FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Multiple Speed Assessment:

A new approach for measuring interpersonal performance and adaptability?

Christoph Nils Herde

Supervisor: Prof. Dr. Filip Lievens

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2019–2020

GHENT
UNIVERSITY

FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Multiple Speed Assessment:

A new approach for measuring interpersonal performance and adaptability?

Christoph Nils Herde

Supervisor: Prof. Dr. Filip Lievens

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2019–2020

GHENT
UNIVERSITY

ACKNOWLEDGEMENTS

Working for my PhD often felt like a journey across unknown territory. Although I started my journey with excitement and confidence, it was clearly not a walk in the park. Sometimes I faced bumpy roads or obstacles, sometimes I had to change directions. Just like sudden rain can surprise you on a sunny day, external influences sometimes confronted me with unexpected challenges I had to master. Sometimes I was questioning whether I would ever arrive at the final destination. Being close to successfully arriving at the end of my journey, I feel grateful and I appreciate the opportunity to experience all the positive and difficult situations, the lessons I learned and all the different places I visited. Literally, parts of this PhD were prepared, written, or presented in Ghent, Duisburg, Anaheim, Leipzig, Leuven, Orlando, London, Cambridge, Stuttgart, Chicago, Barcelona, Geneva, Boston, New York City, as well as several other places in California, Florida, Canada and South East Asia.

For a journey to be enjoyable and successful, you definitely need advice and support from experienced guides who know about the most exciting places, know how to get there, but also know about possible traps and obstacles across the territory. *Filip*, thanks for equipping me with true purpose and showing me paths worth exploring. Thanks for giving me suggestions which areas might be the most interesting to investigate, and for opening doors into various exciting directions. Thanks for sharing so much from your experience and knowledge that helped me to pass frightening obstacles or to successfully reach the destination of my journey towards a PhD and contributing to research. Thanks for investing so much of your precious time into my professional and personal development. Thanks for all the inspiration you gave, thanks for answering questions whenever they arose, and thanks for always giving me the feeling of being highly appreciated. Finally, thanks for showing future possible paths that might be worth exploring.

In this line, thanks also to all members of my doctoral guidance committee. Thanks to *Jonas Lang*, *Frederik Anseel*, *Filip De Fruyt* and *Brian Hoffman* for involving in vivid discussions, for providing helpful feedback and insights that further inspired this project. Special thanks to *Jonas* for much appreciated statistical advice. Further, many thanks to *Myrjam Van de Vijver*. Thanks for all your various contributions to the project, especially regarding the data collection. Thanks for always being open to share your helpful insights.

I also gratefully thank several selection and development practitioners who supported the realization of this project. Thanks to the wonderful people at Hudson. In particular, thanks to *Ellen Volckaert* and *Amelie Vrijdags*. Thank you for supporting this project in so many different ways and for sharing your practitioner's point of view. Thank you and everyone at Hudson for contributing to the fundaments of the data collection. In a similar vein, thanks to SHL and Laureate International Universities. In particular, thanks to *Emily Solberg*, *Jan Harbaugh*, *Mark Strong* and *Gary Burkholder*. Thanks for offering the opportunity to contribute to an exciting project. Thanks for your trust in me and thanks for understanding that research is sometimes an iterative process of going forward, going sideward, or going backwards before the best solution for a problem is ultimately found. On top of that, thanks to *Kimberly Pauwels* and *Sara Teuwen*. Thanks for supporting this project and thanks for contributing to the data collection.

Further, thanks to all students who contributed to this project. *Sofie Ameloot*, thanks for supporting the fundaments of the projects by spending endless hours on watching and cutting video records. *Robin Boudry*, thanks for all your various contributions to the project. Without your technical knowledge and skills, service orientation, and innovative ideas to solve unusual problems, an efficient preparation of the data collections would have been much more difficult. Thanks to all students who supported this project as part of their master thesis and contributed to data codings: *Lisa Van der Schueren*, *Chrissie Polfliet*, *Sigrid*

Lefevre, *Nefert Viaene*, *Afra Vanhalst* and *Ellen Bagdasarian*. Your intense efforts to code data contributed largely to this project and the insights gained. It was a pleasure to see how all of you developed across the course of your thesis projects by gaining new knowledge and further developing your skills.

Going on an exciting journey on your own, however, is boring, lonely, and might leave you much more puzzled or helpless when you face obstacles. Instead, it is great to go the way together with others who enjoy to explore similar areas, face similar challenges, and intent to arrive at the same destination. People with whom you can share your thoughts, hopes, and concerns. People who make similar experiences that can help you to pass obstacles. People who make you laugh and put a smile on your face.

