3-2023

Multiple, speeded assessments under scrutiny: Underlying theory, design considerations, reliability, and validity

Christoph N. HERDE
Singapore Management University, cherde@smu.edu.sg

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

Citation

HERDE, Christoph N. and LIEVENS, Filip. Multiple, speeded assessments under scrutiny: Underlying theory, design considerations, reliability, and validity. (2023). *Journal of Applied Psychology*. 108, (3), 351-373.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/7032

Multiple, Speeded Assessments Under Scrutiny: Underlying Theory, Design Considerations, Reliability, and Validity

Christoph N. Herde

Singapore Management University

Filip Lievens

Singapore Management University

Ghent University, Belgium

This is an unedited manuscript accepted for publication in *Journal of Applied Psychology*. The manuscript will undergo copyediting, typesetting, and review of resulting proof before it is published in its final form. Please cite as:

Herde, C.N., & Lievens, F. (in press). Multiple, Speeded Assessments Under Scrutiny: Underlying Theory, Design Considerations, Reliability, and Validity. *Journal of Applied Psychology*.

Both authors contributed equally to this paper. This research was supported by a research grant from Research Foundation - Flanders (FWO), application number: G092512N and by the Ministry of Education, Singapore under its Academic Research Funding Tier 2 (MOE2019-T2-1-191). We are indebted to Hudson Belgium for their contribution to the development of role-plays, role-player and assessor training, the provision of measures of cognitive ability and personality as well as professional role-players/assessors. Especially, we thank Ellen Volckaert, Amelie Vrijdags as well as Myrjam Van de Vijver for their various contributions to this project. We further thank Sofie Ameloot, Robin Boudry, Amos Tai Yong En, Megan Choy Cheng Mun, and Rebecca Tan Soo-Yi for their contribution to the technical preparations of the data collection as well as all participating role-players and assessors. We also thank Paul R. Sackett, John D. Arnold, and Philipp Schäpers for valuable feedback on earlier versions of this paper.

Portions of this paper are based on the doctoral dissertation of Christoph N. Herde and were presented at the 31st Annual Conference of the Society of Industrial and Organizational Psychology (SIOP), the 50th Conference of the German Society for Psychology (DGPs), and the 79th Annual Meeting of the Academy of Management (AOM).

Address of correspondence: Christoph N. Herde, c.herde@gmx.net

Abstract

Recently, multiple, speeded assessments (e.g., “speeded” or “flash” role-plays) have made rapid inroads into the selection domain. So far, however, the conceptual underpinning and empirical evidence related to these short, fast-paced assessment approaches has been lacking. This raises questions whether these speeded assessments can serve as reliable and valid indicators of future performance. This paper uses the notions of stimulus and response domain sampling to conceptualize multiple, speeded behavioral job simulations as a hybrid of established simulation-based selection methods. Next, we draw upon the thin slices of behavior paradigm to theorize about the quality of ratings made in multiple, speeded behavioral simulations. In two studies, various assessor pools assessed a sample of 96 MBA students in eighteen 3-minute role-plays designed to capture situations in the junior management domain. At the individual speeded role-play level, reliability and validity were not ensured. Yet, aggregated across all assessors’ ratings of all speeded role-plays, the overall score for predicting future performance was high (.54). Validities remained high when assessors evaluated only the first minute (vs. full 3 minutes) or received only a control training (vs. traditional assessor training). Aggregating ratings of performance in multiple, heterogeneous situations that elicit a variety of domain-relevant behavior emerged as key requirement to obtain adequate domain coverage, capture both ability and personality (extraversion and agreeableness), and achieve substantial validities. Overall, these results show the importance of the stimulus and response domain sampling logic and send a strong warning to using “single” speeded behavioral simulations in practice.

Keywords: personnel selection; multiple, speeded assessments; behavioral job simulations; thin slices; generalizability theory; stimulus domain and response domain sampling

Running Head: MULTIPLE, SPEEDED ASSESSMENTS UNDER SCRUTINY

Multiple, Speeded Assessments Under Scrutiny:

Underlying Theory, Design Considerations, Reliability, and Validity

Today's society and business are characterized by a culture of speed and efficiency. For example, people increasingly participate in speed dates, speed networking with suppliers and clients has gained in popularity, and investors make investment decisions based on short pitches of entrepreneurs. The magic words "fast" and "short" have also entered the personnel selection arena. To respond to calls for brief, fast-paced, and more engaging assessment processes that represent today's hectic and fragmented work life (Liff, 2017; Mullenweg, 2014; Pinchback, 2017; Pinsight, 2019), selection practitioners have added new assessment approaches to their portfolio under the umbrella term of multiple, speeded assessments (Herde & Lievens, 2020). Although there exists some variability in practice, three characteristics seem to define multiple, speeded assessments. First, in multiple, speeded assessments, people participate in a large number of short "sessions" (typically around 3 minutes). Second, multiple behavioral job simulations wherein participants face a variety of job situations constitute an essential part of these brief sessions. As noted below, due to these two features, multiple, speeded assessments lie somewhere between Assessment Centers (ACs) and Situational Judgment Tests (SJTs). Third, the evaluation process is streamlined, with assessors usually providing only a single rating per session. Examples of multiple, speeded assessments[1] are short "flash" simulations (Byham, 2016), brief "auditioning" performances (Pinchback, 2017), and short webcam role-plays (Pinsight, 2018).

Apart from practical cost and efficiency considerations, researchers have also promoted the use of a larger number of short and fast assessments because they might permit to more comprehensively sample the diversity of situations of a performance domain

[1] Given these characteristics, the sole use of "flash" interviews (Needleman, 2007) without a behavioral simulation component is not regarded as falling under speeded assessments because interviews typically assess reported behavior (with the exception of the assessment of oral communication).

(Brannick, 2008; Lievens, 2008). For example, in the interpersonal domain, one might confront participants with ten short situations that cover an array of interpersonal challenges (e.g., handing a conflict, communicating bad news, building trust, dealing with complaints; Klein, DeRouin, & Salas, 2006).

So far, empirical research on multiple, speeded behavioral job simulations is scarce. Motowidlo, Hooper and Jackson (2006) investigated relations between behavior in eight short role-plays, implicit trait policies, and personality. However, little insight was obtained about the role-plays because they served as criterion measures. Recently, Ingold, Dönni, and Lievens (2018) investigated the ratings that assessors made on the basis of limited information (i.e., snap judgments) in four simulations. In this study, one group of assessors saw full-length performances, whereas another group watched only the first minutes. On the positive side, assessors' impressions of participants in the first minutes converged reasonably with the other assessors' final ratings and reflected information about some personality traits. On the negative side, there were large idiosyncrasies in assessors' ratings and their criterion-related validity did not reach statistical significance. Similar research was conducted on initial impressions in the rapport-building phase in employment interviews (e.g., Barrick et al., 2012; Barrick, Swider, & Stewart, 2010; Swider, Barrick, & Harris, 2016). Although these prior studies did not deal with multiple, speeded assessments, they attest to the emerging research interest in ratings made based on limited information and suggest that the reliability and validity of such ratings deserve closer scrutiny.

Indeed, many conceptual and empirical questions regarding the growing use of multiple, speeded assessments have been raised (Herde & Lievens, 2020): What are the theoretical fundaments underlying a selection approach that assesses participants' performance in multiple behavioral simulations? Do ratings based upon such multiple, short behavioral simulations reliably assess participants' performance? Do they reveal information

about participants' individual differences like cognitive ability and personality? Do they predict job-related performance (over and above traditional selection procedures)? As compared to such traditional procedures, do multiple, speeded simulations outweigh their design and implementation costs in terms of utility? Can design considerations be identified that improve the quality of multiple, speeded simulations? To the best of our knowledge, there are no empirical investigations that provide answers to these pressing questions. Hence, multiple, speeded assessments seem to be practices that so far have moved ahead of research and rigorous empirical scrutiny.

Therefore, this paper aims to scrutinize the multiple, speeded assessment approach. Our focus is on one type, namely multiple, speeded role-plays. This paper contributes to assessment and selection in at least four ways. At a theoretical level, we introduce the notions of stimulus domain sampling and response domain sampling to conceptualize multiple, speeded behavioral simulations as an integration between SJTs and ACs. Second, to develop hypotheses about how assessors observe and evaluate participants in short and fast behavioral simulations we draw on the "thin slices" of behavior paradigm in social and personality psychology (e.g., Ambady, Bernieri, & Richeson, 2000). Third, we go beyond previous studies on the role of early impressions in behavioral simulations (Ingold et al., 2018) and interviews (Barrick et al., 2010, 2012; Swider et al., 2016) by presenting empirical evidence on the reliability and validity of ratings based on multiple, speeded simulations. Finally, to close the gap between practice and research on multiple, speeded simulations we examine not only whether they work but also identify design considerations under which they work best.

STUDY BACKGROUND

Domain Sampling: Stimuli and Responses

Domain sampling serves as a capstone underlying simulation-based selection methods such as SJTs and ACs. Domain sampling refers to the extent to which the selection method

covers the criterion/content domain to be predicted. To clarify the differences between multiple, speeded behavioral job simulations and classic simulation-based selection methods, we make a further differentiation between stimulus domain sampling and response domain sampling (see also Cronbach, 1984; Guion, 1978; Ryan & Greguras, 1998; Sackett, 1987)[2]. *Stimulus domain sampling* denotes the degree to which the simulation covers the large variety of situations relevant to the domain. For example, to cover the sales management domain, participants might be presented with one key domain-relevant situation: a meeting with a potentially new client. Such an approach scores lower on stimulus domain sampling. Conversely, in an approach higher on stimulus domain sampling, participants might be presented with several domain-relevant situations such as a meeting with a new client, a meeting with an existent client, a meeting with their supervisor, a sales presentation, a coaching session of their sales force, etc. *Response domain sampling* refers to the extent to which participants' responses to simulations cover the large variety of responses relevant to the domain. Related to the sales management example above, a simulation that scores low on response domain sampling presents participants with only a limited, predetermined list of multiple-choice responses to choose from. Conversely, higher levels of response domain sampling are obtained when participants can construct/enact responses themselves (using their full repertoire of verbal and nonverbal behavior) and interactively build upon prior responses.

In traditional AC exercises like group discussions or role-plays, only a few longer behavioral job simulations are used. Hence, these longer simulations can typically present only a limited set of the full domain of situations. As noted by Ryan and Greguras (1998), this less comprehensive approach to stimulus domain sampling runs the risk of being inadequate and lacking representativeness (e.g., some domain situations might not be sampled). Yet, in

[2] Stimulus/response domain sampling is related to stimulus/response fidelity. Yet, there are also differences. Fidelity refers to the veridicality of the simulation (Goldstein Zedeck, & Schneider, 1993); domain sampling denotes the coverage of the domain by the simulation.

these classic AC exercises, response domain sampling is typically comprehensive because participants show behavior in responding to the situations. The behavior is also interactive, thereby increasing the variety of responses shown. Thus, classic AC exercises score relatively high on response domain sampling because participants can enact any kind of behavior that might capture more broadly the universe of possible behavioral responses to the situation(s).

Prototypical SJTs score the opposite as AC exercises on stimulus domain sampling and response domain sampling. That is, the large set of SJT items enable more comprehensive stimulus domain sampling. Yet, typical SJTs provide only a sketchy insight into response behavior because they present a restricted set of predetermined responses.

Multiple, Speeded Behavioral Simulations: Definition and Characteristics

Using the notions of the stimulus and response domain sampling, the recent trend of multiple, speeded behavioral job simulations can be conceptualized as an integration (hybrid) between SJTs and ACs (Herde & Lievens, 2020)[3]. Similar to SJTs, multiple, speeded behavioral simulations aim to systematically cover a large set of situations, thus adopting a comprehensive approach to stimulus domain sampling. However, this does not come at the expense of lower response domain sampling because multiple, speeded simulations seek to capture a variety of interactive behavior in ongoing situations. So, this goal is similar to ACs. Yet, to obtain a comprehensive sampling of the stimulus domain than in ACs, multiple, speeded simulations include a larger number as well as a possibly greater diversity of situations than ACs. Therefore, these situations also have a shorter time span (typically around three minutes). For instance, as noted above, related to the sales management domain, one might use 10 different simulations in which participants face different types of clients,

[3] Although multiple, speeded assessment is a relatively recent trend, there exist similar practices in other fields. In healthcare, Objective Structured Clinical Examinations (OSCE; e.g., Brannick, Erol-Korkmaz, & Prewett, 2011; Harden, Stevenson, Downie, & Wilson, 1975) and Multiple Mini Interviews (MMI; Eva, Rosenfeld, Reiter, & Norman, 2004; Knorr & Hissbach, 2014) are used for certification and admission. OSCEs confront participants with many simulated patients, whereas in MMIs multiple short interviews/simulations are used. In these "stations" (sometimes up to 40; Patrício, Julião, Fareleira, & Carneiro, 2013), participants are evaluated by trained assessors.

subordinates, sales situations, etc. In sum, multiple, speeded assessments refer to the use of a large set of behavioral job simulations with a short duration (typically 3 minutes) and streamlined rating process (one overall rating). The number of simulations is contingent upon the breadth of the domain to be sampled.

Other features of multiple, speeded simulations might vary (Herde & Lievens, 2020). That is, the simulations can be run in a traditional "brick and mortar" fashion or remotely (e.g., with role-players meeting participants in virtual break out rooms). In addition, the short situations included might be integrated (via a common overarching theme) or remain independent. Note further that multiple, speeded simulations refer to a method and not a construct (Arthur & Villado, 2008). So, they are not restricted to the interpersonal domain but might be designed to sample situations from a variety of performance domains. They are thus also not designed to assess specific constructs. That said, it should be clear that underlying individual differences constructs such as personality and cognitive ability might serve as antecedents of the behavior shown in the multiple, speeded simulations.

