Multiple speed assessments: Theory, practice, and research evidence

Christoph N. HERDE
Ghent University

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

Multiple Speed Assessments:

Theory, Practice, and Research Evidence

Christoph Nils Herde

Ghent University

Filip Lievens

Singapore Management University

Address of correspondence: Christoph Nils Herde, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium (christoph.herde@ugent.be). T: +32 9 264 64 55

Abstract

This paper presents Multiple Speed Assessments as an umbrella term to encompass a variety of approaches that include multiple (e.g., 20), short (e.g., 3 minutes) and often integrated interpersonal simulations to elicit overt behavior in a standardized way across participants. Multiple Speed Assessments can be used to get insight into the behavioral repertoire of a target person in situations sampled from a predefined target domain and their intraindividual variability across these situations. This paper outlines the characteristics and theoretical basis of Multiple Speed Assessments. We also discuss various already existing examples of Multiple Speed Assessments (Objective Structured Clinical Examinations, Multiple Mini Interviews, and constructed response multimedia tests) and provide an overview of design variations. Finally, we present current research evidence and future research directions related to Multiple Speed Assessments. Although we present Multiple Speed Assessments in the context of personnel selection, it can also be used for assessment in the educational, personality, or clinical psychology field

Keywords: Multiple Speed Assessment, assessment, personnel selection, intraindividual variability, adaptability

Multiple Speed Assessments:

Theory, Practice, and Research Evidence

These are exciting times for selection researchers and practitioners. Whereas for decades the same instruments (e.g., ability tests, personality inventories, interviews) were used (Cascio & Aguinis, 2008), recently times various new selection approaches and technology have emerged. Examples are screening people's social media content (Roth, Bobko, Van Iddekinge, & Thatcher, 2016) or the use of serious games (Fetzer, 2015). Another development has been the use of multiple short behavior observations in the form of short mini assessment center exercises (e.g., Brannick, 2008; Byham, 2016), or constructed response multimedia tests (e.g., Lievens, De Corte, & Westerveld, 2015). The rise of multiple short behavior observations is not exclusive to personnel selection but extends to other fields as well. In the healthcare context, for example, multiple short behavior observations are used within the Objective Structured Clinical Examination (OSCE; e.g., Brannick, Erol-Korkmaz, & Prewett, 2011) to certify or to screen medical students.

Although in each of these fields multiple short observations are used in different ways, across different contexts, and for different purposes, they all share the same common theme. However, a definition and description of those common characteristics is still missing. Moreover, their underlying theoretical basis has not been articulated. Therefore, this paper aims to make the following theoretical contributions. First, we connect different fields by formally presenting Multiple Speed Assessments as an umbrella term to encompass a variety of approaches that provide participants with multiple, short interpersonal simulations that elicit overt behavior in a standardized way. Second, we explicate the theoretical fundaments that are common to these different approaches. Third, we document the research evidence across these various applications and propose a research agenda to enhance our knowledge about Multiple Speed Assessments.

We start by outlining the key characteristics of the Multiple Speed Assessment approach and clarify the theoretical fundaments of Multiple Speed Assessments. Next, we show how Multiple Speed Assessments can be used as an umbrella term to include a variety of practices and approaches in different fields. Further, we outline different purposes of Multiple Speed Assessments. We also compare Multiple Speed Assessments to similar approaches. We end with presenting the available research evidence and an agenda for future research.

Multiple Speed Assessments: Definition and Characteristics

We define Multiple Speed Assessments as a standardized assessment approach that includes *multiple, short,* and often *integrated interpersonal simulations* to get insight into the behavioral repertoire of a target person in situations sampled from a predefined target domain. Examples are the leadership domain or the interpersonal domain.

Multiple Interpersonal Simulations

To elicit and evaluate participants' behavior, interpersonal simulations represent the hallmark of Multiple Speed Assessments because they allow obtaining samples of participants' actual, overt behavior in the targeted domain. These simulations present the same, standardized situations to all participants and require them to interact with a role-player.

The content of the interpersonal simulations is typically derived from two sources: First, subject matter experts that are familiar with the domain can be asked to generate critical incidents. Second, theoretical frameworks and taxonomies can be used. These taxonomies may be either frameworks that propose fundamental situational characteristics such as DIAMONDS (Rauthmann et al., 2014), Situation 5 (Ziegler, 2014), CAPTION (Parrigon, Woo, Tay, & Wang, 2017), or taxonomies that match the domain to be sampled. For example, interpersonal theory (Kiesler, 1983) might inspire the content of simulations that cover the

interpersonal domain, whereas leadership models such as the Multiple-Linkage Model (Yukl, 1989) might be relevant for simulations about leadership.

Taxonomies and theories can benefit test developers because they highlight which situational characteristics need to be varied across situations. For example, test developers may systematically vary role players' interpersonal disposition in terms of the two fundamental dimensions of dominance and affiliation (Kiesler, 1983) to sample the interpersonal domain. Across simulations, participants would then interact with dominant, submissive, friendly, and unfriendly role players (see Oliver, Hausdorf, Lievens, & Conlon, 2016).

Short Simulations

To obtain samples of a participant's behavioral repertoire in the domain, the *multiple* interpersonal simulations are used and these simulations are *short.* Although below, we provide more specific details about the number and duration of simulations, rules-of-thumb are that each simulation is less than five minutes and that – depending on the diversity of the domain and the situations one wants to cover – between 10 and 20 simulations are sampled. Accordingly, participants encounter a variety of real-life scenarios and characters that may appropriately mirror the domain within a feasible amount of time (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). Importantly, Multiple Speed Assessments thus do *not* degrade assessments to one single short simulation.