Malte, thanks for being such good company across the four years. I enjoyed all of our conversations to shortly detach from research or to dive further into details about assessment, methods, and publishing. Thanks also for all the input you gave to help me integrate into a new country, new culture, and new department. For similar things, thanks *Gudrun* for great conversations about the joy and struggles of a PhD student, for helping out generously when I had a very stressful period, and thanks for broadening my horizon by sometimes sharing unconventional ways to look at certain phenomena.

Many thanks to all of you who shared an office with me. *Catherine, Julie, Saartje* – you made it easy to feel welcome right from day one. I liked how different occupations of our office created very distinct atmospheres. Also thanks for accepting my language barrier and contributing so much to overcome it little by little. Entering this office filled me with joy every time across the four years. In the same way, thanks to *Saar* and *Sander* who I always enjoyed to meet in the office.

Further, thanks to all members of the *professorial staff of PP09* who provided a warm welcome and contributed to a dynamic work environment. Thank you, *Bart Wille*, *Bert*

Weijters, Eva Derous, Katia Levecque, Peter Vlerick, and *Johnny Fontaine*. Special thanks to *Jan, Lien, Anneleen, Sam, Roeliene, Elias* and *Céline*. I always enjoyed chatting with you in the lunchroom or in the hallway. *Jan*, it was great to have conversations about life on and off the job, how to balance both domains and it was always easy to have a great time and a laugh with you. *Lien*, it was so much fun to experience the adventure of the first SIOP conference together. *Anneleen*, thanks for always asking about the current status quo, for listening, and for answering questions. Without any doubt, PP09 has always been a dynamic department with some great people. Thanks to *all past, current, and new members of the department* that I unfortunately met too rarely because I sometimes focused too much on work.

Even with a great guide and mentor, experts who provide academic input and good company across the road, you would never dare to go on an adventurous journey like striving for a PhD without having proper support from people who belief in you and who help you to go this way to the very end.

Words cannot express the thankfulness and gratitude I feel for my wife, *Katharina*. Thanks for accepting and always supporting my striving for this PhD. A husband who is striving for a PhD and sometimes loses himself in perfectionism is certainly not always easy to deal with. Thanks for your patience. Thanks for listening to my problems, thanks for listening to me when my thoughts were running in circles, thanks for listening to me when I was complaining about unsolvable problems, thanks for listening to me while I was finding out that there was no problem at all and thanks for showing me that there can be a solution to every problem. Spending time with you after long working hours always refreshes my mind, fills me with joy, makes me laugh, or calms me down. Thanks for cultivating and sharing the passion for food and travel. Thanks for enriching my life the way you do.

Thanks to my parents *Silvia* and *Rainer*. Thanks for raising me in the way you did, in a warm, save and enriching environment full of love and support. Thanks for always facilitating

my curiosity, for supporting me to develop my own way of looking at phenomena and for providing opportunities to further develop and satisfy my hungry mind. Thanks for continuously encouraging me to start this journey even though I was returning early and dissatisfied from a previous one. Thanks as well to my parents-in-law *Elisabeth* and *Bernd*. Thanks for accepting and warmly welcoming me as a member of your family. Thanks for showing real interest in my work and supporting my unusual journey. Thanks for generously offering support in everyday life to Katharina and me. Thanks for sometimes pushing or kindly forcing me to accept pragmatic solutions for daily hassles.

Finally, thanks to my sisters *Carina*, *Jessica*, *Katja*, my brother *Berni*, and my wife's brothers *Max* and *Christian* as well as to *all their partners and children* and my godmother *Ulla*. Every one of you is unique but you are all the same kind of wonderful. Thanks for always asking about my progress, thanks for listening to my struggles. Thanks to contributing to a vivid, exciting life off the job. Thanks for sometimes distorting the flow of ordinary everyday life with (mostly) enjoyable surprises. Thanks for contributing to great detachment from work by creating moments full of energy and joy to re-charge my batteries. All of the people of my family – there are more beautiful or more exciting places on earth – but there is no other place that you have established and continuously nurture as a place I call "Heimat".

Christoph Herde

Duisburg, September 2019

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Dissertation Objectives

The current state of work life confronts employees with the challenge to perform well in interpersonal situations and to adapt to different (interpersonal) demands. This is due to an increased frequency and complexity of interpersonal interactions on the job (see, for example, Griffin, Neal, & Parker, 2007; Pulakos, Arad, Donovan, & Plamondon, 2000) that has been caused by a number of trends in today's world of work. These trends include the ongoing and ever increasing globalization that more often and more easily brings people together from different countries and cultural backgrounds (e.g., Cascio, 2003; Javidan, Dorfman, de Luque, & House, 2006), a general shift to a more knowledge-based society in Western economies that incorporates more service-oriented businesses (e.g., Zeithaml & Bitner, 1996), or the shift to more project-based work in which employees frequently need to work together with newly formed teams (e.g., Hesketh & Neal, 1999; Kozlowski, Gully, Salas, & Cannon-Bowers, 1996).