HYPOTHESES

The “Thin Slices“ of Behavior Paradigm

Research on the “thin slices” of behavior paradigm in personality and social psychology (e.g., Ambady, 2010; Ambady et al., 2000) provides a useful theoretical fundament for shedding light on assessment approaches that require assessors to observe and evaluate participants in fast and short simulations. Research adhering to this paradigm seeks to investigate how people form judgments about strangers, their individual differences, and performance. Typically, untrained people (“judges”) are asked to rate others (“strangers”) on the basis of little information. This usually comes from dynamic excerpts from strangers’ behavioral stream that last between several seconds to not more than five minutes (Ambady et al., 2000; Ambady & Rosenthal, 1992; Connelly & Ones, 2010). In a seminal study,

Borkenau, Mauer, Riemann, Spinath, and Angleitner (2004) asked judges to watch videotaped strangers in short unstructured situations (e.g., introductions, story-telling, convincing a neighbor to lower the radio volume) and to rate them on personality and intelligence. Other studies extended this paradigm by providing judges with minimal information such as tone of voice clips, short (and sometimes muted) videos and by asking judges to evaluate not only personality/intelligence but also general competence or performance (e.g., Ambady, Hogan, Spencer, & Rosenthal, 1993, Ambady, Krabbenhoft, & Hogan, 2006; Ambady & Rosenthal, 1993; Stallings & Spencer, 1967; Todorov, Mandisodza, Goren, & Hall, 2005).

Past thin slices research examined whether such judgments can predict a diverse set of relevant outcomes. One meta-analysis investigated whether judgments of thin slices predict social and clinical outcomes (Ambady & Rosenthal, 1992). Overall, thin slices judgments predicted these outcomes with an average of r = .39. Another meta-analysis (Ambady et al., 2000) extended the evidence to domains as diverse as testosterone levels (r = .20), type and quality of relationships (r = .27), interviewees' performance (r = .27), and job performance of telephone operators, sales managers, and management consultants (r = .39).

To be able to make such valid judgments on thin slices, this strand of research emphasizes the key role of "good" behavioral information, namely *qualitatively different* pieces of information (Ambady et al., 2000; Back & Nestler, 2016; Carney, Colvin, & Hall, 2007; Connelly & Ones, 2010; Murphy, 2005; Murphy et al., 2015; see also Casey et al., 2009; Knorr & Hissbach, 2014; Rushforth, 2007). This can be done by observing strangers' behavior across multiple, qualitatively different situations. This evidence related to good behavioral information matches well with the aforementioned notions of stimulus domain sampling and response domain sampling. Hence, both the "good behavioral information" and the domain sampling notions concur that more valid assessments can be made when ratings are based on multiple, qualitatively different pieces of criterion-relevant information. This is

what multiple, speeded simulations aim for: assessors evaluate participants in multiple, short situations that elicit behavior related to many qualitatively different criterion domain parts, thereby improving criterion coverage and potentially increasing validity. Thus,

Hypothesis 1 (H1): The overall score on multiple, speeded simulations significantly predicts future domain-relevant criterion performance.

The Role of Stimulus Domain Sampling

In this study, the multiple, speeded simulations aim to cover the junior management domain. According to Motowidlo, Dunnette, and Carter (1990), this large and heterogeneous domain covers solving task-oriented problems, dealing with interpersonally-oriented issues, and convincingly communicating the solutions to these task and interpersonal issues. On the basis of the *stimulus domain sampling* logic, this means that a large and heterogeneous set of domain-relevant situations should be sampled via multiple behavioral simulations (role-plays). For example, some role-plays should focus more on task/decision making (e.g., planning) issues, whereas other role-plays should require people to solve interpersonal and communication issues (e.g., dealing with dejected subordinates or irate customers). Still other role-plays should demand a combination of all of these (see Table 2, for an overview).

Generally, we expect the overall score on the basis of multiple, speeded simulations to be related to both cognitive ability and personality. The link with cognitive ability should be evident given that participants must swiftly understand the problems, suggest appropriate solutions and contingency plans, and make decisions. Therefore, efficient, accurate information processing and quick adjustment to new situations are required, which are quintessential to cognitive ability (Cattell, 1971; Schmidt, 2002; Snyderman & Rothman, 1987; see Kuncel, Hezlett, & Ones, 2004; Ones, Dilchert, & Viswesvaran, 2012). Apart from cognitive ability, we also expect the overall simulation score to be related to interpersonal traits like extraversion and agreeableness. We refer to extraversion and agreeableness as

interpersonal traits because they overlap with the affiliation and dominance dimensions of the Interpersonal Circumplex Model (Leising & Bleidorn, 2011; Markey & Markey, 2006; McCrae & Costa, 1989, 1995). Related to extraversion, the role-plays require to enthusiastically approach others, be talkative, and enjoy interpersonal encounters instead of keeping a distance from others and showing reserved body language (McCrae & John, 1992; see also Wilmot, Wanberg, Kammeyer-Mueller, & Ones, 2019). In addition, the role-plays include situations that activate behavior related to cooperation, negotiation, and interpersonal sensitivity, which are indicative of agreeableness (McCrae & John, 1992)[4]. Thus,

Hypothesis 2 (H2): Participants' overall score on multiple, speeded simulations is significantly related to their level of cognitive ability, extraversion, and agreeableness.

The comprehensive nature of stimulus domain sampling and the heterogeneity of the simulations demonstrate that the simulations included are typically not chosen to be interchangeable measures of each other. For the same reasons, one cannot assume that people will perform at the same level in all situations. For example, someone who scores relatively well in most situations might still perform poorly in some situations (e.g., proposing a solution to a coworker conflict, dealing with a crisis). This is because theory and empirical evidence show that people's behavior emerges from the interaction between the person and the simulated job situations (CAPS, Mischel & Shoda, 1995; Bledow & Frese, 2009; Lance, 2008; Lievens, Tett, & Schleicher, 2009; Speer, Christiansen, Goffin, & Goff, 2014).

This conceptual underpinning leads to several implications. Due to the heterogeneity of the domain, the diversity of the situational demands of the simulations, and the variability in people's performances across situations we do not assume a unidimensional latent variable

4 Our expectation that both cognitive ability and personality can be observed in this multiple, speed simulation matches also with thin slices research. Meta-analytic research showed that even static information (i.e., photographs) enables making judgments that relate to ability scores (r = .28, Zebrowitz, Hall, Murphy, & Rhodes, 2002). This finding was corroborated with dynamic behavioral information (Borkenau et al., 2004; Borkenau & Liebler, 1993; Carney et al., 2007; Murphy, 2007; Murphy, Hall, & Colvin, 2003; Reynolds & Gifford, 2001). In some cases, correlations with ability scores rose to r = .43. Furthermore, meta-analyses showed that thin slices judgments correlated with self-ratings on personality (.20 in Ambady et al., 2000; up to .29 in Connolly, Kavanagh, & Viswesvaran, 2007; see also Connelly & Ones, 2000).

as the “single behavior-generating” mechanism across all simulations (see Bledow & Frese, 2009, p.241). Hence, we expect that simulation scores will not correlate highly (but at best moderately or even lowly) with each other. Relatedly, the overall score that people obtain based on multiple, speeded simulations will depend on the kind and variety of the simulations sampled. As a critical implication, this means that removing specific simulations might change what is being measured and predicted by this overall simulation score. In other words, removing particular simulations (or overlooking to include simulations) might reduce domain coverage and be detrimental for what is being measured, thereby underscoring the importance of stimulus domain sampling.

Hence, one broad way of testing the importance of stimulus domain sampling consists of scrutinizing whether the removal of simulations and thus potentially the reduction of domain coverage indeed changes the cognitive and personality loading of the overall multiple, speeded simulation score. Removing simulations can be compared to test developers overlooking to include a set of simulations. Specifically, we expect that when the overall multiple, speeded simulation score is based on simulations that deal with interpersonal and communication issues but does not include simulations that deal with task-related problem solving/decision making issues, domain coverage will be reduced in that the overall multiple, speeded simulation score will be significantly less related to cognitive ability. Conversely, we expect the opposite when this overall simulation score is based on simulations that deal with solving task-related problems and decision making but does not include simulations that deal with interpersonal and communication issues. In that case, domain coverage will be reduced in that the overall score will be significantly less related to the personality traits of extraversion and agreeableness (see also Gonzalez-Mulé, Mount, & Oh, 2014). Thus,

Hypothesis 3a (H3a): The overall score on multiple, speeded simulations is significantly less related to cognitive ability when simulations that tap into task-related problem solving issues are not included in it.

Hypothesis 3b (H3b): The overall score on multiple, speeded simulations is significantly less related to extraversion and agreeableness when simulations that tap into interpersonal and communication issues are not included in it.

The Role of Response Domain Sampling

According to the response domain sampling principle, it is of paramount importance to ensure that participants can show a large variety of domain relevant responses and thus that they are not constrained to choose a limited, predetermined set of responses (e.g., written multiple-choice options). So, this principle emphasizes ratings and predictions from speeded simulations should be based on qualitatively rich data in the form of participants' behavioral responses to ongoing behavior of the interaction partner.

There are several conceptual and empirical arguments that underpin this response domain logic. First, we expect that behavioral responses to relevant situations have a closer correspondence with future criterion behavior. Conversely, MC responses to written scenarios have been linked to people's *procedural knowledge* of the behavior. Such procedural knowledge has been found to be at best a precursor of actual behavior (i.e., people will not always be able to translate their knowledge into behavioral actions; Lievens & Patterson, 2011). As another conceptual reason, a multitude of information is available to assessors because they can observe verbal, nonverbal, and paraverbal behavior, thereby increasing the predictive power of their assessments. Indirect empirical evidence comes from the validities obtained with webcam SJTs with a constructed behavioral response format. Such webcam SJTs can be expected to outperform written SJTs in terms of response domain sampling because people must react swiftly to a video stimulus by enacting their behavior in front of a

webcam instead of checking a multiple-choice box (e.g., Cucina, Su, Busciglio, Harris Thomas, & Thompson Peyton, 2015; Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2019). None of these studies, however, directly compared the validity of multiple-choice vs. behavioral responses. Therefore, we test the role of response domain sampling by comparing the criterion-related validity of candidates' scores based on their behavioral responses to situations (i.e., multiple, speeded simulations) with their scores based on MC SJT responses, while keeping the domain (situations presented) as similar as possible. Thus,

> *Hypothesis 4 (H4): The overall score on multiple, speeded simulations significantly predicts future domain-relevant criterion performance over and above the overall score on MC responses to similar interpersonal situations in SJT format.*

STUDY 1

Methods

Sample

To gather data about a multiple, speeded simulation approach, we collaborated with a European business school that aimed to reinvigorate the assessment/admission procedure of their MBA program. Therefore, multiple, speeded simulations were implemented for developmental purposes: That is, the entire MBA cohort of this business school (Master in Marketing and Financial Management) participated in our study to identify their strengths and weaknesses as junior managers. The sample encompassed 96 participants (51% females, mean age = 23.63, *SD* = 1.85) from 19 different nations (e.g., 67% Belgian, 5% Chinese, 4% Romanian). All had at least one year of work experience. We excluded one participant from the analyses because she did not take part in the multiple, speeded assessments.

We assessed participants' test taking motivation via a scale with four items from Arvey, Strickland, Drauden, and Martin (1990; *1 = strongly disagree; 5 = strongly agree*; internal consistency reliability = .67). Their mean test-taking motivation was high: 3.96 (*SD*

=.50)[5]. Anecdotal evidence supported participants were motivated to perform well and learn about their strengths and weaknesses (e.g., they wore business attire and seemed nervous).

Procedure

Prior to the multiple, speeded simulations, participants completed proctored computer-based measures (cognitive ability, Big Five, SJT). The multiple, speeded simulations took place in a large hall and lasted 90 minutes: Participants completed 18 different role-plays in which they interacted with 18 different role-players. In the hall, a circle was formed by desks. One role-player was sitting at each desk. Each participant was assigned a different desk number. A bell signaled role-players to start a role-play. After three minutes, another audio signal prompted the role-player to finish the conversation so that participants could move to the next desk where they met a different role-player who introduced a different issue. This carousel procedure was repeated until all participants had completed all 18 role-plays. Role-players typically remained at the same desk and played the same role-play again[6]. Participants' performance in each role-play was rated by the role-player (who thus also served as assessor, see below) immediately after each role-play. Afterwards, participants received feedback reports about their performance in the multiple, speeded simulations. Seven months later, MBA supervisors (instructors) rated participants' performance to provide criterion data.

Measures

Cognitive ability. To assess cognitive ability, we used a traditional matrix-type figural reasoning test (Bogaert, Trbovic, & Van Keer, 2005). Various studies showed that matrix-type figural reasoning tests are good indicators of general mental ability (Jensen, 1998). This figural reasoning test confronted participants with 40 items in 20 minutes. The test manual also supported this test's psychometric properties in terms of internal consistency reliability

[5] n = 49; due to operational problems, test motivation data of 47 people were not gathered.
[6] To reduce fatigue, role-players played only two different role-plays, took breaks, and were regularly replaced by other role-players.

(Cronbach's alpha = .91), split-half reliability (Spearman-Brown formula = .94), and correlations with the Advanced Progressive Matrices (Raven, 1958) of r = .52.

Big Five personality. We assessed personality with the Business Attitudes Questionnaire (BAQ; Vrijdags, Bogaert, Trbovic, & Van Keer, 2014). Each item of the BAQ asks participants to indicate agreement with a work-related statement on a 5-point Likert scale (*1 = totally disagree; 5 = totally agree*). It comprises a total of 150 items, with 6 items each building up one of 25 scales. It was certified by the British Psychological Society (BPS). We used participants' summed scores on the Big Five scales. The test manual reports adequate internal consistency reliabilities ($.91 \leq \alpha \leq .94$), convergent validities with other personality inventories, and criterion-related validities.

Multiple, speeded simulations. We developed 18 different role-plays to sample relevant situations in the junior management domain. To derive the content of these role-plays, we drew from two sources. First, experienced consultants from the HR consultancy firm served as subject matter experts. They qualified as experts because they had provided solutions for selecting and developing junior managers in multiple client projects. Second, we built upon past research of Motowidlo et al. (1990). As noted above, according to Motowidlo et al.'s research, junior managers typically deal with situations that require a combination of problem solving, interpersonal skills, and communication. Hence, role-plays addressed these areas to various degrees (see Table 2).

All role-plays were integrated into an overarching background (i.e., the organization of a charity event). So, each of the 18 role-plays confronted the participant as project manager with a different character from inside or outside the organization that put forward a specific problem. This common background across all role-plays aimed to enhance realism and participants' immersion into the role-plays. It also ensured the amount of information given prior to role-plays was brief (at most one sentence).