Structured Simulations

To ensure a reliable and valid assessment, it is crucial that participants show an adequate amount of relevant behaviors. However, the simulation's content and instructions alone might not guarantee to elicit multiple independent behavioral incidents because the interaction time between role-players and participants is limited in Multiple Speed Assessments.

To deal with this challenge and to ensure sufficient stimulus presentation consistency, role-players in Multiple Speed Assessments use situational cues (aka prompts) that activate relevant behaviors. Prompts are defined as specific actions or statements that are consistently presented across participants (Schollaert & Lievens, 2011, 2012). They are based on the principles of trait activation theory (see also below). The role of such prompts should go beyond ensuring structure and standardization and also facilitate the evaluation process. That is, prompts can be woven into the rating instrument, so that assessors rate participants' behavioral responses to the prompts (Brannick, 2008; Lievens, Schollaert, & Keen, 2015).

Streamlined Evaluation Process

In Multiple Speed Assessments, the evaluation process is streamlined. As one option to accomplish this, there might be only one single evaluation after each simulation to indicate the overall effectiveness of the participants' behavior (e.g., "How well did the participant handle the situation?"). Rating aids such as behavioral checklists or BARS can be used to ensure that observable and relevant behaviors are accounted for in this overall rating (Lievens, 1998). Another option to streamline the process is that the role-player also serves as assessor and vice versa, although one might also use a separate assessor (like in some OSCEs).

To reduce possible assessor related biases (e.g., carry over effects), role-players typically rate the same participant only once (or at best only a couple of times). The former implies that participants interact with one role-player in one simulation and would then go on to meet another role-player who starts the next simulation (see the carousel in Figure 1).

Despite this streamlined rating process, serving as a role-player as well as assessor is cognitively demanding. So, a thorough assessor/role-player training is required. This training builds on frame-of-reference training principles (Roch, Woehr, Mishra, & Kieszczynska, 2012) and thus includes prototypical examples of behaviors that are (in)effective in the given simulation and practice to exercise this via observation and rating aids. Moreover, training for

assessors who also act as role-players should also provide them with the standardized prompts that are used to elicit behavior (Lievens, Schollaert, et al., 2015).

Integrated Simulations

Multiple Speed Assessments often use multiple interpersonal simulations that are integrated and linked to each other via a broader overarching theme. That is, all simulations build upon one common pre-specified background. Examples of such a background context could be the organization of an event (e.g., a charity event, a conference), a move to another location, or the introduction of new administrative procedures (e.g., a digital booking tool in companies). To introduce the background, participants receive briefing documents. They can process this background via a quiz or in-basket prior to participating in the simulations.

Although it is not a necessity of simulations being integrated, this has several advantages. Such an overall context that is common to all simulations reduces the amount of background information that needs to be presented to participants via instructions prior to each simulation. In addition, integrated simulations contribute to higher realism (Lievens & Sackett, 2017), which may prompt participants to engage and immerse into the simulations (Fetzer, 2015). Yet, the common background of all simulations should not lead to performance in one simulation becoming dependent on the performance in a prior one. So, a simulation presents a key problem that is still relatively distinct from other simulations.

<div align="center">

Theoretical Fundaments of Multiple Speed Assessments

</div>

Zero/Minimal-Acquaintance Paradigm

The "zero/minimal-acquaintance" paradigm provides a first conceptual cornerstone for Multiple Speed Assessments. There exists a longstanding and voluminous body of research that asks untrained judges to rate strangers on the basis of minimal information, such as brief behavioral observations of under five minutes ("thin slices", see Back & Nestler, 2016; Funder, 2012). This research showed that such brief behavioral observations enable observers

to make accurate judgments that reveal valid information about a diverse set of outcomes, such as self- and other-ratings of personality, social relations and clinical outcomes, and performance in various fields (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992). In personnel selection, initial impressions have also been found to predict performance and employment decisions (Barrick et al., 2012; Barrick, Swider, & Stewart, 2010; Ingold, Dönni, & Lievens, in press).

Moreover, Multiple Speed Assessments build upon evidence that the accuracy of judgments of multiple variables do not necessarily increase with prolonged observations (Ambady & Rosenthal, 1992; Carney, Colvin, & Hall, 2007) and that observations of less than two minutes are indicative of longer behavioral streams (Murphy et al., 2015). Instead of longer observation time, it seems more beneficial to observe targets in a variety of situations (Back, Schmukle, & Egloff, 2009; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004) that allow to explore the behavioral repertoire (Leising & Bleidorn, 2011), and variability of behavior (Borkenau et al., 2004; Funder & Colvin, 1991; Leikas, Lönnqvist, & Verkasalo, 2014).

A caveat is in order, though: Zero-acquaintance studies differ from selection contexts in terms of contextual characteristics and type of behavior elicited. That is, zero-acquaintance predominantly elicit typical performance, whereas selection contexts activate maximum performance (Breil, Geukes, & Back, 2017; Sackett, Zedeck, & Fogli, 1988).

Trait Activation Theory

Evaluating people in short situations and basing judgments on "thin slices" of behavior runs the risk of not generating enough relevant behavior. To elicit a sufficient amount of relevant behavior among participants, Multiple Speed Assessments also draw from trait activation theory (Lievens, Tett, & Schleicher, 2009; Tett & Burnett, 2003). This theory posits that individual differences are more observable if situations a) aim to activate behavior

relevant for the target construct (i.e., situational trait relevance), and b) are not too strong so that individuals still construe the situation distinctly and, therefore, engage in different types of behavior (i.e., situational strength; see Meyer, Dalal, & Hermida, 2010).

Multiple Speed Assessments apply the principles of trait activation theory at two levels: At the overall simulation level, each simulation is designed to cover part of the target domain. At the within-simulation level, role-players present multiple standardized prompts (see above). The overall content of the simulation and the prompts are developed to introduce relevant mini-situations with the appropriate level of situational strength to elicit behavioral expressions related to the target domain. Accordingly, Multiple Speed Assessments aim to enhance the quality of information about participants' behavior that contribute to accurate judgments (Hirschmüller, Egloff, Schmukle, Nestler, & Back, 2015).

