,* 794–801. http://dx.doi.org/10.1037/0021-9010.92.3.794

Elfenbein, H. A., Barsade, S. G., & Eisenkraft, N. (2015). The social perception of emotional abilities: Expanding what we know about observer ratings of emotional intelligence. *Emotion, 15,* 17–34. http://dx.doi.org/10.1037/a0038436

Fetzer, M., & Tuzinksi, K. (2014). *Simulations for personnel selection*. New York, NY: Springer.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego, CA: Academic Press.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6,* 115–123. http://dx.doi.org/10.1111/1468-2389.00080

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review, 26,* 9–25. http://dx.doi.org/10.1007/s10648-014-9249-3

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511. http://dx.doi.org/10.1037/0021-9010.72.3.493

Gershenson, J. K., Prasad, G. J., & Zhang, Y. (2003). Product modularity: Definitions and benefits. *Journal of Engineering Design, 14,* 295–331. http://dx.doi.org/10.1080/0954482031000091068

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology, 77,* 746–761. http://dx.doi.org/10.1037/0022-3514.77.4.746

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18,* 694–734.

Hammond, K. R. (2010). Intuition, No! Quasirationality, yes! *Psychological Inquiry, 21,* 327–337. http://dx.doi.org/10.1080/1047840X.2010.521483

Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially-available aptitude tests. *Journal of Applied Psychology, 77,* 298–308. http://dx.doi.org/10.1037/0021-9010.77.3.298

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639–684. http://dx.doi.org/10.1111/j.1744-6570.2004.00003.x

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive-ability testing? *American Psychologist, 47,* 1083–1101. http://dx.doi.org/10.1037/0003-066X.47.9.1083

Hembree, R. (1992). Experiments and relational studies in problem solving: A meta-analysis. *Journal for Research in Mathematics Education, 23,* 242–273. http://dx.doi.org/10.2307/749120

Hoffman, B. J., Gorman, A., Atchley, E. K., Blair, C., Meriac, J., & Overstreet, B. (2012). Evidence for the effectiveness of an alternative multi-source feedback measurement methodology. *Personnel Psychology, 65,* 531–563. http://dx.doi.org/10.1111/j.1744-6570.2012.01252.x

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., Lance, C. E., & Sutton, A. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology, 100,* 1143–1168. http://dx.doi.org/10.1037/a0038707

Hogan, R., & Shelton, D. (1998). A socioanalytic perspective on job performance. *Human Performance, 11,* 129–144. http://dx.doi.org/10.1080/08959285.1998.9668028

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice. The effects of frame-of-reference and pre-test information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13,* 75–86. http://dx.doi.org/10.1111/j.0965-075X.2005.00301.x

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984). Revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184–190. http://dx.doi.org/10.1037/0021-9010.79.2.184

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.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21,* 264–276. http://dx.doi.org/10.1111/ijsa.12036

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in interview evaluations. *Journal of Applied Psychology, 83,* 179–189. http://dx.doi.org/10.1037/0021-9010.83.2.179

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321–368. http://dx.doi.org/10.1006/obhd.1993.1015

Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2016). Transparency of assessment centers: Lower criterion-related validity but greater opportunity to perform? *Personnel Psychology, 69,* 467–497. http://dx.doi.org/10.1111/peps.12105

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation and test development*. Mahwah, NJ: Erlbaum.

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology, 98,* 326–341. http://dx.doi.org/10.1037/a0031257

Johns, G. (2006). The essential impact of context on organizational behavior. *The Academy of Management Review, 31,* 386–408. http://dx.doi.org/10.5465/AMR.2006.20208687

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment, 22,* 168–176. http://dx.doi.org/10.1027/1015-5759.22.3.168

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182. http://dx.doi.org/10.1037/0033-295X.95.2.163

Klehe, U. C., König, C. J., Kleinmann, M., Richter, G. M., & Melchers, K. G. (2008). Transparency in structured selection interviews: Consequences for construct and criterion-related validity. *Human Performance, 21,* 107–137.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., . . . Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11,* 121–137. http://dx.doi.org/10.1207/s15324818ame1102_1

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

role of candidates' ability to identify criteria. *Organizational Psychology Review, 1,* 128–146. http://dx.doi.org/10.1177/2041386610387000