A total of 30 role-players (80% females) participated in the multiple, speeded simulations. These were either consultants or graduate students from a large European university. Consistent with role-player training guidelines (Byham, 1977; Lievens, Schollaert, & Keen, 2015), we introduced them to their role and the overarching background (i.e., charity event). Role-players were also taught to use situational cues (Lievens, Schollaert et al., 2015; Schollaert & Lievens, 2011, 2012) to structure the role-plays and elicit behavior. An example of a cue was "I would really like to solve this problem, but I fail to see what I can do more. Can you help me?". Role-players learned the cues by heart. Role-players also practiced and received feedback about their role-playing behavior.

Assessors (i.e., the designated role-player as assessor) provided two to three overall ratings of participants' performance (e.g., *1 = should clearly be improved: starters' level* to *9 = obviously strong: role model behavior*) immediately after each role-play. In case of short role-plays, use of such overall ratings has been recommended (e.g., Brannick, 2008; Lievens, 2008). To ensure that these ratings were based on observable and relevant behaviors, we developed short checklists that listed five to seven behaviors indicative of (in)effective performance specific to each role-play (e.g., weighs alternative solutions that can solve the schedule conflict of the solidarity event and the soccer cup). Across role-plays, the average internal consistency reliability of these ratings was .76. We thus averaged the overall ratings to compute one performance score per role-play.

As mentioned above, this study's multiple, speeded simulations cover a large, heterogeneous domain and include a diversity of situational demands so that people's behavior and performances are expected to vary across the various situations (Lance, 2008; Lievens, 2002; Mischel & Shoda, 1995; Speer et al, 2014). For instance, participants might score well on dealing with a crisis but poorly on comforting a dejected subordinate. Theoretically, this interactional (person-situation) approach underlying this study's multiple,

speeded simulations is consistent with a formative indicator model (Bledow & Frese, 2009; Bollen & Lennox, 1991; Howell, Breivik, & Wilcox, 2007; Jarvis, MacKenzie, & Podsakoff, 2003; MacKenzie, Podsakoff, & Jarvis, 2005). In a formative indicator model, the single measures (in this case role plays) each capture a unique criterion part, are not interchangeable (do not correlate highly), and do not need to have the same antecedents/consequences (e.g., correlations with antecedents and criteria might differ depending on the simulations included, see H3a and H3b). This contrasts to a reflective indicator model (e.g., ability test items), wherein the single measures reflect the same common theme, are interchangeable and correlate highly, and have the same antecedents and consequences.

Given the overall score on multiple, speeded simulations does not represent a unidimensional latent trait that resides within individuals and affects their behaviors (see also Bledow & Frese, 2009), one might wonder whether it is justifiable to compute an overall simulation score. It is possible to compute a composite score in the case of formative indicators if one realizes that the distinct role-play simulations are seen as the defining characteristics of the domain to be sampled (Jarvis et al., 2003; MacKenzie et al., 2005). Hence, it is important to acknowledge that the meaning and validity of this composite score depends on and might change on the basis of the kind/variety of the simulations sampled (see H3a and H3b to test this), thereby underscoring the importance of stimulus domain sampling and thus including a large and diverse set of simulations. So, we averaged the ratings provided by each designated role-player in all eighteen role-plays into a composite measure of overall performance on multiple, speeded simulations.

SJT. To examine the role of response domain sampling we presented participants with 18 written situations, each followed by MC response options (Volckaert & Deruddere, 2013). Given it was not possible to present participants with the exact same situations as in the multiple, speeded simulations, the 18 written situations were “incident isomorphic” (Lievens

& Sackett, 2007) with the multiple, speeded simulation situations. This means that these situations came from the same domain as the multiple, speeded simulations; they thus built on similar incidents, the same characters, and the same overarching case (charity event). Participants also assumed the same role (event coordinator). Yet, contrary to the multiple, speeded simulations, participants were not asked to show a behavioral response but indicated their agreement with MC options (1 = *strongly disagree*, 5 = *strongly agree*). The Appendix presents an example SJT item and a corresponding simulation situation. An overall SJT score was computed by averaging the correct responses across each of the 18 situations. In line with past research (see Campion, Ployhart, & MacKenzie, 2014), the internal consistency of the scores on the 18-item SJT was low (.46). The overall SJT score correlated significantly (r = .32, p = .002) with the overall score on the multiple, speeded simulations.

Control measures. We included participants' gender and age as control variables because in past studies these variables were related to AC and SJT performance (e.g., Clapham & Fulford, 1997; Dean, Roth, & Bobko, 2008; Herde, Lievens, Jackson, Shalfrooshan, & Roth, 2020; Whetzel, McDaniel, & Nguyen, 2008).

Criterion Measure. Seven months later, each participant was rated by the MBA program instructors. During the MBA program, instructors had the opportunity to gain insights about participants across many team and project-based components. For example, action-based learning modules and workshops required participants to work together in self-managed groups, draft reports, and give presentations. Participants also worked on a real-life consultancy project for three months. Instructors often addressed participants' progress and problems, gave advice, and provided feedback.

To validate the multiple, speeded simulations we relied on the relative percentile approach (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996, Goffin, Jelley, Powell, & Johnston, 2009). In this approach, raters assign percentile scores to ratees per criterion

dimension. We used the relative percentile approach because Goffin and colleagues developed it to reduce rating inflation. This is because the reference group to be used for rating consists of the average MBA student (i.e., a percentile score of 50). Prior research confirmed that the relative percentile approach had higher criterion-related validity than conventional absolute rating formats (Goffin et al., 1996, 2009).

Instructors rated four items that represented each of four criterion dimensions: Besides task and contextual performance, teamwork and communication were also included because these are regarded as critical for junior managers (see Motowidlo et al., 1990). In the rating form, each criterion dimension was described with anchors taken from established scales (i.e., for task performance, see Williams & Anderson, 1991; for contextual performance, see Motowidlo & Van Scotter, 1994; for teamwork and communication, see Kyllonen, 2008). The internal consistency of instructor ratings was .87. On average, the four criterion dimensions correlated .65 (all *p*s < .001; see Online Supplement for the correlation table). Therefore, we averaged instructor ratings into an overall performance measure (see Viswesvaran, Schmidt, & Ones, 2005). A confirmatory factor analysis via Mplus 7.4 (Muthén & Muthén, 1998-2015) using the MLR estimator provided support for a one-factor model ($\chi^2(df) = 4.45(2)$, $p = .108$, CFI = .984, SRMR = .024), although the RMSEA was poor (.113, 90% CI = .000-.259).

Note that we also collected ratings from participants' peers (classmates). However, as we received peer ratings from more than one peer only for 28 participants, we decided not to report these results here. Results (available from the first author) show that running our analyses with peer ratings confirmed our conclusions, although validity coefficients were lower (probably due to peer ratings' lower reliability and higher leniency).

Transparency and Openness

We describe our sampling plan, all data exclusions (if any) and all measures in the study, and we adhered to the *Journal of Applied Psychology* methodological checklist.

Analysis data and code are not available because the data were gathered in collaboration with a consultancy firm. Some research materials (role-plays, scoring sheets, training materials, cognitive ability test, and Big-Five personality questionnaire) are not available due to their proprietary nature. The test motivation scale and criterion measures are included in the Online Supplement. Data were analyzed using R, version 3.6.3 (R Core Team, 2015) and the packages *psych* version 2.0.12 (Revelle, 2020) and *lme4* version 1.1-26 (Bates, Mächler, Bolker, & Walker, 2015). The study design was not preregistered because the data were collected for an applied selection project. The hypotheses and analysis were also not preregistered because at that time this was not common practice.

Results

Test of Hypotheses

H1 proposed that the overall score on multiple, speeded simulations predicts domain-relevant criterion performance seven months later. In line with our hypothesis, this overall score significantly predicted criterion performance ($r = .54$, $p < .001$). Therefore, H1 was supported.

There were significant correlations with criterion performance for 11 of the 18 role-plays. The average validity of the ratings from single role-plays for predicting criterion performance was .26. Ratings from single role-plays also showed highly variable correlations with criterion performance (range = .03 - .43, see Table 2). The average intercorrelation among the 18 role-play ratings was .19 (range = -.14 - .51). This variability in relationships with the criterion and the low intercorrelations among simulations show that these individual simulations are not interchangeable and are not equally related to the criterion; yet, taken all together the role-plays predict future performance (which is suggestive of a formative indicator model underlying the ratings made).

The next hypotheses dealt with the role of stimulus domain sampling and how domain coverage depends on the kind of simulations being included in the overall score on multiple, speeded simulations. H2 predicted the overall score to be significantly related to cognitive ability, extraversion, and agreeableness. In line with H2, the overall simulation score correlated positively with participants' cognitive ability (general mental ability test: $r = .27$, $p = .009$), extraversion ($r = .38$, $p < .001$), and agreeableness ($r = .24$, $p = .020$). Openness, emotional stability, and conscientiousness did not relate to the overall multiple, speeded simulation score (ps > .05, see Table 2). Again, the relationships between individual role-plays and these antecedents differed a lot across the role-plays (suggestive of a formative indicator model). That is, cognitive ability and personality measures showed variable correlations with ratings from different role-plays (see Table 2; general mental ability test: $M = .13$, range = -.04-.30; extraversion: $M = .19$, range = .01-.37; agreeableness: $M = .12$, range = -.02-.31).

We then hypothesized domain coverage to reduce when specific simulations were not included in the overall simulation score. To this end, one of the authors and a graduate psychologist sorted the simulations in groups. Two relatively distinct categories could be identified. In seven simulations, participants had to make decisions related to task-related issues (e.g., scheduling conflicts, budget constraints, scarcity of resource). Role-players here were portrayed as not knowing what to do so that participants had to "call the shots". Another category comprised of eight simulations wherein participants had to solve interpersonal and communication issues. In this category, role players were portrayed as angry, furious, insecure, ashamed, dejected, etc. So, participants had to cool down the role-player, be supportive, etc. Finally, there was a category which included three simulations wherein no agreement on the sorting could be reached.

H3a stated the overall score to be significantly less related to cognitive ability when the domain coverage is reduced when simulations that tap into task-related problem solving and decision-making issues are not included in it. To examine H3a we computed an overall score on multiple, speeded simulations with only the ratings on simulations that deal with interpersonal and communication issues. We then computed the correlation between this new composite and cognitive ability. The correlation between this new composite and cognitive ability significantly dropped from .27 to .17 (t= 2.13; p = .036). Importantly, the validity of this new interpersonally loaded composite for predicting the criterion also significantly dropped from .54 to .45 (t = 2.12; p = .036).

To test H3b we did the same but this time the new composite score included only ratings on the simulations that tap into task-related problem solving and decision-making issues. In line with H3b, when the domain coverage was accordingly reduced when simulations that deal with interpersonal and communication issues were not included in it, the overall score on multiple, speeded simulations was significantly less related to extraversion and agreeableness; from .38 to .25 (t = 2.70; p = .008) and from .24 to .06 (t = 3.79; p < .001), for extraversion and agreeableness, respectively. Validity of this new cognitively loaded composite for predicting the criterion also significantly dropped from .54 to .44 (t= 2.10; p = .038).

In sum, these results support H3a and H3b. The change in the nomological network of the overall score on multiple, speeded simulations upon leaving out sets of domain-relevant simulations reduces domain coverage, modifies the conceptual meaning of the overall simulation score, and reduces its validity, thereby underscoring the importance of including a diverse set of domain-relevant simulations.

Our final hypothesis addressed the role of response domain sampling. H4 stated that the overall score on multiple, speeded simulations significantly predicts future performance

over and above the overall score on MC (SJT) responses to similar situations. We ran a hierarchical regression to test this incremental validity hypothesis. Model 1 included gender and age as control variables. Model 2 included the cognitive ability and Big Five measures. In models 3 and 4, we entered the overall SJT score and the overall score on multiple, speeded simulations, respectively. The overall score on multiple, speeded simulations explained an additional 17% of variance in criterion performance ($p < .001$) above all other predictors (see Table 3). We also ran the hierarchical regression with the last two steps reversed. The SJT score did not add incremental variance over and above the overall simulation score (see Online Supplement). This supports the importance of response domain sampling, as specified by H4.

Additional Analyses

To complement the incremental validity analyses we ran utility analyses using the Cronbach-Gleser method (Brogden, 1949; Cronbach-Gleser, 1965; see Hogan & Zenke, 1986) and examined whether and when the multiple, speeded simulations result in return on investment as compared to traditional selection procedures. Although it is most relevant to compare the multiple, speeded role-plays to one traditional longer AC role-play, we also included utility comparisons with one structured interview and one unstructured interview. These utility analyses were based on the following input: The validity of the 18 3-minute multiple, speeded simulations was .54 (see Study 1 results). To conduct a stringent test, the validity of an AC role-play was set at .18 (i.e., validity associated with the upper 95% confidence interval in the meta-analysis of Hoffman, Kennedy, LoPilato, Monahan, & Lance, 2015) instead of at the mean validity of .12. The validities of the structured and unstructured interview were set at .28 and .21, respectively (McDaniel et al., 1994). The *SD*y value was based on 40% (see Hunter & Schmidt, 1983, p.476) of MBA students' average first year

annual income ($91,586)[7]. There were 90 participants and the selection ratio was set at .33 (i.e., mean across European MBA programs)[8]. The ordinate of the normal curve at the cut-off score was 0.3485. All cost estimates (e.g., training/ assessor fees, role-play development costs) were provided by the consultancy firm. Results (see appendix) showed that in case 90 participants are assessed, one lengthier role-play yields a utility of $541,521, whereas the 18 3-minute multiple, speeded simulations produced a utility of $1,222,289. The utility of the structured and unstructured interview was $948,994 and $711,320, respectively.

We also ran utility analyses in which we varied the number of participants (i.e., 36, 54, 72, 90, 108, 126; see Figure A1 in the Appendix). As a key conclusion, this program of 18 speeded simulations for assessing 90 participants started reaching a break-even point with one traditional role-play when the number of participants at least doubled the number of assessors (in this case, there were 18 assessors). For this study's multiple, speeded assessments to reach a break-even point with one structured interview, the number of participants had to be about four times larger than the number of assessors. In addition, the more participants in the multiple, speeded simulations, the more utility they generated because they then capitalized on the large number of assessors who can rate many participants in a short time span. Thus, multiple, speeded simulations are most appropriate in large-scale selection (public sector, police force, hospitals, banks, army, etc.). It also explains why OSCEs and MMIs have become widespread in educational settings.