Principle of Aggregation

Apart from ensuring that relevant behavior is activated, the principle of aggregation (Epstein, 1979) serves as another safeguard in Multiple Speed Assessments. According to this principle, reliability increases if multiple behavioral observations are aggregated across many different occurrences or situations. Such an aggregation process maximizes the portion of systematic variance in behavioral ratings that is shared across situations (Epstein, 1979; Kuncel & Sackett, 2014). Likewise, behavioral ratings from single assessors are prone to assessor-specific error variance (idiosyncrasies). So, aggregating across behavioral ratings from multiple assessors should increase reliability (Eisenkraft, 2013).

<div align="center">

Examples of Multiple Speed Assessments

</div>

Objective Structured Clinical Examination

In the healthcare education context, multiple short behavior observations are used in OSCEs (Harden, Stevenson, Downie, & Wilson, 1975). The OSCE was introduced to enrich the assessment of clinical performance and communication of medical students. In the context

of certification, an OSCE presents students or residents with a large variety of clinical

scenarios that frequently involve standardized patients. For example, participants are asked to

assess a clinical history, perform physical examinations, or suggest the most appropriate

treatment.

Multiple Mini Interviews

Inspired by the OSCE, many healthcare education institutions have also introduced

Multiple Mini Interviews (MMI; Eva, Rosenfeld, Reiter, & Norman, 2004) to select

applicants for admission to study/residency programs. As the term MMIs suggests, applicants

participate in multiple short interviews. Yet, some MMIs also sample applicants' overt

behavior in short interpersonal simulations (Knorr & Hissbach, 2014).

Constructed Response Multimedia Tests

In the personnel selection context, constructed response multimedia tests have been

developed that present multiple short video clips to participants (e.g., De Soete, Lievens,

Oostrom, & Westerveld, 2013; Lievens, De Corte, et al., 2015; Oostrom, Born, Serlie, & van

der Molen, 2010). The actor in these video clips speaks directly into the camera. Once a video

fragment stops, participants have to respond as if they were to interact with the actor.

Participants' responses are then recorded via webcams.

Variations of Multiple Speed Assessments

These different examples illustrate that Multiple Speed Assessments can have a

different make-up, even though they share the same characteristics. Below, we discuss these

possible variations (see also Table 1 that matches these Multiple Speed Assessments onto key

predictor method factors, Lievens & Sackett, 2017).

Stimulus and Response Format

Multiple Speed Assessments can be administered in various stimulus and response

formats. One option is the face-to-face ("brick and mortar") test administration. Role-players

and participants then interact face-to-face with each other, with different simulations taking place at different tables in one large room or in separated rooms. This resembles the prototypical make up of OSCEs and MMIs (Knorr & Hissbach, 2014; Patrício, Julião, Fareleira, & Carneiro, 2013). As an alternative, online/remote/videoconference Multiple Speed Assessments take place as real-time interactions between role-players and participants via video chat. Initial evidence indicates that face-to-face and videoconference Multiple Speed Assessments produce similar results: Tiller et al. (2013) found no significant differences in MMI mean scores and comparable reliabilities and participant reactions. Moreover, cost savings for videoconference MMIs were about 84 %.

Whereas these earlier formats involve synchronous communication, participants might also watch standardized multimedia clips that introduce the problem situation and then immediately react upon each clip via a webcam. Although such asynchronicity precludes assessing dynamic interactions between role-players and participants, it might increase the efficiency of test administration. Recent research revealed that these constructed response multimedia tests provide valid assessments of future behavior (e.g., Cucina, Su, Busciglio, Harris Thomas, & Thompson Peyton, 2015; Lievens, De Corte, et al., 2015; Oostrom, Born, Serlie, & van der Molen, 2010, 2011).

Type of Domain

Multiple Speed Assessment comprehensively samples from a predefined domain through a variety of different interpersonal simulations that all activate domain relevant behavior but vary in terms of key situational characteristics. However, the type of domain can differ a lot. For example, constructed response multimedia tests have been developed to sample a diverse set of domains such as entry level police officer performance (Lievens, De Corte, et al., 2015) or interpersonal leadership (Oostrom et al., 2011).

Type of Simulations

Depending on the domain to be sampled, it is possible to rely only upon one type of simulation or to integrate different types of simulations to elicit domain relevant behavior. Examples of Multiple Speed Assessments with only one type of simulation are constructed response multimedia tests that consist of (asynchronous) role plays (e.g., Lievens, De Corte, et al., 2015). Examples of Multiple Speed Assessments with multiple different simulation types are MMIs that integrate role-plays, short presentations, fact findings, or other possible simulations with interviews (Knorr & Hissbach, 2014).

Number and Duration of Simulations

Multiple Speed Assessments use multiple simulations to comprehensively sample a prescribed domain. Reviews show that across different applications (a) the number of simulations varies between 3 and 40, (b) a simulation does not last longer than seven minutes, and (c) simulations of five to six minutes ensure reliable assessments[1] (Knorr & Hissbach, 2014; Patrício et al., 2013; Rees et al., 2016). In addition, decisions about the exact number and duration of simulations should always depend upon cost constraints, intended domain coverage, and desired score reliability (see Wang & Grimm, 2012).

<div align="center">

Purposes of Multiple Speed Assessment

</div>

Assessment of Overall Behavior Across Situations

During Multiple Speed Assessments, participants' overt behavior is observed and evaluated in multiple simulations that cover the target domain. Therefore, how people behave in each of these simulations gives an indication of their behavioral repertoire. An overall score can also be computed that averages behavioral evaluations across all simulations. As shown in Figure 2, this enhanced predictor domain coverage should allow good predictions of future behavior due to the higher point-to-point correspondence with the targeted domain (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

---

[1] Applicants also report being satisfied with a duration of six and eight minutes (Cameron & MacKeigan, 2012).