Kluemper, D. H., McLarty, B. D., & Bing, M. N. (2015). Acquaintance ratings of the Big Five personality traits: Incremental validity beyond and interactive effects with self-reports in the prediction of workplace deviance. *Journal of Applied Psychology, 100,* 237–248. http://dx.doi.org/10.1037/a0037810

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100,* 128–161. http://dx.doi.org/10.1037/a0036734

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100,* 399–416. http://dx.doi.org/10.1037/a0037674

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96,* 202–210. http://dx.doi.org/10.1037/a0020375

Le, H. (2013, April). *Proposing a taxonomy for simulation tests*. Poster presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are we talking about the same person here?" Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. *Social Psychological and Personality Science, 4,* 468–474. http://dx.doi.org/10.1177/1948550612462414

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67,* 241–293. http://dx.doi.org/10.1111/peps.12052

Levashina, J., Morgeson, F. P., & Campion, M. A. (2012). Tell me some more: Exploring how item verifiability and cognitive ability influence responses to biodata questions in a high-stakes selection context. *Personnel Psychology, 65,* 359–383. http://dx.doi.org/10.1111/j.1744-6570.2012.01245.x

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41,* 1604–1627. http://dx.doi.org/10.1177/0149206312463941

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181–1188. http://dx.doi.org/10.1037/0021-9010.91.5.1181

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,* 1043–1055. http://dx.doi.org/10.1037/0021-9010.92.4.1043

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology, 100,* 1169–1188. http://dx.doi.org/10.1037/apl0000004

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Bingley, United Kingdom: JAI Press. http://dx.doi.org/10.1108/S0742-7301(2009)0000028006

Lord, R. G., & Maher, K. J. (1990). Alternative information processing models and their implication for theory, research, and practice. *The Academy of Management Review, 15,* 9–28.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

MacCann, C., Wang, P., Matthews, G., & Roberts, R. D. (2010). Examining self-report versus other reports in a situational judgment test of emotional abilities. *Journal of Research in Personality, 44,* 673–676. http://dx.doi.org/10.1016/j.jrp.2010.08.009

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91. http://dx.doi.org/10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96,* 327–336. http://dx.doi.org/10.1037/a0021983

McFarland, L. A., & Ployhart, R. E. (2015). Social media in organizations: A theoretical framework to guide research and practice. *Journal of Applied Psychology, 100,* 1653–1677. http://dx.doi.org/10.1037/a0039244

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64,* 53–87. http://dx.doi.org/10.1111/j.1744-6570.2010.01202.x

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36,* 121–140. http://dx.doi.org/10.1177/0149206309349309

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246–268. http://dx.doi.org/10.1037/0033-295X.102.2.246

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a SJT. *Journal of Applied Psychology, 95,* 321–333. http://dx.doi.org/10.1037/a0017975

Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96,* 762–773. http://dx.doi.org/10.1037/a0021832

Oswald, F. L., & Putka, D. J. (in press). Statistical methods for big data. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge.

Outtz, J. L. (1998). Testing medium, validity and test performance. In M. Hakel (Ed.), *Beyond multiple choice* (pp. 41–57). Mahwah, NJ: Erlbaum.

Parshall, C. G., Davey, T., & Pashley, P. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Norwell, MA: Kluwer Academic Publishers. http://dx.doi.org/10.1007/0-306-47531-6_7

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *The Academy of Management Review, 33,* 629–648. http://dx.doi.org/10.5465/AMR.2008.32465704

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85,* 880–887. http://dx.doi.org/10.1037/0021-9010.85.6.880

Robie, C., Schmit, M. J., Ryan, A. M., & Zickar, M. J. (2000). Effects of item context specificity on the measurement equivalence of a personality inventory. *Organizational Research Methods, 3,* 348–365. http://dx.doi.org/10.1177/109442810034003

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163–184. http://dx.doi.org/10.1111/j.1745-3984.2003.tb01102.x

Roth, P. L., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of Black-White differences in overall and exercise scores. *Personnel Psychology, 61,* 637–661. http://dx.doi.org/10.1111/j.1744-6570.2008.00125.x

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., Jr., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87,* 369–376. http://dx.doi.org/10.1037/0021-9010.87.2.369