Discussion

Study 1 tested the role of stimulus and response domain sampling for multiple, speeded simulations to be effective. Results showed that sampling multiple, heterogeneous and relevant situations (bundled in two broad groups) as well as of assessing a variety of behaviors that participants chose to exhibit (instead of their MC SJT choices) matter to cover

[7] https://www.eduopinions.com/blog/what-to-study/what-is-the-average-salary-for-an-mba-graduate/
[8] https://www.mbacrystalball.com/business-schools/selectivity-rate-top-mba-business-schools/

the criterion domain, capture both ability and personality differences, and achieve substantial validities. The high observed validity of .54 and substantial incremental validity over MC-based SJTs highlight that this was successful.

More precisely, our results provided evidence that the multiple, speeded simulations tap at least two distinct domains and that both contribute to predicting future performance. So, we illustrated the importance of stimulus domain sampling by showing that at least these two broad groups of situations from the junior management domain should be covered. Although prior research (e.g., Motowidlo et al., 1990) and subject matter expertise (experienced consultants) inspired the selection of the 18 specific situations to be covered in our multiple, speeded assessments, our results did not test the domain stimulus sampling logic at the fine-grained level of the individual simulations.

As a second and related conclusion, our results are suggestive of a formative (instead of reflective) indicator model underlying this study's multiple, speeded simulation ratings: The individual role-play simulations (a) did not correlate highly because they were heterogeneous and invoked various situational demands, (b) did not have the same antecedents or consequences, and (c) removing domain-relevant[9] role-plays changed the conceptual meaning of the overall score on multiple, speeded simulations and reduced validity[10].

Taken together, these results send a strong warning to practice. Whereas speeded assessment is often equated with the use of *single* simulations, our results show that high validities are not ensured. Instead, we recommend to carefully sample *multiple* and *diverse* simulations. This is especially important when covering a large domain such as junior

[9] As mentioned by an anonymous reviewer, the situations should be domain-relevant because situations that are not criterion-relevant might introduce error into the overall score and lower validity.

[10] Note that this support for a formative indicator model underlying multiple, speeded simulation ratings is only suggestive and awaits further replication. Given our limited sample size and related sampling error, similar results might be obtained from a reflective model with multiple dimensions, varying item loadings and item-specific error (Edwards, 2011).

management performance. So, Study 1 recommends relying on stimulus and response domain sampling to obtain qualitatively different pieces of behavioral information of participants across multiple, speeded simulations. The aim of Study 2 was to build on these results and identify other conditions that affect the quality of speeded simulations.

STUDY 2

Background and Research Questions

As noted above, speeded behavioral simulations are getting popular because they provide a shorter and fast-paced alternative to traditional lengthier behavioral simulations. The key message of Study 1 was twofold. On one hand, speeded behavioral simulations can obtain high validities and incremental validity. On the other hand, important design considerations must be followed in terms of covering the domain via a large set of heterogeneous simulations. Especially the latter might lift the total testing time again to the same level as with using one traditional lengthier AC role-play. Therefore, it is important to examine whether speeded behavioral simulations can also be designed in different formats and to determine the effects of such design variations on rating quality.

Given that speeded behavioral simulations originated in practice, its characteristics are not fixed in plaster. First, in Study 1, assessors rated 3-minute simulation performances because this time span was considered to give both role-players and assessors the opportunity to discuss the issues at hand. Yet, the thin slices research tradition suggests that this time span can be even shorter. Apart from investigating the effects of the quality of information (see Study 1), thin slices research also scrutinized the effects of the quantity of available information. In particular, research on slice length reveals that 1 minute of observation might be optimal for rating many personality and cognitive variables (Carney et al., 2007; see also Murphy et al., 2015). Moreover, meta-analytic research (Ambady & Rosenthal, 1992) showed that the accuracy of predicting various social and clinical outcomes does not significantly

differ between 30s and 5-min observations. Reducing the time needed for remote assessors to evaluate speeded simulations also benefits their practical use. It might mean that assessors - or in the future perhaps even algorithms - need to code and evaluate only one minute (although interactions might in principle take longer than one minute). Given the thin slices research evidence and these practical considerations, Study 2 examines the effect of slice length (first minute vs. full three minutes) as a first factor.

Second, in multiple, speeded simulations, investments in rating standardization are made following the long tradition in I/O psychology that documents the benefits of assessor training and rating aids (e.g., Lievens & Sackett, 2017; Melchers, Lienhardt, Von Aarburg, & Kleinmann, 2011; Roch, Woehr, Mishra, & Kieszczynska, 2012). Such rater standardization via trained raters and rating aids should ensure rating quality. Conversely, in lab studies in thin slices research, judges typically did not receive any training and rarely used rating aids. As a result, single rater reliabilities were relatively low in thin slices studies (Ambady et al., 2000; Connelly & Ones, 2010). Despite these reliability levels, significant validities[11] in terms of relations to cognitive ability, personality, or performance were observed (see above). To reflect these two traditions (i.e., the I/O psychology tradition and thin slices research), Study 2 includes rater standardization (assessor training and rating aids vs. no assessor training and no rating aids) as a second factor.

Finally, in Study 1, role-players also served as assessors. This was consistent with how companies and consultancy firms so far have adopted speeded simulations (to save time and costs, see e.g., Pinsight, 2018; 2019). However, it is best practice in ACs to separate assessor and role-player tasks because otherwise assessors' ratings might subtly factor in "satisfying" effects of the interaction itself (e.g., candidates reacting with compliant behavior when the

[11] This is partly due to thin slices studies often reporting results averaged across many judges. For instance, Ambady and Rosenthal (1992) calculated that the median number of judges in their meta-analysis was 37 (range: 2-446). This aggregation across judges reduces idiosyncrasies, and boost reliabilities and validities (Eisenkraft, 2013). In operational multiple, speed simulations, such a large number of assessors is not feasible and therefore investments in training and rating aids are made.

role-playing assessor shows more dominant behavior; Sadler, Ethier, & Woody, 2011; Tiedens & Fragale, 2003). That is why Study 2 uses only remote assessors and compares their ratings' validities to those of Study 1's role-playing assessors.

In short, Study 2 contrasts different research traditions (I/O psychology vs. thin slices tradition) to test two factors (length of observation and rater standardization) that might affect the reliability and validity of speeded simulation ratings. Given the discrepancy between these traditions, we put forward the following research questions instead of hypotheses:

Research Question 1: Is the interrater reliability of assessor ratings the highest in the condition when assessors rate 3-minute simulation performances under high rater standardization as compared to the other three conditions?

Research Question 2: Is the criterion-related validity of assessor ratings the highest in the condition when assessors rate 3-minute simulation performances under high rater standardization as compared to the other conditions?

Method

Sample

Study 2 relied on the same speeded simulation performances of the 96 participants of Study 1. Yet, in Study 2, a new pool of 60 assessors (70% females, mean age = 21.78, *SD* = 2.94) rated the videotapes of these performances. Assessors were students recruited from four European universities in the same country. All but two of them were studying for a degree in Psychology or Business Administration. Eighteen percent had some prior assessor experience. All assessors were paid and received a certificate.

Design

The 60 assessors were randomly assigned to four conditions. These four conditions resulted from a 2x2 design (level of rater standardization x length of observation).

Level of rater standardization. This factor had two conditions. In the high rater standardization condition, assessors received a 6-hour assessor training. This training was the same as the training given to assessors in Study 1. This training thus included core aspects of behavior-driven (Byham, 1977) and frame-of-reference training (Roch et al., 2012). It started with a lecture and exercises on observation, registration, classification, and evaluation. The training also familiarized assessors with the situational cues that the role-players used to elicit relevant behaviors. Another important part was that assessors were familiarized with the rating procedure and the rating aids (short checklists that listed behaviors indicative of (in)effective performance per role-play). Next, assessors practiced evaluating participant performances in the role-plays they were specialized in: They first watched videotaped performances and then independently provided evaluations. Assessors then met to reach consensus. This procedure was repeated for a total of three practice tapes.

Conversely, in the low rater standardization condition, assessors did not receive a systematic assessor training. To avoid Hawthorne effects, assessors followed a control training of similar length. Critically, however, this control training did not incorporate elements of behavior-driven or frame-of-reference training and did neither familiarize assessors with the situational cues nor with the behavioral checklists. In line with prior control trainings (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002), it included lectures and exercises on selection methods (e.g., ACs and SJTs). Assessors also independently practiced evaluating participants in videotaped role-plays. Yet, to prevent imposing a common frame-of- reference they did not discuss their ratings and did not receive feedback.

To examine possible differences in assessor satisfaction and motivation as consequence of these different trainings, assessors completed a survey (see Noordzij, Van Hooft, Van Mierlo, Van Dam, & Born, 2013) wherein they assessed their satisfaction with the (a) trainer, (b) training content, (c), materials, (d) organization, (e) own contribution, (f) and

training's usefulness (*1 = extremely dissatisfied* to *5 = extremely satisfied*). They also provided an overall evaluation "How would you rate the training program, on a scale from 1 (*very bad*) to 10 (*very good*)?". After the rating sessions, assessors also filled in an adapted version of Arvey et al.'s (1990) test motivation scale. Mann Whitney U tests revealed no significant differences on these satisfaction and motivation measures across trainings (all *ps* > .05). Generally, assessors were satisfied with the trainer (M = 4.81, SD = 0.39), content (M = 4.39, SD = 0.61), materials (M = 4.58, SD = 0.55), organization (M = 4.57, SD = 0.57), own contribution (M = 4.10, SD = 0.72), and usefulness (M = 4.42, SD = 0.66), and they evaluated the training favorably (M = 8.20, SD = 1.09). Their motivation was also high (M = 4.25, SD = 0.48, internal consistency reliability = .66).

Length of observation. This factor (i.e., slice length) had two conditions. In one condition, assessors saw only the first minute of the role-play, whereas in the other condition they observed the full 3 minutes.

Procedure

Assessors were randomly assigned to two to four role-plays so that we obtained ratings from two to three independent assessors per role-play and condition. When watching the recorded performances, assessors followed the same rating procedure as in Study 1. Note that rewinding or pausing role-play conversations was prohibited. To limit potential biases (e.g., order effects), we took various precautions: (a) distributing all records for the different role-plays per assessor across distinct blocks so that each contained records of only one role-play, (b) counterbalancing the records per role-play across assessors, and (c) randomly presenting participants per role-play. For 20% of the performances, cameras did not successfully record videos due to technical reasons so that only audio records were available. In these cases, assessors used audio records to evaluate performance. We re-ran our analyses

after dropping all ratings based upon audio records. Results and conclusions were similar. Thus, below we report our results using all ratings.

Measures

Assessors used the same rating scale as in Study 1. So, they provided two to three overall ratings of participants' performance and used behavioral checklists. Across role-plays, assessors, and conditions, the average internal consistency reliability of these ratings was .69 (range= .64-.78). In line with Study 1, overall ratings were averaged to compute one performance score per role-play. These role-play performance scores provided by each assessor per role-play were first averaged across assessors[12] and then averaged across all eighteen role-plays into an overall score on multiple, speeded simulations (see Study 1). Finally, data from the same criterion measures as in Study 1 were used.

Results

Research Question 1: Interrater Reliability of Multiple, Speeded Simulation Formats

Our first research question dealt with the interrater reliability of ratings across the various multiple, speeded simulation formats that were reflected in our four conditions. We started by computing single-rater and average interrater ICCs for ratings per role-play by condition (see Table 4). Across all four conditions, single-rater reliabilities (ICC[2,1]) were relatively low. Logically, average interrater reliabilities were higher than single-rater reliabilities but they reflected the same trend. That is, the differences in single-rater reliabilities across conditions were small and 95% confidence intervals overlapped most of the times. As shown at the bottom of Table 4, averaged single-rater reliabilities (across all role-plays) varied between .27 (low rater standardization, 3 min observation time) and .42 (low rater standardization, 1 min observation time). Average single-rater reliability was highest in the condition in which assessors evaluated 1-minute performances under low rater

[12] Given the number of assessors across role-plays slightly differed across conditions, we re-ran our validity analyses with the overall simulation score based on ratings from one randomly selected assessor. Results and conclusions were similar.

standardization (.42), followed by the condition in which assessors evaluated 3-minute performances under high rater standardization (.38; see RQ1).

An interrater reliability index is insightful but in multiple, speeded simulations it provides only an initial look into the amount of reliable variance. That is, interrater reliability mainly deals with assessor-related sources of variance such as assessor main effects (leniency/stringency), etc. However, in multiple, speeded simulations, there are other systematic and unsystematic sources of variance. To decompose ratings into these different sources of variance, we conducted per condition a generalizability analysis (Brennan, 2001; Vispoel, Morris, & Kilinc, 2018) that modeled participants, role-plays, and assessors as crossed-random factors to examine the relative contribution of all of these sources of variance (and their interactions) to the observed variance in the ratings made. We fitted linear random effects models with restricted maximum likelihood estimators (Putka, Le, McCloy, & Diaz, 2008) by using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) for R (R Core Team, 2015). We also present the generalizability coefficient as a summary statistic. As shown at the very bottom of Table 4, the results show the same trend as the ICCs. The largest generalizability coefficient was found for the two "extreme" conditions, namely high rater standardization and 3 min observation time (see RQ1) and low rater standardization and 1 min observation time, whereas the smallest one was observed for the high rater standardization, 1 min observation time condition. Detailed tables about the generalizability analyses can be found in the Online Supplement.

Research Question 2: Validity of Multiple, Speeded Simulation Formats

Our second research question dealt with the criterion-related validity of ratings across the various conditions. Table 5 summarizes the (incremental) validities by condition. Generally, validities mirrored the results of Study 1 that used role-playing assessor ratings (see last column in Table 5). The overall score on multiple, speeded simulations from all four

conditions significantly predicted criterion performance (*r*s between .43 [high rater standardization, 1 min observation time] and .56 [high rater standardization, 3 min observation time], see RQ2). In addition, in all four conditions, the overall simulation score was significantly related to cognitive ability, the SJT, and extraversion. Only ratings from the 3 min observation time conditions significantly correlated with agreeableness ($r = .25$, $p = .015$ and $r = .26$, $p = .010$, for high and low rater standardization conditions, respectively).