Assessment of Participants' Intraindividual Variability Across Situations

Apart from using participants' average score across all simulations, the behavioral observations per simulation (or across several simulations) can also be used for shedding light onto participants' intraindividual variability in behavior[2] across situations (Lievens et al., 2018). This fits in the emerging consensus that both people's consistency and within-person variability across situations are important. For example, the Cognitive-Affective Personality System Theory (Mischel & Shoda, 1995) posits that people's intraindividual variability across situations is not indicative of error variance but represents substantive variance in how people uniquely construe a specific situation and show subsequent behavior (see also Fleeson & Jayawickreme, 2015). As Multiple Speed Assessments sample a specific domain via simulations that systematically vary in terms of key situational characteristics, one can examine how participants vary their behavior across different situations such as different leadership (e.g., Yukl, 1989) or interpersonal demands (Kiesler, 1983). To examine whether variability across different simulations does indeed capture meaningful variability across different situations instead of error variance, variability indicators derived from Multiple Speed Assessments can be correlated with (a) validated indicators of variability, such as self- and other reports of adaptability or learning agility, or b) relevant outcomes, such as job or training performance (see Lievens, 2017).

An assessment of the following two aspects of people's intraindividual variability seem most promising (Baard, Rench, & Kozlowski, 2014; Jundt, Shoss, & Huang, 2015). First, Multiple Speed Assessments might be implemented for zooming into people's interpersonal adaptability across situations (Oliver & Lievens, 2014). As participants interact

---

[2] Although we refer to intraindividual variability in behavior, there is a link with performance. For example, if people vary and adapt their behavior in line with the situational demands (act more dominant as a leader, act friendlier as a team member), their performance will be high (with no variability). If they are not able to vary or adapt their behavior in this case (act dominant with Person A, act dominant with person B), their performance will vary (high in leadership situations, low in team situations).

with different role-players in different interpersonal situations, one can scrutinize how people vary and adapt their interpersonal behavior in line with the situational demands. Second, Multiple Speed Assessments allow assessing participants' learning agility (e.g., DeRue, Ashford, & Myers, 2012). That is, one might assess whether participants learn quickly from prior situations and improve along the entire Multiple Speed Assessment experience.

Application Areas

Multiple Speed Assessments can be used in a variety of assessment contexts. In this paper, we focus on the use of Multiple Speed Assessments in personnel selection and educational settings (e.g., OSCEs, MMIs). Yet, a Multiple Speed Assessment approach might also be used to inform research on interventions that influence short-term personality development (Roberts et al., 2017). Similarly, in clinical applications, patients can be asked to go through a large variety of role-plays to assess how they uniquely (e.g., rigidly) construe those situations and act upon those construals (Lievens, 2017).

Comparisons of Multiple Speed Assessments to Similar Approaches

Assessment Center Exercises and Situational Judgment Tests

We regard Multiple Speed Assessments as a hybrid (Lievens & Sackett, 2017) between assessment centers and traditional situational judgment tests. Both of these methods also require participants to respond to multiple situations that sample a target domain. However, as compared to assessment centers, Multiple Speed Assessments integrate overt behavioral stimuli (role-player actions) and responses (participants' behavioral reactions) from a larger number of simulations with a higher level of stimulus presentation consistency (standardized role-player prompts) and larger domain coverage (multiple short situations). Multiple Speed Assessments differ from traditional close-ended situational judgment tests by focusing on overt behavior and by using human assessors as raters.

Situational and Past Behavior Interview Questions

Situational and past behavior interview questions share basic characteristics with Multiple Speed Assessments but also differ considerably. Similar to Multiple Speed Assessments, such interview questions confront participants with multiple short situations. However, in contrast to Multiple Speed Assessments, interview questions do not sample overt behavior (with the exception of oral communication). Situational interview questions tend to assess job knowledge and past behavior interview questions seem to tap into job experience (Levashina, Hartwell, Morgeson, & Campion, 2014). Note also that all interview questions are usually asked and evaluated by only one (or sometimes two) interviewer, whereas Multiple Speed Assessments involve multiple role players (assessors).

Agenda For Future Research

Table 2 summarizes the empirical evidence on the various already existing Multiple Speed Assessments. Although generally the evidence is encouraging, knowledge gaps still exist. Therefore, we outline an agenda for future research on Multiple Speed Assessments.

Reliability of Multiple Speed Assessments

In Multiple Speed Assessments, role-players receive a thorough training, elicit multiple relevant behavioral acts with prompts, and use observation aids. In addition, Multiple Speed Assessments sample behavioral ratings of participants in a large diversity of situations that are provided by multiple assessors. This aggregation process aims to dissolve potential idiosyncrasies on behalf of assessors (Eisenkraft, 2013; Epstein, 1979). So, in light of the "law" of aggregation, the key point is that the overall Multiple Speed Assessment evaluation (thus aggregated across multiple situations) should serve as a reliable indicator of domain related behavior. Future research should disentangle the relative contribution of the reliable and unreliable variance components of Multiple Speed Assessment ratings. That is, one should examine the amount of variance that participants, assessors, simulations, and various forms of interactions among these sources explain (Jackson, Michaelides, Dewberry, & Kim,

2016; Putka & Hoffman, 2013). Such analyses help to understand why Multiple Speed Assessments "work". Is it because behavior is sampled across multiple simulations? Or because it is rated by different assessors? Or because aggregate behavioral ratings across simulations *and* assessors are used?