To successfully master these challenges, organizations might apply personnel selection and development procedures that contribute to a better equipped workforce. In detail, organizations might assess how participants perform in situations that involve interpersonal encounters and require adaptation of one's behavior to different (interpersonal) demands. One of the possible approaches in personnel selection and development are simulation-based procedures. Simulation-based procedures build upon the principle of behavioral consistency. That is, as far as the simulations represent critical elements of situations that are faced on the job, it is assumed that behavior observed in the simulations predicts behavior in these situations on the job (e.g., Wernimont & Campbell, 1968; see also Lievens & DeSoete, 2012). To assess applicants' or employees' performance in interpersonal situations or their

adaptation to different interaction partners, one might thus develop simulation-based procedures that sample various job-related situations with different interpersonal encounters.

In a simplified manner, one might cluster different forms of simulation-based procedures by their fidelity (e.g., Goldstein, Zedeck, & Schneider, 1993; see also Lievens & De Soete, 2012). Fidelity describes the degree to which a simulation captures the targeted job-related situation (physical fidelity), or more specifically, the degree to which a simulation captures the targeted job-related situation in terms of (a) knowledge, characteristics, and abilities that are indeed required in the job-related situation, (b) the response mode by which situations on the job need to be handled, and (c) knowledge, characteristics, and abilities that are not required in the job-related situation (psychological fidelity; Goldstein et al., 1993; see also Lievens & De Soete, 2012). As a simplified distinction, one can distinguish between low- and high-fidelity simulations. Low-fidelity simulations assess participants' procedural knowledge or behavioral intentions in a specific domain, whereas high-fidelity simulations sample actual behaviors from participants (e.g., Thornton III & Rupp, 2006).

Examples for low-fidelity simulations are Situational Judgment Tests (SJTs) that traditionally present participants with written, high-contextualized descriptions of job-related situations and a number of possible response options. Participants then usually need to rate, rank, or choose a best or worst option to deal with the depicted situation (Motowidlo, Dunnette, & Carter, 1990). Across the last decades, SJTs gained a track record of criterion-related validity (Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb III, 2007) and have generated positive reactions from participants (Kanning, Grewe, Hollenberg, & Hadouch, 2006).

In line with the low-fidelity paradigm and the notion of behavioral consistency, SJTs have traditionally been designed to assess procedural knowledge or behavioral intentions related to a heterogeneous set of behaviors that sample a specific criterion-domain (e.g., a

specific job-field). Thus, different items of the same SJT were often considered to assess procedural knowledge about a heterogeneous set of constructs (e.g., Schmitt & Chan, 2006). Often, SJTs do even present construct heterogeneous response options for a single situation (e.g., Motowidlo, Crook, Kell, & Naemi, 2009; Whetzel & McDaniel, 2009).

Recently, however, researchers have intensified efforts to investigate SJTs that focus on the assessment of procedural knowledge related to one single construct each (e.g., Lievens & Motowidlo, 2016; Motowidlo, Hooper, & Jackson, 2006). Compared to traditional SJTs that sample a construct-heterogenous set of behaviors, construct-driven SJTs that focus on the assessment of a single, pre-specified construct imply several conceptual advantages. These advantages include a stronger theoretical fundament to the development and interpretation of the content validity of SJTs and easier to interpretable patterns of relations to corresponding or non-corresponding measures to facilitate construct-validation approaches (Lievens & Motowidlo, 2016). Thereby, a better understanding of what is being measured in SJTs might be obtained that might further benefit the design of training interventions. Finally, construct-driven SJTs might be more broadly applicable in practice, because SJTs that assess procedural knowledge in a specific construct-domain might be more generic than SJTs that sample procedural knowledge about critical situations in specific job-fields (Lievens & Motowidlo, 2016). Recent empirical research has shown positive evidence for the construct-related and criterion-related validity of construct-driven SJTs (e.g., Bledow & Frese, 2009; Mussel, Gatzka, & Hewig, 2018; Oostrom, de Vries, & de Wit, 2019).