To examine possible differences in validities across conditions, we ran tests for the difference of dependent correlations (see Steiger, 1980). We warn that these tests should be interpreted with caution given the large number of tests being conducted. Ratings from the conditions of high rater standardization and 3 min observation time ($z = 2.53$, $p = .012$) and of low rater standardization and 3 min observation time ($z = 2.00$, $p = .045$) correlated significantly higher with criterion performance than ratings of the high rater standardization and 1 min observation time condition. Table 5 presents the detailed results. The trend is that especially the condition of high rater standardization and 1 min observation time showed slightly lower correlations.

Finally, we investigated possible differences in incremental validity of the overall score on multiple, speeded simulations across conditions. So, we ran multiple regressions per condition. The entry order was the same as in Study 1. Across conditions, results confirmed the substantial incremental validity of the overall simulation score. It explained between 13% (high rater standardization, 1 min observation time) and 20% (high rater standardization, 3 min observation time) in criterion performance above all other predictors.

Discussion

Study 2 examined two factors that might affect the rating quality of multiple, speeded simulations, leading to four different formats. Our choice of factors contrasted two research

traditions (I/O psychology vs. thin slices tradition) with diverging paradigms regarding observation time and rater standardization. Three main conclusions emerged.

First, in Study 2, single-rater reliabilities (between .27 and .42 across role-plays) were relatively low in all conditions, as compared to Connelly, Ones, Ramesh, and Goff's (2008; see also Connelly & Ones, 2004) average meta-analytical value of .71 for AC role-plays. Yet, this benchmark applies to single-rater reliability for an AC role-play that typically lasts longer, whereas observation time in the multiple, speeded simulations was at most 3 minutes. If we compare this study's reliabilities to meta-analytic thin slices research that reports values between .23 and .50 (Ambady et al., 2000; Connelly & Ones, 2010), Study 2's reliabilities are in the same range. As another comparison, Ingold et al. (2018) reported average single rater reliabilities of .15 for initial impression ratings (first two minutes of AC exercises).

As a second key conclusion, validities of the multiple, speeded simulation score from remote assessors across each of the four simulation formats corroborated the high validities obtained with role-playing assessors in Study 1, attesting to the robustness of results across different assessor types and formats. So, validities of ratings from role-playing assessors were thus also similar to those of remote assessors.

Third, the overall score on multiple, speeded simulations was about as valid in the two "extreme" conditions (i.e., assessors with a control training and no rating aids evaluating only the first minute and assessors with a thorough training and rating aids evaluating the full 3 minutes). Conversely, validity was lowest in the 1-minute high rater standardization condition. How can these results be explained? Past thin slices research (see e.g., Ambady & Rosenthal, 1993) scrutinized whether frugal and fast thin slices judgments are affected by distraction and deliberate reasoning. As a general conclusion, distractor tasks did not impede accuracy. However, accuracy did decrease when people were asked to verbally reason about their judgments and motivate them. So, frugal judgments seem to be more accurate when they

are made without deliberation. This pattern mirrors what we found in Study 2. The validity of thin slice judgments in the 1-min condition decreased when people received training and detailed rating aids. We speculate that this increased deliberation might have resulted in them overestimating specific cues (see Ambady, 2010). Conversely, when assessors made their judgments rapidly, this might not have been the case. If future studies confirm these speculations and findings, our Study 2 findings challenge the viability of our common training and rating aid practices in the face of multiple, *1-minute* assessments and call for other approaches to harness the accuracy of rapid judgments. At the same time, the fact that such a 1-minute observation time still leads to high validities might open opportunities to streamline multiple, speeded simulations in practice.

GENERAL DISCUSSION

Recently, multiple, speeded simulations have made inroads into the selection realm. So far, the conceptual underpinning and empirical evidence related to these short, fast-paced assessment approaches has been lacking. In addition, in practice the term "multiple, speeded assessments" seem to be a moving target without a clear definition. Therefore, we conceptualized this novel assessment approach as a hybrid of established simulation-based selection methods (ACs and SJTs), thereby introducing the notions of stimulus and response domain sampling. We also connected the thin slices of behavior paradigm to multiple, speeded simulations to theorize about design factors and present evidence related to the quality of the ratings.

Main Conclusions

The first aim of this paper was to determine whether multiple, speeded simulations "work" (Are they reliable and valid?). The answer can be summarized as a nuanced "yes, but under specific conditions". On one hand, at the individual speeded role-play level, reliability and validity are not ensured. Single-rater reliabilities were relatively low and echoed the

reliabilities in thin slices research (Ambady et al., 2000; Connelly & Ones, 2010). Thus, although role-players and assessors followed an intensive training, used rating aids, and relied upon behavior elicitation and evaluation via standardized cues, there remains too much error variance (idiosyncratic assessor effects) in their ratings. In addition, individual ratings from single role-plays showed variable relations to cognitive ability, extraversion and agreeableness. Finally, evidence of the predictive validity of ratings in single, short simulations was more encouraging but the same variability across role-plays was found.

On the other hand, the picture changed when we aggregated ratings across multiple simulations. In this case, we found substantial relations between the overall score on multiple, speeded simulations and cognitive ability, extraversion, as well as agreeableness. The predictive validity results mirrored this pattern. When ratings were aggregated across all role-plays, this overall simulation score significantly (.54) predicted performance seven months later. Further, this overall simulation score added large amounts of incremental variance for predicting performance, beyond measures of cognitive ability, personality traits, and an SJT.

Another aim of this paper consisted of investigating factors under which speeded simulations might work better (or worse). Carefully sampling multiple, heterogeneous situations that elicit a variety of domain-relevant behavior emerged as the key requirement to obtain adequate domain coverage, capture both ability and personality differences (extraversion and agreeableness), and achieve substantial validities. Consistent with thin slices research, validities of overall simulation scores remained high when assessors evaluated only the first minute and received only a control training.

Implications for Theory

We conceptualized multiple, speeded behavioral simulations as a hybrid between SJTs and ACs. Hence, it bridges the research literatures related to these two classic simulation-based selection methods. Multiple, speeded simulations capitalize on their respective

strengths: It shares with SJTs that it aims to comprehensively cover the stimulus domain by presenting a large and diverse number of situations. Yet, it goes beyond SJTs' limited response option choice by adding interactive behavioral responses in ongoing situations. So, it shares this interactive, behavioral focus with ACs to broaden the variety of possible behavioral responses that participants can show. The excellent observed validity obtained (.54) attests to a successful integration of these two seminal selection methods.

Our studies also speak to recent discussions as to whether initial impressions *can* be a reliable and valid source of variance in selection procedures (Barrick et al., 2010, 2012; Ingold et al., 2018; Swider et al., 2016). Drawing from the stimulus and response domain sampling logic and thin slices research, this study adds insights to this emerging knowledge base by identifying at least three conditions to improve the validity of speeded simulations.

Quality vs. quantity of information. Study 1 showed the importance of presenting multiple, qualitatively different situations (that could eventually be bundled in two broad groups) to candidates to make valid predictions. Only when the various behavioral simulations invoked such substantively different situational demands (and thus did not serve as interchangeable measures of one another) multiple, speeded simulations allowed obtaining adequate and representative domain coverage. Conversely, Study 2 revealed that the quantity of the behavioral information (i.e., the length of the performance to be observed and rated) plays a lesser role. Strikingly, ratings based upon 1-min snapshots were still valid and only rarely differed from ratings based upon the full 3 minutes, which extends conclusions from thin slices research (e.g., Ambady & Rosenthal, 1992; Carney et al., 2007; Murphy et al., 2015) to the assessment field. So, as a first condition, in multiple, speeded simulations, information quality (observing behavior in short, different situations and thus obtaining more and different information) seems more important than information quantity (observing

behavior for a longer time in the same situation and thus obtaining more information in general).

Training assessors vs. using multiple assessors. This second aspect reflects a longstanding difference between I/O psychology and thin slices research. Whereas assessor training programs and rating aids have been a hallmark of I/O psychology, thin slices researchers have rarely undertaken efforts to train judges and use rating aids. Conversely, to reduce assessor-specific idiosyncrasies ratings were often aggregated across a large number of judges (Eisenkraft, 2013). One provocative result of Study 2 was that validities of untrained assessors' ratings were often not significantly different from those of trained assessors. Conversely, there was more support for the principle of aggregation as a key factor in speeded simulations (Epstein, 1979). If ratings were aggregated across multiple, different assessors that evaluate candidates in different situations (see first condition), validities substantially increased. Thus, as a second condition, in multiple, speeded simulations, it seems more important to aggregate ratings across assessors/simulations than to provide extensive training to assessors that are specialized in one or two role-plays, even though we acknowledge that assessor training is also motivated by other reasons (e.g., legal requirements).

In situ vs. ex-situ assessors. AC best practices make a separation between role-player vs. assessor and do not merge these roles in a role-playing assessor. Recently, Rauthmann and Sherman (2020) equate such role-playing assessors with "in-situ" assessors because they are physically present and involved in the interaction, and they personally experience the interaction partner. Conversely, "ex situ" assessors are neither part of the interaction nor personally involved because they evaluate the interaction off-line (e.g., on video). Strikingly, our Study 2 results did not lend support to separate role-player and assessor roles because validities of in-situ assessors' ratings were not significantly lower than those of ex-situ

assessors. So, as a third condition, this study suggests that role-playing assessors can be used in multiple, speeded simulations.

Avenues For Future Research

As we conceptualized multiple, speeded simulations as a hybrid between SJTs and ACs, it is relevant for future studies to compare multiple, speeded simulations to its two "parents". For example, one could present the exact same set of situations in an open-ended written SJT vs. speeded assessment format. This permits verifying whether people can "translate" the procedural knowledge that they tap upon when completing SJTs into behavioral actions in speeded role-plays. Related to ACs, we need more utility analyses (see Study 1) to compare the return on investment of multiple, speeded simulations to various AC exercises and formats.

Second, the fact that participants go through a large number of short simulations offers opportunities for measuring adaptability and learning agility (Baard, Rench, & Kozlowski, 2014; Dalal, Bhave, & Fiset, 2014; DeRue, Ashford, & Myers, 2012; Jundt, Shoss, & Huang, 2015). This study found that participants differed in their performance across role-plays. Such intraindividual variability across situations is not only due to random error, but also suggests individuals systematically construe situations in different ways, leading to different behavior across them (Dalal et al., 2014; Fournier, Moskowitz, & Zuroff, 2008; Gibbons & Rupp, 2009; Lance, 2008; Mischel & Shoda, 1995). In future research, an intraindividual variability index could be computed across speeded simulations as a novel "construct" for prediction purposes (Lievens et al., 2018). To measure learning agility, future studies might construct the simulation sequence and the situational cues planted in them so that insights can be gleaned whether and how people quickly and flexibly learn (Lang & Bliese, 2009).

Third, efforts should be undertaken to more systematically sample a given domain. In this study, the domain reflected situations that junior managers might face (Motowidlo et al.,

1990). Future studies might also rely on situation taxonomies (Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014; Ziegler, Horstmann, & Ziegler, 2019; see also Carson, 1969; Kiesler, 1983; Yukl, 2010). We see less value in trying to identify the specific simulations that are the most predictive and then only deploy these simulations (instead of the full range of simulations) because such efforts run counter the stimulus and response domain sampling logic (see also the evidence suggestive of an underlying formative indicator model).

Fourth, we need to zoom into the applicant experience behind speeded assessments. One of the reasons for their growing popularity is that they mirror the fragmented and hectic nature of modern work and life. When speeded assessments are integrated into a common theme, they are also expected to be more "immersive" and "engaging" (Herde & Lievens, 2020). Yet, these assumptions await empirical confirmation.

Limitations

First, we included only speeded behavioral simulations (role-plays). In principle, other simulations such as short presentations or fact-findings can be implemented under the umbrella term of multiple, speeded assessments (see Knorr & Hissbach, 2014). Our multiple, speeded simulations were also set up in a "brick-and-mortar" fashion. So, we cannot extend our conclusions to online formats like webcam role-plays/ SJTs with constructed responses (e.g., Cucina et al., , 2015; Lievens et al., 2019). Yet, all Study 2 ratings were provided by remote assessors and validities of their ratings confirmed those of role-playing assessors.

Second, this study focused on the junior management domain. Other domains should also be sampled in multiple, speeded simulations, such as the leadership, interpersonal, integrity, or decision-making. Multiple, speeded simulations might also focus on specific job families such as customer service, sales, or call centers. Extrapolating on our results, we expect multiple, speeded simulations to produce good validity results at the aggregate level as

long as the multiple simulations sample a large and diverse set of domain-relevant situations and relevant behavioral responses.

Third, this study obtained criterion data by asking MBA instructors to evaluate the participants seven months later. Although these criterion data do not reflect academic performance (GPA), we can also not equate them to job performance. Therefore, future studies are needed to examine the validity of multiple, speeded assessments for predicting supervisory job performance.

Implications for Practice

In today's selection practice, multiple, speeded assessments exist in various formats. Generally, our results send a strong warning to organizations that equate speeded assessment with one short simulation or a limited number of such speeded behavioral simulations. A first key conclusion is thus that selection practices should not be degraded to such a *single*, short behavioral simulation. Conversely, a *large* set of short and *diverse* behavioral simulations should be used. Thus, for multiple, speeded simulations to capitalize on stimulus domain sampling and response domain sampling, participants should be confronted with multiple, flash simulations that elicit behavior relevant for a large set of qualitatively different parts of the domain. It is difficult to make general recommendations about the exact number of behavioral simulations to be developed because this depends on the breadth of the domain. As a second key conclusion, instead of relying on ratings of single assessors, selection decisions should be based on ratings that are aggregated across assessors and behavioral simulations. Only then, it is ensured that the multiple, speeded simulations provide valid insights about future performance and do this over and above SJTs, cognitive ability, and personality.

Table 6 summarizes these and other design considerations related to multiple, speeded simulations that flow from our two studies. As shown, Table 6 lists design considerations and rating process considerations. If organizations decide to "jump on the multiple, speeded

assessment train", we suggest that they follow up on these recommendations. Although our recommendations are evidence-based, we acknowledge that this is only the first empirical study that investigates the reliability and validity of short interpersonal simulations and potential factors that might affect them. So, further replication and other studies are needed to refine and extend these recommendations.