Validity, Added Value, and Utility of Multiple Speed Assessments

Multiple Speed Assessments use multiple simulations to comprehensively cover a predefined domain, which should ensure adequate levels of criterion-related validity of the overall aggregated rating. Besides this overall rating, Multiple Speed Assessments also introduce an economic way to obtain various indicators of people's intraindividual variability across the simulations. In any case, future research needs to determine the predictive validity of the aggregated ratings and indicators of intraindividual variability. At a more specific level, we should explore which domains can be best predicted by Multiple Speed Assessments. Does the behavior elicitation via interactions between role-players and participants lead to some domains (e.g., leadership and interpersonal domains) being better predicted than others (see research on the "good trait"; Back & Nestler, 2016; Funder, 2012)?

Given that Multiple Speed Assessments require considerable administrative and human resources, it is of interest to investigate how they relate to and add incremental validity above other simulation-based assessment methods to predict job performance. In fact, a crucial question is how short simulations that are the building blocks of Multiple Speed Assessments compare to a few long-lasting simulations that are usually applied in assessment centers in terms of predicting performance (with overall test-time held constant).

From a utility perspective, it is also key to investigate how additional investments in test-time and human resources affect the criterion-related validity of Multiple Speed Assessments. For example, does validity increase with a longer duration of each simulation?

Or does it increase by increasing the number of simulations and/or by increasing the number of assessors per simulation? When do such increases reach a tipping point?

Finally, future research should focus on validating Multiple Speed Assessments' evaluation of people's intraindividual variability. How does people's short-term behavioral variability within simulations and across simulations relate to their intraindividual variability as examined by experience sampling methods in the real world (Fleeson & Gallagher, 2009; Lievens et al., 2018) and to self- and other reports of interpersonal adaptability? How do different performance trajectories across simulations relate to self- and other-reports of learning agility or physiological indicators of stress resilience? If we identify concrete, stable situation-behavior linkages within Multiple Speed Assessments that relate to future job behavior, we will gain important knowledge about the utility of Multiple Speed Assessments. Moreover, this will also advance our understanding of intraindividual variability and its relation to outcomes such as adaptability, successful leadership, or psychological adjustment.

Participant Perceptions of Multiple Speed Assessments

Another avenue consists of examining how participants react to Multiple Speed Assessments. We need to know whether participants view multiple short simulations as face valid (i.e., resembling key characteristics of the target domain). Essentially, this means exploring whether test-takers perceive multiple short simulations as representative of today's fragmented and hectic world of work. Multiple Speed Assessments vividly introduce different situations and characters via multiple integrated simulations. Participants might therefore perceive this contextualized approach as realistic (Lievens & Sackett, 2017), which may increase their engagement and immersion into the situations (Fetzer, 2015).

A related question is whether participants feel to have sufficient opportunity "to show what they got" in Multiple Speed Assessments. On one hand, participants may perceive the short duration of simulations as an impediment to show relevant behavior. On the other hand,

in Multiple Speed Assessments they have multiple, independent chances to perform because they face different assessors in the simulations. Participants can thus compensate ineffective behaviors in a single simulation in other simulations. They also know that idiosyncratic biases from single assessors are averaged out in the overall rating.

Multiple Speed Assessments and Subgroup Differences

Especially in high-stakes testing situations, it is crucial to investigate whether Multiple Speed Assessments (dis)advantage participants of specific subgroups (in terms of gender, ethnicity, age, etc.). For example, does the interpersonal nature of simulations in Multiple Speed Assessments favor females because females score higher on extraversion and agreeableness (Costa, Terracciano, & McCrae, 2001; Feingold, 1994)? Does the hectic nature of Multiple Speed Assessments disadvantage older people? Given that Multiple Speed Assessments use short simulations, we need to find out whether assessors are more prone to stereotypes and biases based upon rapidly accessible stimuli like gender, age, or ethnicity.

<div align="center">**Conclusion**</div>

This paper formally presented Multiple Speed Assessments as an umbrella term to encompass a variety of approaches that include multiple, short, and often integrated simulations to get insight into the behavioral repertoire of a target person in situations sampled from a given domain. Multiple Speed Assessments aim to offer standardized behavioral-based assessments of people's performance in a given domain and their intraindividual variability across the various situations of that domain. Multiple Speed Assessments should encourage researchers and practitioners to better describe, explain, and predict behavior in today's fast-paced world.

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., & 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

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 University Press.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, *97*, 533–548. https://doi.org/10.1037/a0016229

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

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

Breil, S. M., Geukes, K., & Back, M. D. (2017). Using situational judgment tests and

assessment centres in personality psychology: Three suggestions. *European Journal of

Personality*, *31*, 442–443.

Byham, W. (2016). *Assessment centers for large populations*. Presented at the International

Congress on Assessment Center Methods, Bali, Indonesia.

Cameron, A. J., & MacKeigan, L. D. (2012). Development and pilot testing of a multiple

mini-interview for admission to a pharmacy degree program. *American Journal of

Pharmaceutical Education*, *76*, 10. https://doi.org/10.5688/ajpe76110

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

Cascio, W. F., & Aguinis, H. (2008). Staffing Twenty-first-century Organizations. *The

Academy of Management Annals*, *2*, 133–165.

https://doi.org/10.1080/19416520802211461

Casey, P. M., Goepfert, A. R., Espey, E. L., Hammoud, M. M., Kaczmarczyk, J. M., Katz, N.

T., … 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

Costa, P., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322–331. https://doi.org/10.1037//0022-3514.81.2.322

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

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity–validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment*, *21*, 239–250. https://doi.org/10.1111/ijsa.12034

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

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. doi:10.1046/j.1365-2923.2004.01776.x

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*, 429–456.

Fetzer, M. (2015). Serious games for talent selection and development. *TIP: The Industrial-Organizational Psychologist*, *52*, 117–125.

Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, *97*, 1097–1114. https://doi.org/10.1037/a0016786

Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

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

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60*, 773–794.

Harden, R. M., 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

Hirschmüller, S., Egloff, B., Schmukle, S. C., Nestler, S., & Back, M. D. (2015). Accurate judgments of neuroticism at zero acquaintance: A question of relevance. *Journal of Personality*, *83*, 221–228. https://doi.org/10.1111/jopy.12097

Husbands, A., Rodgerson, M. J., Dowell, J., & Patterson, F. (2015). Evaluating the validity of an integrity-based situational judgement test for medical school admissions. *BMC Medical Education*, *15*. https://doi.org/10.1186/s12909-015-0424-0

Ingold, P. V., Dönni, M., & Lievens, F. (in press). 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*.

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

Johnston, A. N. B., Weeks, B., Shuker, M.-A., Coyne, E., Niall, H., Mitchell, M., & Massey, D. (2017). Nursing students' perceptions of the Objective Structured Clinical Examination: An integrative review. *Clinical Simulation in Nursing*, *13*, 127–142. https://doi.org/10.1016/j.ecns.2016.11.002

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

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., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*, 38–47. https://doi.org/10.1037/a0034147

Leikas, S., Lönnqvist, J.-E., & Verkasalo, M. (2014). Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal behaviors. *Journal of Personality and Social Psychology*, *103*, 1007–1022.

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

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured

    employment interview: Narrative and quantitative review of the research literature.

    Personnel Psychology, *67*, 241–293. https://doi.org/10.1111/peps.12052

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A

    review. *International Journal of Selection and Assessment*, *6*, 141–152.

    https://doi.org/10.1111/1468-2389.00085

Lievens, F. (2017). Assessing personality–situation interplay in personnel selection: Toward

    more integration into personality research. *European Journal of Personality*, *31*, 424–

    440. https://doi.org/10.1002/per.2111

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of

    selection procedures: Effects of response fidelity on performance and validity. *Journal

    of Management*, *41*, 1604–1627. https://doi.org/10.1177/0149206312463941

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., & 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., 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 In J.J. Martocchio & H. Liao (Eds.), *Research in Personnel and Human Resources Management* (Vol. 28, pp. 99–152). Bingley: JAI Press.

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

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

Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., … 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

Oliver, T., Hausdorf, P., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management*, *42*, 1992–2017.

Oliver, T., & Lievens, F. (2014). Conceptualizing and assessing interpersonal adaptability. In D. Chan (Ed.), *Individual adaptability to changes at work: New directions in research* (pp. 52–72). New York: Taylor & Francis.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology*, *19*, 532–550. https://doi.org/10.1080/13594320903000005

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia

situational test with a constructed-response format: Its relationship with personality,

cognitive ability, job experience, and academic performance. *Journal of Personnel

Psychology*, *10*, 78–88. https://doi.org/10.1027/1866-5888/a000035

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*, 642-681.

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

Patterson, F., Tiffin, P. A., Lopes, S., & Zibarras, L. (2018). Unpacking the dark variance of

differential attainment on examinations in overseas graduates. *Medical Education, 52*,

736-746. https://doi.org/10.1111/medu.13605

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016).

The predictive validity of a situational judgement test and multiple-mini interview for

entry into postgraduate training in Australia. *BMC Medical Education*, *16*, 87.

https://doi.org/10.1186/s12909-016-0606-4

Pau, A., Jeevaratnam, K., Chen, Y. S., Fall, A. A., Khoo, C., & Nadarajah, V. D. (2013). The

Multiple Mini-Interview (MMI) for student selection in health professions training – A

systematic review. *Medical Teacher*, *35*, 1027–1041.

https://doi.org/10.3109/0142159X.2013.829912

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

Rauthmann, J., Gallardo-Pujol, D., Guillaume, E., Todd, E., Nave, C., Sherman, R. A., …

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

Rees, E. L., Hawarden, A. W., Dent, G., Hays, R., Bates, J., & Hassell, A. B. (2016).

Evidence regarding the utility of multiple mini-interview (MMI) for selection to

undergraduate health programs: A BEME systematic review: BEME Guide No. 37.

Medical Teacher, *38*, 443–455. https://doi.org/10.3109/0142159X.2016.1158799

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic

review of personality trait change through intervention. *Psychological Bulletin*, *143*,

117–141. https://doi.org/10.1037/bul0000088

Roberts, C., Clark, T., Burgess, A., Frommer, M., Grant, M., & Mossman, K. (2014). The

validity of a behavioural multiple-mini-interview within an assessment centre for

selection into specialty training. *BMC Medical Education*, *14*, 169.

https://doi.org/10.1186/1472-6920-14-169

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

Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in

employee-selection-related decisions: A research agenda for uncharted territory.

Journal of Management, *42*, 269–298. https://doi.org/10.1177/0149206313503018

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

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and

maximum job performance. *Journal of Applied Psychology*, *73*, 482–486.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach:

Selection test development based on a content-oriented strategy. *Personnel*

Psychology, *39*, 91–108. https://doi.org/10.1111/j.1744-6570.1986.tb00576.x

Schollaert, 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

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job

performance. *Journal of Applied Psychology*, *88*, 500–517.

https://doi.org/10.1037/0021-9010.88.3.500

Tiller, D., O'Mara, D., Rothnie, I., Dunn, S., Lee, L., & Roberts, C. (2013). Internet-based

multiple mini-interviews for candidate selection for graduate entry programmes.