Given that SJTs can be economically administered via paper-pencil or computer-based formats, they might in principle be well-suited to let organizations assess applicants or employees across different geographical regions. Such a procedure to sample from global talent pools appears crucial for organizational success in today's war for talent (see Cascio & Aguinis, 2008). However, given their high-level of contextualization, one might wonder

whether SJTs can indeed be developed in a way that participants from different regional or cultural groups perceive test items in a similar way and attribute equal meaning to them (Lievens, 2006; Ployhart & Weekley, 2006). Until now, empirical evidence for the cross-cultural transportability of SJTs is mixed (Lievens, Corstjens et al., 2015; Such & Schmidt, 2004). However, past research only applied SJTs that were designed in a specific reference country and then investigated their cross-cultural/regional transportability into other countries (imposed etic approach). In the field of personality psychology, several examples attested to the so-called combined emic-etic approach that integrates cross-regional/cultural input across all stages of test development (Cheung, Fan, Cheung, & Leung, 2008; Schmit, Kihm, & Robie, 2000). Until now, no empirical studies have been conducted with a combined emic-etic approach to the development of simulation-based selection procedures such as (construct-driven) SJTs. Such an approach is untested so far for SJTs, but it might serve to develop SJTs for assessing procedural knowledge about interpersonal performance and (interpersonal adaptability) across geographical regions. Hence, we propose Objective 1 of this dissertation:

Objective 1: Provide an empirical test of the combined emic-etic approach to develop SJTs to measure procedural knowledge about interpersonal performance and (interpersonal) adaptability across geographical regions.

High-fidelity simulations such as assessment center exercises have frequently been applied to provide insights into assessments of performance in interpersonal settings, and have gained a track record of criterion-related validity (e.g., Becker, Höft, Holzenkamp, & Spinath, 2011; Gaugler, Rosenthal, Thornton III, & Bentson, 1987; Hermelin, Lievens, & Robertson, 2007; Hoffman, Kennedy, LoPilato, Monahan, & Lance, 2015). In recent years, selection and development practitioners have also added new high-fidelity simulations to their portfolio to respond to calls for short, fast-paced, and more engaging assessment experiences that mirror today's hectic and fragmented work life (Liff, 2017). For example, multiple short

interpersonal simulations are used in different forms and with different design variations to sample individuals' behavioral repertoire in a predefined domain. Based upon this overarching principle, these different approaches might be captured under the umbrella term of Multiple Speed Assessments. Examples for Multiple Speed Assessments are Objective Structured Clinical Examinations that are most prominently used in the context of certification of medical students (Brannick, Erol-Korkmaz, & Prewett, 2011), Multiple Mini-Interviews that are most prominently used in the context of selection of medical students (Knorr & Hissbach, 2014), or constructed response multimedia tests that are most prominently used in the context of personnel selection research and practice (e.g., Lievens, De Corte, & Westerveld, 2015; Pinsight, 2018). However, given that these different approaches have been developed and used in different fields, a clear definition as well as an overview of their shared characteristics is currently missing. On top of that, their common theoretical fundament has not been formulated so far. Finally, future progress in research and practice might further be facilitated by an overview of the different design variations of Multiple Speed Assessments as well as possible application areas and an agenda for future research. Hence, we propose Objective 2 of this dissertation:

Objective 2: Provide a conceptual overview of Multiple Speed Assessments, including their shared characteristics, theoretical fundaments, possible design variations as well as application areas, and an agenda for future research.

Although specific forms of Multiple Speed Assessments have already been used and evaluated in the context of medical student selection and certification, there is a lack of empirical investigations of face-to-face Multiple Speed Assessments that aim to capture the domain of job- or leadership-related performance that includes interpersonal performance and adaptability. Therefore, we currently lack knowledge about the reliability and validity of Multiple Speed Assessments that aim to assess participants' job- or leadership-related

performance with special regards to interpersonal criteria and (interpersonal) adaptability. For example, can judgments made in short interpersonal simulations show sufficient interrater reliability? How do different sources of variance, such as participant main effects or participants x simulation interaction effects contribute to reliable variance in Multiple Speed Assessments? How many independent assessors and simulations are necessary to gain an overall reliable estimate of participants' performance in the targeted domain? Do judgments in short simulations reveal meaningful information about participants' individual differences such as cognitive ability and personality? Do judgments in short simulations predict performance in the targeted domain and add incremental variance beyond more traditional predictors? To answer these questions, we propose Objective 3 of this dissertation:

Objective 3: Provide knowledge about the reliability and validity of a face-to-face format of Multiple Speed Assessments to sample the leadership domain, which includes components of interpersonal performance and (interpersonal) adaptability.