Finally, to develop the large set of behavioral simulations, we suggest that practitioners follow a systematic procedure to map the performance (criterion) domain in the simulations. Logically, this stepwise procedure echoes the development of AC exercises and SJTs (as described by Thornton, Mueller-Hanson, & Rupp, 2016). It starts with obtaining a thorough insight into the performance domain. Such insight can be derived from work analyses and/or theoretical models (e.g., leadership models). For example, in this study, we drew from work of Motowidlo et al. (1990) to break down the junior management domain in three subdomains (i.e., solving and deciding on task-oriented problems, dealing with interpersonal issues, and convincing communication). In the next step, Subject Matter Experts (SMEs) are consulted individually or in focus groups to generate critical job situations per relevant subdomain. To provide inspiration to SMEs we experienced it is useful to provide them with a matrix in which each subdomain is crossed with relevant interaction partners (e.g., supervisor, subordinate, colleague, client, supplier, media, general public). In the following step, test developers use these critical job situations to build the short simulations. This includes designing realistic stimulus material, participant instructions, role-player instructions, general background, etc. (see Thornton et al., 2016). As a last step, other SMEs can rate to what extent the short simulations capture the (sub)domains to verify domain coverage (Colquitt, Sabey, Rodell, & Hill, 2019).

Conclusion

This study investigated the theoretical underpinnings and empirical evidence behind speeded assessments. These assessments have the potential to ensure adequate stimulus domain sampling and response domain sampling by requiring participants to show interactive behavior in a large variety of different, brief situations. Assessor ratings from speeded simulations were reliable and valid indicators of performance, but *only if* these ratings were aggregated across a large set of heterogeneous situations. In other words, shortcuts in assessment science do not seem to exist: Speeded assessment approaches truly need to be conceptualized as *multiple*, speeded assessments.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32*, 201–271. http://dx.doi.org/10.1016/S0065-2601(00)80006-4

Ambady, N., Hogan, D. B., Spencer, L. M., & Rosenthal, R. (1993). *Ratings of thin slices of behavior predict organizational performance.* 5th Annual Convention of the American Psychological Society, Chicago, IL.

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

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

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

Arthur Jr, W., & 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. https://doi.org/10.1037/0021-9010.93.2.435

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Baard, S. K., Rench, T. A., & Kozlowski, S. W. J. (2014). Performance adaptation: A theoretical integration and review. *Journal of Management, 40*, 48–99. https://doi.org/10.1177/0149206313488210

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 98–124). Cambridge, UK: Cambridge University Press.

Barrick, M. R., Dustin, S. L., Giluk, T. L., Stewart, G. L., Shaffer, J. A., & Swider, B. W. (2012). Candidate characteristics driving initial impressions during rapport building: Implications for employment interview validity. *Journal of Occupational and Organizational Psychology, 85*, 330–352. https://doi.org/10.1111/j.2044-8325.2011.02036.x

Barrick, M. R., Swider, B. W., & Stewart, G. L. (2010). Initial evaluations in the interview: Relationships with subsequent interviewer evaluations and employment offers. *Journal of Applied Psychology, 95*, 1163–1172. https://doi.org/10.1037/a0019918

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*. https://doi.org/10.18637/jss.v067.i01

Bick, J. (2007, January 2). Businesses try a form of speed dating. *The New York Times*. Retrieved from https://www.nytimes.com/2007/01/02/technology/02iht-network.4077677.html

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*, 229-258. https://doi.org/10.1111/j.1744-6570.2009.01137.x

Bogaert, J., Trbovic, N., & Van Keer, E. (2005). *Ability Test Suite—Level III – Manual*. Ghent, Belgium: Hudson.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314. https://doi.org/10.1037/0033-2909.110.2.305

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

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

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 131–133. https://doi.org/10.1111/j.1754-9434.2007.00025.x

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education, 45*, 1181–1189. https://doi.org/10.1111/j.1365-2923.2011.04075.x

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brogden, H.E. (1949) When testing pays off. *Personnel Psychology, 2*, 171 185.

Byham, W. C. (1977). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 89–125). Pergamon Press.

Byham, W. (2016, October). *Assessment centers for large populations.* Presented at the International Congress on Assessment Center Methods, Bali, Indonesia.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. (2014). The state of research on situational judgement tests: A content analysis and directions for future research. *Human Performance, 27*, 283–310. https ://doi.org/10.1080/08959 285.2014.929693

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

Carson, R. C. (1969). *Interaction concepts of personality*. Chicago, IL: Aldine.

Casey, P. M., Goepfert, A. R., Espey, E. L., Hammoud, M. M., Kaczmarczyk, J. M., Katz, N. T., Neutens, J. J., Nuthalapaty, F. S., & Peskin, E. (2009). To the point: Reviews in medical education—the Objective Structured Clinical Examination. *American Journal of Obstetrics and Gynecology, 200*, 25–34. https://doi.org/10.1016/j.ajog.2008.09.878

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Clapham, M. M., & Fulford, M. D. (1997). Age bias in assessment center ratings. *Journal of Managerial Issues, 9*, 373-387. https://www.jstor.org/stable/40604153

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*, 1243–1265. https://doi.org/10.1037/apl0000406

Connelly, B. S., & Ones, D. S. (2008, April). Interrater unreliability in assessment center ratings: A meta-analysis. In B. J. Hoffman (Chair), *Reexamining assessment centers: Alternate approaches*. Symposium conducted at the 23rd Annual Conference of the Society of Industrial and Organizational Psychology (SIOP), San Francisco, CA, USA.

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

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. I*ndustrial and Organizational Psychology, 1,* 121-124. https://doi.org/10.1111/j.1754-9434.2007.00022.x

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

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper Collins.

Cronbach, L.J., & Gleser, G.C. (1965) *Psychological Tests and Personnel Decisions.* Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cucina, J. M., Su, C., Busciglio, H. H., Harris Thomas, P., & Thompson Peyton, S. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment*, *23*, 197–209. https://doi.org/10.1111/ijsa.12108

Dalal, R. S., Bhave, D. P., & Fiset, J. (2014). Within-person variability in job performance: A theoretical review and research agenda. *Journal of Management*, *40*, 1396–1436. https://doi.org/10.1177/0149206314532691

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685-691. https://doi.org/10.1037/0021-9010.93.3.685

DeRue, D. S., Ashford, S. J., & Myers, C. G. (2012). Learning agility: In search of conceptual clarity and theoretical grounding. *Industrial and Organizational Psychology*, *5*, 258–279. https://doi.org/10.1111/j.1754-9434.2012.01444.x

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*, 370-388. https://doi.org/10.1177/1094428110378369

Eisenkraft, N. (2013). Accurate by way of aggregation. *Journal of Experimental Social Psychology*, *49*, 277–279. https://doi.org/10.1016/j.jesp.2012.11.005

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*, 1097–1126. http://dx.doi.org/10.1037/0022-3514.37.7.1097

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: The multiple mini-interview. *Medical Education*, *38*, 314–326. https://doi.org/10.1046/j.1365-2923.2004.01776.x

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2008). Integrating dispositions, signatures, and the interpersonal domain. *Journal of Personality and Social Psychology*, *94*, 531–545. https://doi.org/10.1037/0022-3514.94.3.531

Fridman, A. (2015, July 5). 4 Reasons speed is everything in business. *Inc.* Retrieved from https://www.inc.com/adam-fridman/4-reasons-speed-is-everything-in-business.html

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management, 35*, 1154–1180. https://doi.org/10.1177/0149206308328504

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology*, *11*, 23–33. https://doi.org/10.1007/BF02278252

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*, 251–268. https://doi.org/10.1002/hrm.20278

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 2–34). San Francisco, CA: Jossey-Bass.

Gonzalez-Mulé, E., Mount, M. K., & Oh, I. S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*, 1222-1243. https://doi.org/10.1037/a0037547

Guion, R. M. (1978). A note on Taylor's "EEOC guidelines on content validity." *Journal of Assessment Center Technology, 1*, 15–17.

Harden, R. McG., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal, 1*, 447–451. https://doi.org/10.1136/bmj.1.5955.447

Herde, C. N., & Lievens, F. (2020). Multiple Speed Assessments: Theory, practice, & research evidence. *European Journal of Psychological Assessment, 36*, 237-249.. https://doi.org/10.1027/1015-5759/a000512

Herde, C. N., Lievens, F., Jackson, D. J., Shalfrooshan, A., & Roth, P. L. (2020). Subgroup differences in situational judgment test scores: Evidence from large applicant samples. *International Journal of Selection and Assessment, 28*, 45-54. https://doi.org/10.1111/ijsa.12269

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

Hogan, J., & Zenke, L. L. (1986). Dollar-value utility of alternative procedures for selecting school principals. *Educational and Psychological Measurement, 46*, 935–945. https://doi.org/10.1177/001316448604600413

Howell, R.D., Breivik, E., & Wilcox, J.B. 2007. Reconsidering formative measurement. *Psychological Methods,12,* 205-218. https://doi.org/10.1037/1082-989X.12.2.205

Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. *American Psychologist, 38*, 473–478. https://doi.org/10.1037/0003-066X.38.4.473

Ingold, P. V., Dönni, M., & Lievens, F. (2018). A dual-process theory perspective to better understand judgments in assessment centers: The role of initial impressions for dimension ratings and validity. *Journal of Applied Psychology, 103*, 1367–1378. https://doi.org/10.1037/apl0000333

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y.-J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*, 976–994. https://doi.org/10.1037/apl0000102

Jarvis, C.B., Mackenzie, S.B., & Podsakoff, P.M. 2003. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research 30*, 199-218. https://doi.org/10.1086/376806

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jundt, D. K., Shoss, M. K., & Huang, J. L. (2015). Individual adaptive performance in organizations: A review. *Journal of Organizational Behavior, 36*, S53–S71. https://doi.org/10.1002/job.1955

Kiesler, D. J. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review, 90*, 185–214. http://dx.doi.org/10.1037/0033-295X.90.3.185

Klein, C., DeRouin, R. E., & Salas, E. (2006). Uncovering workplace interpersonal skills: A review, framework, and research agenda. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (Vol. 21, pp. 80–126). New York, NY: Wiley & Sons. https://doi.org/10.1002/9780470696378.ch3

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: Same concept, different approaches. *Medical Education, 48,* 1157–1175. https://doi.org/10.1111/medu.12535

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148–161. https://doi.org/10.1037/0022-3514.86.1.148

Kyllonen, P. C. (2008). *The research behind the ETS® Personal Potential Index (PPI).* Princeton, NJ: ETS.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1,* 84–97. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lang, J. W. B., & Bliese, P. D. (2009). General mental ability and two types of adaptation to unforeseen change: Applying discontinuous growth models to the task-change paradigm. *Journal of Applied Psychology, 94,* 411–428. https://doi.org/10.1037/a0013803

Leising, D., & Bleidorn, W. (2011). Which are the basic meaning dimensions of observable interpersonal behavior? *Personality and Individual Differences, 51,* 986–990. https://doi.org/10.1016/j.paid.2011.08.003

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

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

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology, 1,* 112–115. https://doi.org/10.1111/j.1754-9434.2007.00020.x

Lievens, F., Lang, J. W. B., De Fruyt, F., Corstjens, J., Van de Vijver, M., & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. *Journal of Applied Psychology, 103,* 753–771. https://doi.org/10.1037/apl0000280

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96,* 927-940. https://doi.org/10.1037/a0023496

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. https://doi.org/10.1037/0021-9010.92.4.1043

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102,* 43-66. https://doi.org/10.1037/apl0000160

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology, 104,* 715-726. https://doi.org/10.1037/apl0000367

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. https://doi.org/10.1037/apl0000004

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in Personnel and Human Resources Managemen*t (Vol. 28, pp. 99–152). Bingley: Emerald Group Publishing.

Liff, J. P. (2017, April). *Next generation assessment: The state of innovations in selection science*. Panel discussion conducted at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL, USA.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90*, 710-730. https://doi.org/10.1037/0021-9010.90.4.710

Markey, P. M., & Markey, C. N. (2006). A spherical conceptualization of personality traits. *European Journal of Personality, 20*, 169–193. https://doi.org/10.1002/per.582

McCrae, R. R., & Costa, P., Jr. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56*, 586–595. http://dx.doi.org/10.1037/0022-3514.56.4.586

McCrae, R. R., & Costa, P., Jr. (1995). Trait explanations in personality psychology. *European Journal of Personality, 9*, 231–252. https://doi.org/10.1002/per.2410090402

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

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. https://doi.org/10.1037/0033-295X.102.2.246

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647. http://dx.doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*, 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79*, 475–480. https://doi.org/10.1037/0021-9010.79.4.475

Mullenweg, M. (2014). The CEO of automattic on holding" auditions" to build a strong team. *Harvard Business Review, 92*, 39–42.

Murphy, N. A. (2005). Using thin slices for behavioral coding. *Journal of Nonverbal Behavior, 29*, 235–246. https://doi.org/10.1007/s10919-005-7722-x

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin, 33*, 325–339. https://doi.org/10.1177/0146167206294871

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality, 71*, 465–493. https://doi.org/10.1111/1467-6494.7103008

Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., & Nguyen, L. (2015). Reliability and validity of nonverbal

thin slices in social interactions. *Personality and Social Psychology Bulletin*, *41*, 199–213. https://doi.org/10.1177/0146167214559902

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th Edition). Los Angeles, CA: Muthén & Muthén.