Medical Education, *47*, 801–810. https://doi.org/10.1111/medu.12224

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with

standardized patients: State of the art. Teaching and Learning in Medicine, 2, 58–76.

https://doi.org/10.1080/10401339009539432

Wang, L., & Grimm, K. J. (2012). Investigating reliabilities of intraindividual variability

    indicators. *Multivariate Behavioral Research*, *47*, 771–802.

    https://doi.org/10.1080/00273171.2012.715842

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied*

    Psychology, *52*, 372. http://dx.doi.org/10.1037/h0026244

Wood, T. J., Chan, J., Humphrey-Murto, S., Pugh, D., & Touchie, C. (2017). The influence of

    first impressions on subsequent ratings within an OSCE station. *Advances in Health*

    Sciences Education, *22*, 969–983. https://doi.org/10.1007/s10459-016-9736-z

Woolf, K., Haq, I., McManus, I. C., Higham, J., & Dacre, J. (2008). Exploring the

    underperformance of male and minority ethnic medical students in first year clinical

    examinations. *Advances in Health Sciences Education*, *13*, 607–616.

    https://doi.org/10.1007/s10459-007-9067-1

Yukl, G. (1989). *Leadership in Organizations* (2nd ed). Englewood Cliffs, NJ: Prentice Hall.

Ziegler, M. (2014). *B5PS. Big Five Inventory of Personality in Occupational Situations.*

    Mödling, Austria: Schuhfried GmbH.

Table 1

Overview of Different Variations of Multiple Speed Assessments

	Constructed Response Multimedia Test	Objective Structured Clinical Examination	Multiple Mini Interview
Stimulus format	Dynamic audiovisual stimuli	Face-to-face interactive stimuli	
Contextualization	High contextualization		
Response format	Audiovisual constructed	Face-to-face interaction	
Response evaluation consistency	Calibrated judgment		
Information source	Behavior exhibited by the candidate		
Target sample	Job applicants (e.g., entry-level police officers)	Healthcare students, residents	Selection of applicants for (healthcare) study/residency programs
Type of simulations	(asynchronous) Role-plays	Clinical scenarios often involving standardized patients	Mainly interviews, but also role plays, fact finding exercises, presentations, etc.
Domain	Job-related behavior, interpersonal leadership	Clinical performance and communication in healthcare settings	required behavioral repertoire for healthcare programs and prospective job
Number of simulations	4-24	4-40 (Patrício et al., 2013)	3-12 (Knorr & Hissbach, 2014), mean: 9.2 (Rees et al., 2016)
Duration of simulations	≤ 5 min	6-20 min most frequently 3-6 min (Patrício et al., 2013)	5-15 min (Knorr & Hissbach, 2014) mean: 7.3 min (Rees et al., 2016)

Note. The descriptions of Multiple Speed Assessments resemble prototypical examples. OSCEs do traditionally complement behavioral based "procedure" stations with "question" stations that require participants to answer questions about previous procedure stations (Harden et al., 1975). In this table, we only refer to procedure stations because question stations do not sample overt behavior.

Table 2

Summary of Empirical Evidence for Different Examples of Multiple Speed Assessment

|	Constructed Response Multimedia Test	OSCE	MMI
Reliability			
Can assessors make reliable ratings of behavior in short simulations? How does the use of prompts increase the reliability of the ratings?	Inter-rater reliability: $.68 \leq ICC \leq .92$ (Cucina et al., 2015; DeSoete et al., 2013; Lievens et al., 2015; Oostrom et al., 2010, 2011)	Inter-rater reliability: $.20 \leq r \leq .95$ (Casey et al., 2009)	Inter-rater reliability: $.54 \leq ICC \leq .83$; $.74 \leq \alpha \leq .84$; $.62 \leq r \leq .91$; $.52 \leq G \leq .85$ (Knorr & Hissbach, 2014)
Are behavioral ratings aggregated across multiple simulations reliable?	Internal consistency: $.80 \leq \alpha \leq .83$ (Lievens et al., 2015; Oostrom et al., 2010, 2011)	Internal consistency: $\alpha = .62$ $G = .49$ (Brannick et al., 2011)	Internal consistency: $.61 \leq \alpha \leq .96$ $.32 \leq G \leq .88$ Test-retest reliability: $.34 \leq r \leq .70$ (Knorr & Hissbach, 2014)
What is the relative contribution of different reliable and unreliable variance components (i.e., assessors, simulations, etc.) to Multiple Speed Assessment ratings?	ICCs increase from using 1 to 3 raters (Cucina et al., 2015)	Main source of measurement error: variation in participants' performance across stations (Van der Vleuten & Swanson, 1990) Adding stations may be more efficient than adding raters (Brannick et al., 2011)	Variance attributable to candidate differences: 10-74%, frequently < 30 % Increasing number of stations has larger impact on reliability than increasing number of assessors Similar reliabilities for 5/6 vs. 8 minute station MMIs (Knorr & Hissbach, 2014)
Validity and Added Value			
How well do Multiple Speed Assessments predict performance?	Selection decision $r = .24*$ (DeSoete et al., 2013) $r = .31*$ (Lievens et al., 2015) Objective measures of job performance $r = .15*$ (Cucina et al., 2015) $r = .26*$ (Oostrom et al., 2010) Supervisor ratings $r = .01$ (Cucina et al., 2015) $r = .13$ (Oostrom et al., 2010) Training performance $r = .12*$ (Cucina et al., 2015) $r = .26*/.30*$ (Lievens et al., 2015)	Variable evidence from low to high correlations (e.g., Casey et al., 2009; Rushforth, 2007)	In-programme performance $-.05 \leq r \leq .57*$ Post-graduation performance $-.10 \leq r \leq .65*$ (Knorr & Hissbach, 2014)