Finally, although high-fidelity simulations have frequently been applied to provide insights into assessments of performance in interpersonal settings and have gained a track record of criterion-related validity (Becker et al., 2011; Gaugler et al., 1987; Hermelin et al., 2007; Hoffman et al., 2015), the interpersonal dynamics that occur in high-fidelity simulations between participants and other human actors have been rarely studied in the past (Lievens & Klimoski, 2001). As a rare exception, Oliver, Hausdorf, Lievens, and Conlon (2016) showed that participants' interpersonal behavior in high-fidelity simulations is influenced by the interpersonal behavior of role-players and task-related situational demands. Further, Oliver et al. (2016) examined how participants' interpersonal behavior relates to performance ratings in high-fidelity simulations as a function of different interpersonal and task demands. As a limitation, however, past investigations of interpersonal behavior and dynamics in high-fidelity simulations have not acknowledged that interpersonal behavior shows substantial

intraindividual variability across time on a continuous moment-to-moment level (Markey, Lowmaster, & Eichler, 2010; Sadler, Ethier, Gunn, Duong, & Woody, 2009; Tracey, 2004). This is due to the fact that past studies have usually assessed interpersonal behavior with single-point estimates that only assign a single, overall score of behavior during a specific high-fidelity simulation. Given that the intraindividual variability in interpersonal behavior also likely drives interpersonal dynamics between human actors, we currently lack accurate knowledge about the nature of interpersonal dynamics in high-fidelity simulations as well as their relation to performance ratings in high-fidelity simulations and job-related performance. In a related manner, one might argue that the interpersonal dynamics that are shown within high-fidelity simulations might indicate how individuals adapt their interpersonal behavior to different interpersonal demands or human actors, which matches the core of definitions of interpersonal adaptability (see Oliver & Lievens, 2014; Pincus et al., 2014; Sadler, Ethier, & Woody, 2011 for similar arguments). We thus propose Objective 4 of this dissertation:

Objective 4: Provide knowledge about the interpersonal behavior of participants and the interpersonal dynamics they establish with other human actors in high-fidelity simulations at the continuous moment-to-moment level as well as their relations to ratings of performance in high-fidelity simulations, interpersonal adaptability and task performance in interpersonal settings.

Dissertation Outline

The current dissertation consists of four chapters that each aim to address one of the objectives mentioned above and one overall discussion chapter. Each of the chapters was written to be read individually. Therefore, overlaps in terms of content are both possible and intended.

Chapter 2, entitled “Situational Judgment Tests as Measures of 21st Century Skills: Evidence across Europe and Latin America“, addresses Objective 1. It provides a theoretical

introduction about possible advantages of SJTs to measure procedural knowledge of various skills that are crucial for success in the world of work of the 21st Century, including skills that tap into interpersonal performance and adaptability. Further, this chapter outlines how a combined emic-etic approach can be used for developing SJTs that can be applied across geographical regions and investigates whether SJT scores can indeed be compared across regions. This chapter has already been published in Journal of Work & Organizational Psychology.

Chapter 3, entitled "Multiple Speed Assessments: Theory, Practice, and Research Evidence", addresses Objective 2. It outlines the shared characteristics and theoretical fundaments of Multiple Speed Assessments. Further it showcases various examples of Multiple Speed Assessments and possible design variations. Finally, current research evidence and future research directions related to Multiple Speed Assessments are presented. This chapter has been accepted for publication and is currently an Advance Online Article in European Journal of Psychological Assessment.

Chapter 4, entitled "Multiple Speed Assessments Under Scrutiny: Are Their Ratings Reliable and Valid?", addresses Objective 3. It draws from the zero acquaintance/thin slices paradigm to derive hypotheses about the reliability and validity of judgments in short, structured interpersonal simulations. Regarding reliability, this chapter investigates the interrater reliability of judgments in short, structured interpersonal simulations, and it examines the relative importance of various sources of reliable and unreliable sources of variance in these ratings. Further, it is investigated how many simulations and independent assessors are necessary to gain an overall reliable estimate of performance. Regarding validity, this chapter investigates relations between judgments in short and fast interpersonal simulations on the one hand and the cognitive ability and personality of participants on the other hand. On top of that, this chapter investigates the predictive and incremental validity of

Multiple Speed Assessments beyond more traditional predictors that tap into similar construct domains. This chapter has been accepted as full paper and presented at the 79th Annual Meeting of the Academy of Management. Further, this paper was judged by anonymous reviewers to be one of the best accepted papers in the conference program. It is now in further preparation for submission to an A1-journal.

Chapter 5, entitled "A closer look at Intraindividual Variability in Interpersonal Behavior and Interpersonal Dynamics in High-Fidelity Simulations" addresses Objective 4. It draws from Interpersonal Theory and the Interpersonal Circumplex Model as theoretical fundament for interpersonal dynamics in high-fidelity simulations, such as Multiple Speed Assessments. Further, this chapter investigates intraindividual variability in interpersonal behavior as well as interpersonal dynamics between participants and role-players in four distinct high-fidelity simulations. To do so, an assessment approach is utilized that assesses interpersonal behavior and interpersonal dynamics at the continuous, moment-to-moment level. Further, this chapter explores relations between interpersonal dynamics to performance in high-fidelity simulations, interpersonal adaptability and task-related performance. Portions of this paper were presented at the 33rd Annual Conference of the Society of Industrial and Organizational Psychology. This chapter is now in further preparation for submission to an A1-journal.