Needleman, S. E. (2007). Speed Interviewing Grows As Skills Shortage Looms. *The Wall Street Journal*, *B15*, 2. Retrieved from https://www.wsj.com/articles/SB119430485859183090

Noordzij, G., van Hooft, E. A., van Mierlo, H., van Dam, A., & Born, M. P. (2013). The effects of a learning-goal orientation training on self-regulation: A field experiment among unemployed job seekers. *Personnel Psychology*, *66*, 723-755. https://doi.org/10.1111/peps.12011

Ones, D. S., Dilchert, S., & Viswesvaran, C. (2012). Cognitive abilities. In N. Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0010

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*, 642. https://doi.org/10.1037/pspp0000111

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher*, *35*, 503–514. https://doi.org/10.3109/0142159X.2013.774330

Pinchback, J. (2017, October 4). *Introducing talent auditions*. LinkedIn Talent Connect presentation. Retrieved from https://www.youtube.com/watch?v=1UwVTOqlPwI

Pinsight. (2018, October 19). Virtual assessment centers. Retrieved from https://www.pinsight.com/our-platform/

Pinsight. (2019, July 2). *Shorter & Faster: Recent Trend in Assessment Centers*. Retrieved from https://www.youtube.com/watch?v=oRzF4FK47Ms

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*, 114–133. https://doi.org/10.1037/a0030887

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*, 959–981. http://dx.doi.org/10.1037/0021-9010.93.5.959

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rauthmann, J., Gallardo-Pujol, D., Guillaume, E., Todd, E., Nave, C., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Personality Processes and Individual Differences*, *107*, 677–718. http://dx.doi.org/10.1037/a0037250

Rauthmann, J. F., & Sherman, R. A. (2020). The situation of situation research: Knowns and unknowns. *Current Directions in Psychological Science*, *29*, 473-480. https://doi.org/10.1177/0963721420925546

Raven, J. C. (1958). Advanced progressive matrices (2nd ed.). London: Lewis.

Revelle W (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois.

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin*, *27*, 187–200. https://doi.org/10.1177/0146167201272005

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of*

Occupational and Organizational Psychology, *85*, 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., ... & Thornton, G. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management, 41*, 1244-1273. https://doi.org/10.1177/0149206314567780

Rushforth, H. E. (2007). Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Education Today, 27*, 481–490. https://doi.org/10.1016/j.nedt.2006.08.009

Ryan, A. M., & Greguras, G. J. (1998). Life is not multiple choice: Reactions to the alternatives. In M. Hakel (Ed.), *Beyond multiple-choice: Alternatives to traditional testing* (pp. 183–202). Mahwah, NJ: Erlbaum.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology*, *40*, 13–25. https://doi.org/10.1111/j.1744-6570.1987.tb02374.x

Sadler, P., Ethier, N., & Woody, E. (2011). Interpersonal complementarity. In L. M. Horowitz, & S. N. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 123–142). New York: Wiley.

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. https://doi.org/10.1037/0021-9010.87.4.735

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, *15*, 187–210. https://doi.org/10.1080/08959285.2002.9668091Schollaert, E., & Lievens, F. (2011). The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment*, *19*, 190–197. https://doi.org/10.1111/j.1468-2389.2011.00546.x

Schollaert, E., & Lievens, F. (2012). Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance*, *25*, 255–271. https://doi.org/10.1080/08959285.2012.683907

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*, 674–687. http://dx.doi.org/10.1037/0022-3514.67.4.674

Smith, R. E., Shoda, Y., Cumming, S. P., & Smoll, F. L. (2009). Behavioral signatures at the ballpark: Intraindividual consistency of adults' situation–behavior patterns and their interpersonal consequences. *Journal of Research in Personality*, *43*, 187–195. https://doi.org/10.1016/j.jrp.2008.12.006

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, *42*, 137–144. https://doi.org/10.1037/0003-066X.42.2.137

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology, 99*, 282-295. https://doi.org/10.1037/a0035213

Stallings, W. M., & Spencer, R. E. (1967). *Ratings of instructors in accountancy 101 from videotape clips.* Research Report No. 265. Office of Instructional Resources: Measurement and Research Division, University of Illinois.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251. https://doi.org/10.1037/0033-2909.87.2.245

Swider, B. W., Barrick, M. R., & Harris, T. B. (2016). Initial impressions: What they are, what they are not, and how they influence structured interview outcomes. *Journal of Applied Psychology, 101,* 625–638. https://doi.org/10.1037/apl0000077

Tiedens, L. Z., & Fragale, A. R. (2003). Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology, 84,* 558–568. https://doi.org/10.1037/0022-3514.84.3.558

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308,* 1623–1626. https://doi.org/10.1126/science.1110589

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods, 23,* 1–26. https://doi.org/10.1037/met0000107

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108–131. https://doi.org/10.1037/0021-9010.90.1.108

Volckaert, E., & Dereuddre, S. (2013). *Flexible Competency Assessment FCA – Elktronische Simulatieoefening – Handleiding*. Ghent, Belgium: Hudson.

Vrijdags, A., Bogaert, J., Trbovic, N., & Van Keer, E. (2014). *Business Attitudes Questionnaire (Psychometric technical manual)*. Ghent, Belgium: Hudson..

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. https ://doi.org/10.1080/0895928080 2137820

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17,* 601–617. https://doi.org/10.1177/014920639101700305

Wilmot, M. P., Wanberg, C. R., Kammeyer-Mueller, J. D., & Ones, D. S. (2019). Extraversion advantages at work: A quantitative review and synthesis of the meta-analytic evidence. *Journal of Applied Psychology, 104,* 1447–1470. https://doi.org/10.1037/apl0000415

Yukl, G. (2010). *Leadership in Organizations*. Upper Sadle River, NJ: Prentice Hall.

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin, 28,* 238–249. https://doi.org/10.1177/0146167202282009

Ziegler, M., Horstmann, K. T., & Ziegler, J. (2019). Personality in situations: Going beyond the OCEAN and introducing the Situation Five. *Psychological Assessment, 31,* 567–580. http://dx.doi.org/10.1037/pas0000654

Table 1

Means, Standard Deviations, and Intercorrelations Between Study 1 Variables

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
Controls													
1 Gender	96	-	-	-									
2 Age	96	23.63	1.85	-.15	-								
Predictors													
3 GMA	95	16.86	6.70	.13	-.37**	-							
4 SJT	95	3.54	0.24	.22*	-.26*	.22*	-						
5 Extraversion	95	90.82	12.54	.03	-.17	.10	.19	-					
6 Agreeableness	95	92.82	13.54	.04	-.27**	.14	.14	.53**	-				
7 Openness	95	83.76	14.18	-.06	.02	.01	.10	.51**	.38**	-			
8 Emotional stability	95	86.55	12.39	-.22*	.01	.07	.02	.51**	.30**	.38**	-		
9 Conscientiousness	95	92.60	10.08	-.07	.06	-.01	.08	.07	.06	.27**	.13	-	
10 Overall Simulation Score	96	5.51	0.73	.00	-.38**	.27**	.32**	.38**	.24*	.11	.05	-.06	
11 Criterion performance	95	57.47	21.00	-.13	-.31**	.12	.24*	.07	.19	.00	-.10	-.03	.54**

Note. $N = 96$. Gender: male = 1, female = 2. Overall simulation score was a composite of ratings provided by the designated role-player from all 18 role-plays. * $p < .05$, ** $p < .01$.

Table 2

Relations Between Role-play Ratings, Overall Score on Multiple, Speeded Simulations, Cognitive Ability, Personality, and Criterion Performance in Study 1

Role-play		*n*	*M*	*SD*	GMA	E	A	Criterion performance
1	Role-player wants to find extra volunteers but does not want to run into a conflict with other event services (T).	95	6.79	1.40	.27**	.20	.01	.20
2	Role-player is dissatisfied with many aspects of last year's event and threatens to take action to forbid the event (I).	94	6.42	1.65	.03	.15	.27**	.24*
3	Role-player is hostile and wants to boycott the event because several activities threaten the attendees' safety (I).	96	5.81	1.42	.12	.30**	.16	.13
4	Role-player feels inexperienced and insecure because her efforts to increase sales do not pay off (I).	92	5.16	1.55	.10	.27**	.08	.26*
5	Role-player mentions a popular sport event is scheduled on the same day and that this should be solved (T).	96	5.41	1.84	.21*	.17	-.01	.43**
6	Role-player promised a band to play at the event but the committee had already decided hosting a different band (T).	95	4.89	1.76	.05	.05	-.01	.32**
7	Role-player wants to brainstorm about solving the problem of shortage of volunteers (T).	92	5.86	1.50	.15	.11	-.02	.03
8	Role-player criticizes participant, asking to make quick decisions regarding specific entertainment issues (R)	94	6.21	1.46	.12	.16	.11	.37**
9	Role-player (finance coordinator) asks to make a choice among various options, while staying within the budget (T).	94	5.36	1.33	-.01	.01	.03	.14
10	Role-player is inexperienced, feels close to burnout, and considers resigning from her job (I).	67	5.41	1.55	.07	.27*	.31*	.21
11	Role-player (a police inspector) is angry because the current event proposal does not meet safety regulations (I).	94	5.68	1.38	.14	.28**	.26*	.30**
12	Role-player is angry about another employee who does not meet his task expectations (I).	94	5.45	1.35	-.04	.12	.10	.30**
13	Role-player feels disengaged and is unmotivated to switch to another catering option (I).	88	4.12	1.47	.06	.21	.09	.21
14	Role-player suggests to completely change the event activities, although many preps have already been done (R).	92	4.91	1.91	.17	.37**	.12	.30**
15	Role-player lost the registration list and has problems to acknowledge it because of potential face loss (I).	91	5.41	1.35	.15	.21*	.26*	.34**
16	Role-player (beverage supplier) mentions that a final order for beverages was never placed and his schedule is full (T).	90	5.47	1.24	.26*	.17	.15	.13
17	Role-player (from ICT) is furious and questions the need to take up extra IT tasks (R).	95	5.02	1.40	.15	.10	.19	.35**
18	Role-player mentions a double booking was made regarding the order of plates and cutlery (T).	93	5.72	1.52	.30**	.23*	.10	.28**
Overall		**96**	**5.51**	**0.73**	**.27****	**.38****	**.24***	**.54****

Note. Overall = overall score on the 18 speeded simulations (i.e., composite of ratings provided by the designated role-playing assessor across 18 role-plays). E = Extraversion; A = Agreeableness. I = role plays related to interpersonal/communication issues; T = role-plays related to decision making about task-related issues and R = rest category* $p < .05$, ** $p < .01$.

Table 3

Results of Multiple Regression Analysis to Predict Criterion Performance in Study 1

Predictors	*b*	β	R^2 *(adj.)*	Δ R^2	*F*	*df*	*p*	Sig. *F* change
Model 1			.13 (.11)	.13	6.67	2,91	.002	.002
Gender	-7.35	-.18						
Age	-3.84**	-.34						
Model 2			.18 (.10)	.05	2.26	8,85	.031	.558
Gender	-9.50*	-.23						
Age	-3.17*	-.28						
GMA	0.10	.03						
Extraversion	0.10	.06						
Agreeableness	0.24	.16						
Openness	-0.03	-.02						
Conscientiousness	-0.03	-.01						
Emotional stability	-0.37	-0.22						
Model 3			.21 (.13)	.03	2.54	9,84	.012	.045
Gender	-11.09*	-.26						
Age	-2.76*	-.24						
GMA	0.01	.00						
Extraversion	0.04	.02						
Agreeableness	0.25	.16						
Openness	-0.04	-.02						
Conscientiousness	-0.07	-.03						
Emotional stability	-0.35	-.21						
SJT	18.44*	.21						
Model 4			.38 (.31)	.17	5.12	10,83	< .001	< .001
Gender	-7.12	-.17						
Age	-1.39	-.12						
GMA test	-0.19	-.06						
Extraversion	-0.29	-.17						
Agreeableness	0.26	.17						
Openness	0.00	.00						
Conscientiousness	0.02	.01						
Emotional stability	-0.20	-.12						
SJT	9.06	.11						
Overall simulation score	14.54**	.50						

Note. $N = 94$. Gender: male = 1, female = 2; Overall simulation score was a composite of ratings provided by the designated role-player from all 18 role-plays. * $p < .05$, ** $p < .01$

Table 4

Summary of Reliability and Generalizability Analyses of Study 2

	3 minute observation time				1 minute observation time			
	High standardization		Low standardization		High standardization		Low standardization	
Role-play	ICC 2,1	ICC 2,k	ICC 2,1	ICC 2,k	ICC 2,1	ICC 2,k	ICC 2,1	ICC 2,k
1	.24	.49	.20	.33	.18	.39	.45	.62
	[.12; .37]	[.29; .64]	[-.06; .43]	[-.13; .61]	[.00; .36]	[.01; .63]	[.16; .63]	[.28; .78]
2	.47	.64	.15	.34	.10	.19	.18	.31
	[.33; .59]	[.50; .75]	[.02; .28]	[.06; .54]	[-.04; .25]	[-.09; .40]	[-.06; .41]	[-.13; .58]
3	.40	.67	.43	.60	.11	.20	.58	.73
	[.29; .51]	[.55; .76]	[.27; .56]	[.42; .72]	[-.05; .29]	[-.11; .45]	[.45; .68]	[.62; .81]
4	.56	.72	.34	.60	.43	.60	.45	.71
	[.43; .66]	[.60; .80]	[.11; .52]	[.27; .76]	[.20; .60]	[.33; .75]	[.27; .59]	[.52; .81]
5	.64	.78	.36	.53	.58	.73	.41	.58
	[.53; .73]	[.70; .85]	[.21; .50]	[.35; .67]	[.45; .68]	[.62; .81]	[-01; .66]	[-.01; .79]
6	.39	.56	.09	.24	.25	.40	.06	.12
	[.23; .52]	[.38; .68]	[-.01; .21]	[-.02; .45]	[.09; .40]	[.17; .57]	[-.04; .19]	[-.09; .32]
7	.29	.45	.30	.46	.39	.56	.60	.75
	[.09; .46]	[.16; .63]	[.14; .44]	[.24; .61]	[.20; .54]	[.34; .70]	[.33; .75]	[.50; .85]
8	.30	.46	.32	.48	.21	.44	.39	.56
	[.03; .50]	[.06; .67]	[.16; .46]	[.27; .63]	[.04; .37]	[.11; .64]	[.09; .59]	[.17; .74]
9	.29	.45	.11	.26	.18	.31	.33	.49
	[.13; .43]	[.23; .61]	[.00; .23]	[-.01; .47]	[-.06; .40]	[-.12; .57]	[.08; .51]	[.14; .68]
10	.37	.54	.44	.70	.09	.23	.42	.60
	[.08; .56]	[.15; .72]	[.33; .54]	[.60; .78]	[-.01; .22]	[-.02; .46]	[.20; .59]	[.34; .74]
11	.53	.69	.26	.51	.53	.69	.56	.72
	[.36; .66]	[.53; .79]	[.11; .39]	[.28; .66]	[.29; .68]	[.45; .81]	[.41; .68]	[.58; .81]
12	.28	.44	.06	.16	.24	.39	.46	.63
	[.01; .49]	[.02; .66]	[-.01; .14]	[-.03; .33]	[.08; .39]	[.15; .56]	[.27; .60]	[.42; .75]
13	.34	.50	.46	.63	.69	.82	.54	.70
	[-.03; .58]	[-.07; .74]	[.31; .58]	[.48; .73]	[.59; .77]	[.74; .87]	[.41; .65]	[.58; .79]
14	.54	.78	.35	.52	.25	.50	.63	.77
	[.44; .63]	[.70; .84]	[.14; .52]	[.24; .68]	[.07; .41]	[.19; .68]	[.52; .72]	[.68; .84]
15	.46	.72	.51	.67	.42	.59	.19	.42
	[.36; .57]	[.62; .80]	[.36; .62]	[.53; .77]	[.27; .55]	[.43; .71]	[.05; .34]	[.13; .61]
16	.18	.30	.28	.54	.32	.58	.46	.63
	[.02; .33]	[.03; .49]	[.08; .45]	[.21; .71]	[.11; .49]	[.28; .74]	[.30; .59]	[.46; .74]
17	.33	.50	.21	.36	.29	.55	.47	.73
	[.18; .47]	[.30; .64]	[-.04; .44]	[-.08; .61]	[.18; .40]	[.40; .67]	[.34; .59]	[.61; .81]
18	.23	.38	.07	.14	.16	.28	.34	.50
	[.07; .39]	[.13; .56]	[-.04; .22]	[-.08; .35]	[.00; .32]	[.00; .48]	[.18; .48]	[.30; .65]
M	**.38**	**.56**	**.27**	**.45**	**.30**	**.47**	**.42**	**.58**
SD	**.13**	**.14**	**.14**	**.17**	**.17**	**.19**	**.15**	**17**
Generalizability coefficient	**.40**		**.26**		**.24**		**.39**	

Note. $N = 96$. Values in parentheses indicate 95% confidence intervals.