Rousseau, D. M. (2006). Is there such a thing as evidence-based management? *The Academy of Management Review, 31,* 256–269. http://dx.doi.org/10.5465/AMR.2006.20208679

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number right scoring, and test-taking strategy. *Journal of Educational Measurement, 14,* 15–22. http://dx.doi.org/10.1111/j.1745-3984.1977.tb00024.x

Ryan, A. M. (2001). Explaining the Black-White test score gap: The role of test perceptions. *Human Performance, 14,* 45–75. http://dx.doi.org/10.1207/S15327043HUP1401_04

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65,* 693–717. http://dx.doi.org/10.1146/annurev-psych-010213-115134

Sanchez, R., & Mahoney, J. T. (1996). Modularity, flexibility and knowledge management in product and organizational design. *Strategic Management Journal, 17,* 63–76. http://dx.doi.org/10.1002/smj.4250171107

Schilling, M. A. (2000). Toward a general modular systems theory and its application to interfirm product modularity. *The Academy of Management Review, 25,* 312–334.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87,* 735–746. http://dx.doi.org/10.1037/0021-9010.87.4.735

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science, 24,* 87–92. http://dx.doi.org/10.1177/0963721414560811

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, S. L. (1995). Frame-of-reference effects on personality scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607–620. http://dx.doi.org/10.1037/0021-9010.80.5.607

Schmitt, N., & Quinn, A. (2010). Reductions in measured subgroup mean differences: What is possible? In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 425–451). New York, NY: Routledge.

Sears, G., Zhang, H. H., Wiesner, W. D., Hackett, R., & Yuan, Y. (2013). A comparative assessment of videoconference and face-to-face employment interviews. *Management Decision, 51,* 1733–1752. http://dx.doi.org/10.1108/MD-09-2012-0642

Schaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65,* 445–494. http://dx.doi.org/10.1111/j.1744-6570.2012.01250.x

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Erlbaum.

Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions transparent on relations with typical performance predictors. *Human Performance, 20,* 187–203. http://dx.doi.org/10.1080/08959280701332992

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285. http://dx.doi.org/10.1207/s15516709cog1202_4

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behavior employment interview questions. *Journal of Occupational and Organizational Psychology, 75,* 277–294. http://dx.doi.org/10.1348/096317902320369712

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500–517. http://dx.doi.org/10.1037/0021-9010.88.3.500

Thornton, G. C., III. (1992). *Assessment centers in human resource management*. Reading, MA: Addison Wesley.

Uhlmann, E., Leavitt, K., Menges, J., Koopman, J., Howe, M., & Johnson, R. (2012). Getting explicit about the implicit: A taxonomy of implicit measures and guide for their use in organizational research. *Organizational Research Methods, 15,* 553–601. http://dx.doi.org/10.1177/1094428112442750

Ulrich, K. T., & Eppinger, S. D. (2004). *Product design and development*. New York, NY: McGraw-Hill/Irwin.

Van Iddekinge, C. H., Raymark, P. H., Roth, P. L., & Payne, H. S. (2006). Comparing the psychometric characteristics of ratings of face-to-face and videotaped structured interviews. *International Journal of Selection and Assessment, 14,* 347–359. http://dx.doi.org/10.1111/j.1468-2389.2006.00356.x

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98,* 281–300. http://dx.doi.org/10.1037/a0017908

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22,* 269–286. http://dx.doi.org/10.1177/001316446202200203

Waung, M., Hymes, R. W., & Beatty, J. E. (2014). The effects of video and paper resumes on assessments of personality, applied social skills, mental capability, and resume outcomes. *Basic and Applied Social Psychology, 36,* 238–251. http://dx.doi.org/10.1080/01973533.2014.894477

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21,* 291–309. http://dx.doi.org/10.1080/08959280802137820

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York, NY: Academic Press.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29,* 231–258. http://dx.doi.org/10.1177/014920630302900206

Zimmerman, R. D., Triana, M. C., & Barrick, M. R. (2010). Predictive criterion-related validity of observer-ratings of personality and job-related competencies using multiple raters and multiple performance criteria. *Human Performance, 23,* 361–378. http://dx.doi.org/10.1080/08959285.2010.501049

Received July 7, 2015
Revision received July 21, 2016
Accepted July 25, 2016 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.