Chapter 6 ends with a general discussion of all previous chapters. It thus summarizes all main results and conclusions from each of the previous chapters, integrates them, and proposes further avenues for future research.

References

Becker, N., Höft, S., Holzenkamp, M., & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology, 10,* 61–69. https://doi.org/10.1027/1866-5888/a000031

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

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

Cascio, W. F. (2003). Changes in workers, work, and organizations. In R. J. Klimoski, W. C. Borman, & D. R. Ilgen (Eds.), *Handbook of Psychology* (Vol. 12, pp. 401–422). Hoboken, NJ: Wiley & Sons.

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

Cheung, F. M., Fan, W., Cheung, S. F., & Leung, K. (2008). Standardization of the cross-cultural Chinese Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment. *Acta Psychologica Sinica, 40,* 839–852. http://dx.doi.org/10.3724/SP.J.1041.2008.01639

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

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

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.

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*, 327–347. https://doi.org/10.5465/amj.2007.24634438

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405–411. https://doi.org/10.1111/j.1468-2389.2007.00399.x

Hesketh, B., & Neal, A. (1999). Technology and performance. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 21–55). San Francisco, CA: Jossey-Bass.

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

Javidan, M., Dorfman, P. W., de Luque, M. S., & House, R. J. (2006). In the eye of the beholder: *Academy of Management Perspectives, 20*, 67–90. https://doi.org/10.5465/AMP.2006.19873410

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view. *European Journal of Psychological Assessment*, *22*, 168–176. https://doi.org/10.1027/1015-5759.22.3.168

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

Kozlowski, S. W. J., Gully, S. M., Salas, E., & Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In M. Beyerlein, S. Beyerlein, & D. Johnson (Eds.), *Advances in interdisciplinary studies of work teams: Team leadership* (Vol. 3, pp. 251–289). Greenwich, CT: JAI Press.

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 279–300). Mahwah, NJ: Erlbaum.

Lievens, F., Corstjens, J., Sorrel, M. Á., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? International Journal of Selection and Assessment, 23, 361-372. https://doi.org/10.1111/ijsa.12120

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., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *Handbook of Assessment and Selection* (pp. 383–410). Oxford: University Press.

Lievens, F, & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245–286). Chicester: John Wiley & Sons, Ltd.

Lievens, F, & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3–22. https://doi.org/10.1017/iop.2015.71

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.

Markey, P., Lowmaster, S., & Eichler, W. (2010). A real-time assessment of interpersonal complementarity. *Personal Relationships, 17*, 13–25. https://doi.org/10.1111/j.1475-6811.2010.01249.x

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

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. http://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281–288. https://doi.org/10.1007/s10869-009-9106-4

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

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment, 34*, 328–335. https://doi.org/10.1027/1015-5759/a000346

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. https://doi.org/10.1177/0149206314525207

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., de Vries, R. E., & de Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance, 32*, 1–29. https://doi.org/10.1080/08959285.2018.1539856

Pincus, A. L., Sadler, P., Woody, E., Roche, M. J., Thomas, K. M., & Wright, A. G. C. (2014). Multimethod assessment of interpersonal dynamics. In C. J. Hopwood & R. F. Bornstein (Eds.), *Multimethod clinical assessment* (pp. 51–91). New York: Guilford.

Pinsight. (2018). Virtual assessment centers. Retrieved November 27, 2018, from https://www.pinsight.com/

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests: Theory, Measurement, and Application* (pp. 345–350). Mahwah, NJ: Erlbaum.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624. https://doi.org/10.1037//0021-9010.85.4.612

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology, 97*, 1005–1020. https://doi.org/10.1037/a0016232

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

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153–193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x

Schmitt, N., & Chan, D. (2006). Situational judgement tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application* (pp. 135–155). Mahwah, NJ: Lawrence Erlbaum.

Such, M. J., & Schmidt, D. B. (2004). *Examining the effectiveness of empirical keying: A cross-cultural perspective.* Presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, IL.

Thornton III, G. C., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Tracey, T. J. G. (2004). Levels of interpersonal complementarity: A simplex representation. *Personality and Social Psychology Bulletin*, *30*, 1211–1225. https://doi.org/10.1177/0146167204264075

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372-376. http://dx.doi.org/10.1037/h0026244

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188–202. https://doi.org/10.1016/j.hrmr.2009.03.007

Zeithaml, V. A., & Bitner, M. J. (1996). *Services marketing*. New York: McGraw-Hill.