In line with prior studies (Jackson, Michaelides, Dewberry, & Kim et al., 2016; Putka & Hoffman, 2013), we aimed to generalize ratings across assessors and defined reliability in relative terms (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). So, sources of variance were regarded as reliable if they contributed to the similarity in participants' relative position compared to other participants based upon ratings from different assessors. Thus, participant main effects and participant * role-play interaction effects were considered as sources of reliable variance (see Lance, 2008; Putka & Hoffman, 2013; Speer et al., 2014). In contrast, assessor main effects, assessor * participant interaction effects, assessor * role-play interaction effects, as well as the assessor * participant * role-play interaction effect that is confounded with the residual were considered as sources of unreliable variance. Generalizability coefficient = expected single-rater reliability for any given role-play rating.

Table 5

Summary of Validity Analyses for Overall Score on Multiple, Speeded Simulations of Study 2

	3 minute observation time		1 minute observation time		
	High standardization	Low standardization	High standardization	Low standardization	Study 1 results
M	5.25	5.87	3.78	4.54	5.51
SD	0.65	0.63	0.51	0.66	0.73
Correlations					
GMA	.35**[a]	.28**[a]	.26*[a]	.30**[a]	.27**[a]
SJT	.34**[ab]	.39**[a]	.22*[c]	.26*[bc]	.32**[ac]
Extraversion	.37**[ab]	.39**[a]	.23*[c]	.28**[bcd]	.38**[ad]
Agreeableness	.25*[ab]	.26*[a]	.10[c]	.16[abc]	.24*[abc]
Openness	.08[ab]	.15[b]	.04[ab]	.03[a]	.11[ab]
Emotional Stability	.10[a]	.05[a]	.04[a]	.08[a]	.05[a]
Conscientiousness	-.06[a]	-.03[a]	.03[a]	.02[a]	-.06[a]
Criterion Performance	.56**[a]	.54**[a]	.43**[b]	.50**[ab]	.54**[ab]
Multiple regression					
R^2 *(adj.) of full model*	.41 (.34)	.40 (.33)	.34 (.27)	.38 (.30)	.38 (.31)
ΔR^2	.20	.19	.13	.17	.17
Sig. *F* change	< .001	< .001	< .001	< .001	< .001

Notes. The overall score on multiple, speeded simulations per condition was computed by aggregating ratings from all assessors and role-plays per condition. * $p < .05$, ** $p < .01$. [a-c]: correlations marked with different index letters within each row indicate significant differences in correlations via tests for difference of dependent correlations (Steiger, 1980). Given the multitude of these tests, differences should be interpreted with caution. ΔR^2 and Sig. *F* change refer to incremental validities of the overall score on multiple, speeded simulations over and above gender, age, GMA, Big Five, and SJT.

As a comparison, we also re-ran the multiple regressions with overall simulation scores that only aggregated ratings from one random assessor per role-play. We also re-ran the multiple regressions without controls. Results and conclusions were similar. Full multiple regression results of all four conditions as well as validities for single role-plays and for separate criterion components can be found in the Online Supplement.

Table 6

Evidence-based Recommendations For Multiple, Speeded Simulations

Design Considerations	Evidence
• Carefully determine the domain of interest.	Study 1
• Cover the domain comprehensively by including a large number of diverse behavioral job simulations (stimulus domain sampling). Do *not* use only one or two speeded simulations.	Study 1
• Ensure each speeded simulation's duration is between one to three minutes.	Study 2
• Ensure the total number of participants expected to sit the speeded simulations at least doubles the number of assessors used (e.g., 15 simulations require 15 assessors and thus at least 30 participants).	Study 1 (utility analysis)
Rating Process Considerations	
• Use role-playing assessors.	Study 1 vs. 2
• Hold role-play specific trainings for assessors and role-players because they are specialized in specific speeded simulations.	Study 2
• Focus on observing interactive behavior as a response to speeded simulations (response domain sampling) instead of assessing procedural knowledge of behavior.	Study 1
• Streamline the evaluation process by requiring assessors to evaluate only overall performance per speeded simulation. Next, aggregate ratings across assessors and speeded simulations. Do *not* rely on ratings of single assessors in single simulations.	Study 1 and 2

Appendix

Example of Isomorphic Situation across SJT and Speeded Simulation

Item stem of SJT

I fear that 16 June is a bad date because it is the same day as the F.A. cup final between Manchester United and Liverpool. I anticipate that this means 50 to 100 fewer people showing up for the Solidarity Run because many will want to see the match rather than do the Solidarity Run. I'm sure that some people from my unit, who attend the Solidarity Run every year, will hesitate about participating this year because of the F.A. cup final. I propose we postpone the event by one week.

Regards,

Paul

(followed by multiple-choice response options)

Situation given to participant at the start of speeded simulation 5

"Good morning, you sent an e-mail that today you wanted to discuss the fact that the F.A. cup final was on the same day as the event."

(followed by role-play)

Utility Analyses

Table A1

Utility Analysis Comparing Multiple, Speeded Simulations to Traditional Selection Procedures for Assessing 90 Participants

	18 speeded role-plays	1 role-play	1 structured interview	1 unstructured interview
N (Number of participants)	90	90	90	90
Validity (rxy)	0.54	0.18	0.28	0.21
Sdy	36634.4	36634.4	36634.4	36634.4
Ordinate Cut score	0.3485	0.3485	0.3485	0.3485
SR	0.33	0.33	0.33	0.33
C	$217,125	$28,125	$8,563	$6,563
Exercise development costs	$180,000	$15,000	$2,000	$0
Assessor training and fees	$37,125	$13,125	$6,563	$6,563
Total benefits	$1,880,244	$626,748	$974,941	$731,206
Total costs	$657,955	$85,227	$25,947	$19,886
Utility	**$1,222,289**	**$541,521**	**$948,994**	**$711,320**
Increase in utility of MSA as compared to traditional AC exercise		226%	129%	172%

Table A2

Cost Computation of Utility Analysis in Table A1

	18 speeded role-plays	1 role-play	1 structured interview	1 unstructured interview
Cost Computation				
Number of methods (e.g., role-play, interview)	18	1	1	1
Development cost of the method	$10,000	$15,000	$2,000	$0
Total: Development costs	**$180,000**	**$15,000**	**$2,000**	**$0**
Number of freelancers required (e.g., assessors, role-players, interviewers)	18	2	1	1
Training cost of 1 freelancer	$1,500	$1,500	$1,500	$1,500
Total: Personnel Training costs	**$27,000**	**$3,000**	**$1,500**	**$1,500**
Number of freelancers	18	2	1	1
Number of sessions needed for all participants to be rated	5	90	90	90
Number of hours needed per session per freelancer for observing and rating	1.5 [a]	0.75 [a]	0.75 [a]	0.75 [a]
Number of hours required per freelancer (for rating all participants)	7.5	67.5	67.5	67.5
Fee of freelancers per hour	**$75**	**$75**	**$75**	**$75**
Total: Personnel fees	**$10,125**	**$10,125**	**$5,063**	**$5,063**
Total: Personnel training and fees	**$37,125**	**$13,125**	**$6,563**	**$6,563**

Note. Costs that are fixed across the different approaches (e.g., participant recruitment, briefings, food & beverage) were not included. Costs were derived from input of consultancy firms.

[a] It was estimated that on average traditional role-plays last 45 minutes (30 minutes for the role-play and 15 minutes for rating). An interview was estimated to last about 45 minutes (including questioning and rating), whereas a speeded role-play was estimated to last 5 minutes (3 minutes for the role-play and 2 minutes for rating). Eighteen speeded role-plays then last 90 minutes.

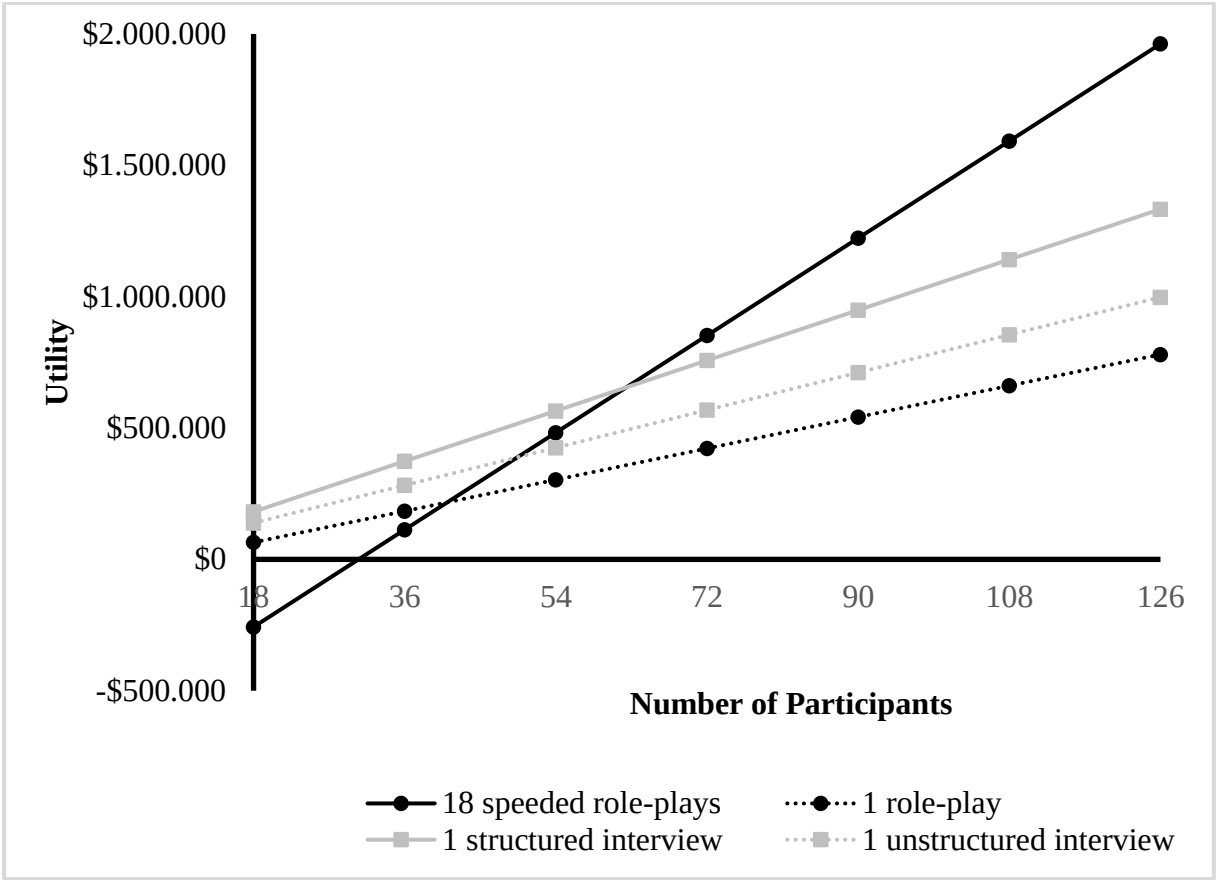

*Figure A1***.** Utility of Multiple, Speeded Assessments vs. Other Procedures for Differing Number of Participants.

Data Transparency Table

Parts of the data reported in this manuscript 1 (MS1) were collected as part of a larger data collection. Another paper, addressing the accuracy and criterion-related validity of automated scoring via text mining and machine learning in these multiple short simulations, is currently under review (MS2, submitted). Another paper (MS3) being currently in revision addresses interpersonal and temporal dynamics between role-players and participants in multiple short simulations as well as the nature of these dynamics (noise vs. substance) in terms of consequences for participants and organizations. Importantly, all three manuscripts focus on non-overlapping research questions and objectives.

The table below displays where each data variable appears per manuscript.

Variable	**MS1**	**MS2 (under review)**	**MS3 (in revision)**
GMA	X	X	
SJT	X		
Personality self-report	X	X	
Role-play performance rated by role-players	X	X	X
Role-play performance rated by remote assessor sample 1	X	X	X
Role-play performance rated by remote assessor sample 2	X		
Role-play performance rated by remote assessor sample 3	X		
Role-play performance rated by remote assessor sample 4	X		
Role-play performance derived by machine learning algorithms		X	
Dominance and affiliation (Continuous Assessment of Interpersonal Dynamics, CAID)			X
Dominance and affiliation (Social Behavior Inventory)			X
Criterion performance	X	X	X