How do Multiple Speed Assessments relate to other forms of simulation-based assessment methods (assessment center exercises, situational judgment tests, etc.)?	Written constructed response multimedia test $r = .41*$ (Lievens et al., 2015) Single role play $r = .39*$ (DeSoete et al., 2013)		Relation between two MMIs: $r = .75$ Constructed response multimedia test Audio/textual response format: $r = .15/.51$ (Knorr & Hissbach, 2014) SJTs $.26* \leq r \leq .53*$ (Husbands et al., 2015; Patterson et al., 2016; Roberts et al., 2014)
Do Multiple Speed Assessments add incremental validity above traditional tests and traditional simulation-based assessment methods?	Job placement success $\Delta R^2 = .04*$ Supervisor ratings $\Delta R^2 = .00$ (Oostrom et al., 2010) Selection decision $\Delta R^2 = .03*$ (Lievens et al., 2015) Training performance $\Delta R^2 = .03-.08*$ (Lievens et al., 2015)		In-programme and licensing examination performance (Knorr & Hissbach, 2014) Applied knowledge test $\Delta R^2 = .01$ Clinical decision making skills examination $\Delta R^2 = .02*$ OSCE $\Delta R^2 = .10*$ (Patterson et al., 2016)
Participant perceptions			
Are multiple short simulations regarded as face valid?		Positive perceptions (e.g., Johnston et al., 2017; Rushforth, 2007)	Positive perceptions, participants prefer MMIs to traditional interviews (Rees et al., 2016)
Do participants view multiple short simulations as procedurally fair and as providing good opportunity to perform?		Tentative evidence that students acknowledge procedural fairness and opportunity to perform, but perceive time as inadequate (e.g., Johnston et al., 2017; Rushforth, 2007)	Mixed evidence regarding satisfaction with time per station Tentative evidence that participants appreciate stations offering "clean slates" (Rees et al., 2016) and that participants identify good opportunities to perform (Pau et al., 2013)
Subgroup differences			
Do Multiple Speed Assessments favor subgroups related to gender, age, or ethnicity?	Gender $-.31* \leq d \leq .24$ (Cucina et al., 2015; DeSoete et al., 2013; Lievens et al., 2015; Oostrom et al., 2010, 2011) Age $-.14 \leq r \leq .23*$ (DeSoete et al., 2013; Oostrom et al., 2010, 2011)	Gender Females seem to outscore males (e.g., Woolf et al., 2008; average $d = .37*$) Age $r = -.33*$ (Patterson et al., 2018) Ethnic majority vs. minority Ethnic minority seems to score lower (e.g., Woolf et al., 2008;	Majority of studies indicates equal scores across gender, age, or socio-economic subgroups (Rees et al., 2016)

|		
	Ethnic majority vs. minority d = .14 (DeSoete et al., 2013) d = .44* (Lievens et al., 2015) White-Black -.10 ≤ d ≤ .00 White-Hispanic .11* ≤ d ≤ .22* (Cucina et al., 2015)	average d = .27*)
Do short simulations increase the relative influence of stereotypes/heuristics/biases in assessors' judgments?		First impressions show at least moderate level of accuracy Relations of first impressions with systematic evaluation: r = .83*, and with expert rating: r = .59 (Wood et al., 2017)

Note. * *p* < .05. Results in this table are uncorrected. Positive *d* coefficients indicate higher scores for females, Whites, and ethnic majority members.

Figure 1. Schematic example of a Multiple Speed Assessment.
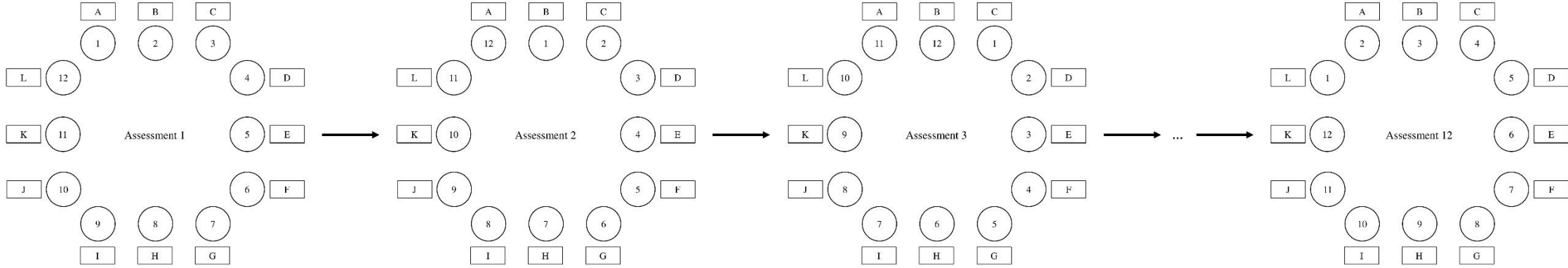
In this example, 12 assessees (circles) simultaneously walk through a Multiple Speed Assessment that contains 12 simulations (rectangles). After

each simulation, each participant goes on to a different simulation where they face again a role-player. Role-players may be seated on different

tables or in different (virtual) rooms. This procedure repeats until all participants participated in all simulations.

Figure 2. Schematic illustration of domain sampling by Multiple Speed Assessments compared to traditional approaches.
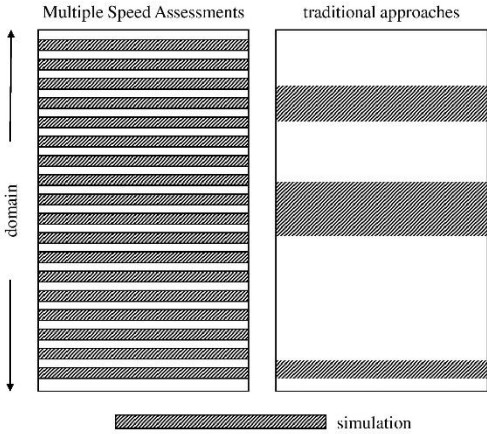
The short duration of each simulation enables Multiple Speed Assessments to sample the domain more comprehensively than traditional approaches that build upon long-lasting, but fewer simulations.