CHAPTER 2: SITUATIONAL JUDGMENT TESTS AS MEASURES OF 21st CENTURY SKILLS: EVIDENCE ACROSS EUROPE AND LATIN AMERICA[1]

Over the years, various governmental, employment, and academic organizations have identified a list of skills to successfully master the challenges of the 21st century. So far, an adequate assessment of these skills across countries has remained challenging. Limitations inherent in the use of self-reports (e.g., lack of self-insight, socially desirable responding, response style bias, reference group bias, etc.) have spurred on the search for methods that could complement or even substitute self-report inventories. Situational judgment tests (SJTs) have been proposed as one of the complements/alternatives to the traditional self-report inventories. SJTs are low-fidelity simulations that confront participants with multiple domain relevant situations and request to choose from a set of predefined responses. Our objectives are twofold: (a) outlining how a combined emic-etic approach can be used for developing SJT items that can be used across geographical regions and (b) investigating whether SJT scores can be compared across regions. Our data come from Laureate International Universities (N = 5,790) and comprise test-takers from Europe and Latin America who completed five different SJTs that were developed in line with a combined emic-etic approach. Results showed evidence for metric measurement invariance across participants from Europe and Latin America for all five SJTs. Implications for the use of SJTs as measures of 21st Century Skills are discussed.

[1] This chapter is based on: Herde, C. N., Lievens, F., Solberg, E. G., Harbaugh, J. L., Strong, M. H., & Burkholder, G. J. (2019). Situational judgment tests as measures of 21st Century Skills: Evidence across Europe and Latin America. *Journal of Work and Organizational Psychology, 35*, 65-74. https://doi.org/10.5093/jwop2019a8

Introduction

Since several decades, various educational and (non)profit organizations around the globe have compiled lists of skills needed for the next generation to survive in an ever changing, turbulent, and complex world. Although the final lists of these large-scale international efforts often differ in their name ("survival skills", "21st Century Dkills") and content, they all share the characteristic that the skills identified go beyond technical and functional aptitude. The most common examples of such 21st Century Skills are, therefore, collaboration and teamwork, creativity and imagination, critical thinking, and problem solving (see, for overviews, Binkley et al., 2012; Geisinger, 2016).

Besides identifying the list of 21st Century Skills, an equally important issue deals with how these skills are best measured. Specifically, challenges deal with using a methodology that does not lead to biases and that enables comparing the results obtained across the various geographical regions. Along these lines, it is of pivotal importance that measurement effects do not cloud the standing of the regions on the 21st Century Skills (constructs). In the past, self-reports were typically used for determining people's standing on each of the skills. However, the self-report methodology suffers from various pitfalls. One drawback is that self-reports assume people possess the necessary self-insight to rate themselves on each of the statements that operationalize the 21st Century Skills. Another drawback is that people tend to engage in response distortion in that they might overstate how they score on the statements (socially desirable responding). Other documented limitations relate to response style bias (extreme responding that differs across groups, such as different cultures; e.g., Hui & Triandis, 1989; Johnson, Kulesa, Cho, & Shavitt, 2005) or reference group bias (responding that is dependent on the chosen group of reference, such as one's own cultural group; e.g., Heine, Lehman, Peng, & Greenholtz, 2002).

These limitations have resulted in the search for other methods for measuring 21st Century Skills (Kyllonen, 2012; see also Ainley, Fraillon, Schulz, & Gebhardt, 2016; Care, Scoular, & Griffin, 2016; Ercikan & Oliveri, 2016; Greiff & Kyllonen, 2016; Herde, Wüstenberg, & Greiff, 2016; Lucas, 2016). In PISA (OECD, 2014), three such approaches were suggested (for a summary, see Kyllonen, 2012). The first method dealt with the use of anchoring vignette items. Anchoring vignette items first ask respondents to evaluate several other targets on a specific target construct. Only afterwards, a respondent provides a self-rating on the target construct. The respondent's self-rating is then rescaled based upon the evaluation standards that are extracted from the other ratings (e.g., Hopkins & King, 2010). As a second approach, forced-choice methods were proposed. Forced-choice items do not ask respondents to evaluate isolated statements about themselves on a Likert-scale. Instead, they confront respondents with a choice between options that are intended to be of similar social desirability. Recent research attested to the broad applicability of forced-choice items (Brown & Maydeu-Olivares, 2011; Stark, Chernyshenko, & Drasgow, 2004). Third, situational judgment tests (SJTs) were proposed. SJTs confront respondents with multiple, domain-relevant situations and request to choose from a set of predefined responses (Motowidlo, Dunnette, & Carter, 1990).

Importantly, these approaches aim to alleviate the limitations inherent in the typical self-report inventories, while at the same time ensuring that the average ratings on the 21st Century Skills can be compared across geographical regions. Note that SJTs do not actually measure 21st Century Skills. Instead, SJTs assess people's procedural knowledge ("knowing what to do and how to do it") of engaging in behavior that operationalizes a given 21st Century Skill (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006).

In this study, we focus on the use of SJTs as measures of 21st Century Skills. Our objectives are twofold. First, we outline how a combined emic-etic approach can be used for developing SJT items that can be used across geographical regions. Second, we investigate whether SJT scores derived from a SJT that was developed in line with a combined emic-etic approach can indeed be compared across regions. We do so by conducting analyses of measurement invariance across regions of Europe and Latin America. Analyses of measurement invariance reveal whether different (regional or cultural) groups interpret test items in the same way and attribute the same meaning to them. Therefore, analyses of measurement invariance are crucial to disentangle measurement effects from true score differences between (regional or cultural) groups (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Our study is situated in an educational context. We use the data from Laureate International Universities, which is a global network of universities that, at the time of the study, operated in 25 countries and had over one million students globally. Similar to the efforts described above, Laureate International Universities started in 2015 to identify, define, and measure foundational competencies and behavioral skills required by graduating students to be successful in entry-level professional jobs across industries and geographical regions. SJT items were also developed to assess those foundational competencies. On the basis of the SJT scores, students receive feedback regarding their strengths and weaknesses as well as actionable tips to help them improve. It is also important that regions can be compared on their average standing on the various competencies.

The structure of this paper is as follows: First, we shortly define SJTs and illustrate their most important characteristics. Second, we explain why these special characteristics of SJTs may pose problems for measurements across geographical regions. Third, we describe how a combined emic-etic approach of test development might serve to limit these problems.

Fourth, we provide an empirical test of the combined emic-etic approach to develop SJTs to measure 21st Century Skills across geographical regions of Europe and Latin America. Fifth, we discuss our results and implications for further research and practice.

Study Background

SJTs: Definition, Characteristics, and Brief History

In SJTs, candidates are presented with short domain-relevant situational descriptions and various response options to deal with the situations. Upon reading the short situational descriptions, candidates are asked to pick one response option from a list, rank the response options ("What would you prefer doing most, least?"), or rate the effectiveness of these options (Motowidlo et al., 1990). Most SJTs still take the form of a written test because the scenarios are presented in a written format. In video-based or multimedia SJTs, a number of video scenarios describing a person handling a critical situation is developed (McHenry & Schmitt, 1994). Recently, organizations are also exploring 2D-animated, 3D-animated, and even avatar-based SJTs (see, for an overview, Weekley, Hawkes, Guenole, & Ployhart, 2015).

SJTs are not new inventions. Early SJT versions go back to before WWII. In 1990, Motowidlo and colleagues reinvigorated interest in SJTs. Since then, SJTs have become attractive selection instruments for practitioners who are looking for cost-effective instruments. As compared to other sample-based predictors, SJTs might be easily deployed via the internet in a global context due to their efficient administration (Ployhart, Weekley, Holtz, & Kemp, 2003). Moreover, in domestic employment contexts, SJTs have demonstrated adequate criterion-related and incremental validity and potential to reduce adverse impact (Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb III, 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001).

SJTs in an International Context: Potential Problems

Although SJTs have been advanced as alternative method for assessing 21st Century Skills across geographical regions, such an outcome is far from assured. For example, Ployhart and Weekley (2006) mentioned the following key challenge: "it is incumbent on researchers to identify the cross-cultural generalizability – and limits – of SJTs… One might ask whether it is possible to create a SJT that generalizes across cultures. Given the highly contextual nature of SJTs, that poses a very interesting question." (p. 349). Indeed, SJT items are directly developed or sampled from the criterion behaviors that the test is designed to predict (Chan & Schmitt, 2002). Therefore, SJT items are highly contextualized because the situations are embedded in a particular context or situation that is representative of future tasks.

Lievens (2006) reviewed prior research on SJTs in a cross-cultural context and identified SJT item characteristics that might affect the cross-cultural use of SJTs (see Lievens, Corstjens et al., 2015). The contextualized nature of SJT items makes them particularly prone to cultural differences because the culture wherein one lives acts like a lens, guiding the interpretation of events and defining appropriate behaviors (Heine & Buchtel, 2009; Lytle, Brett, Barsness, Tinsley, & Janssens, 1995). This contextualized nature of SJTs might create boundary conditions for the use across geographical regions in at least four ways (Lievens, 2006). First, the contextualization in SJTs is shown in the kind of problem situations (i.e., the item stems) that are presented to candidates. When SJTs are used in an international context, the issue then becomes whether there are differences in terms of the situations (critical incidents) generated across regions. Some situations will simply not be relevant in one region, whereas they might be very relevant in another region (e.g., differences in organizing meetings across countries). Second, similar differences might occur on the level of how to react to the problem situation. That is, some response options might be

relevant in one region, whereas they might not occur in another region. The meeting example can again be used here, with openly not agreeing with the boss being an unrealistic response option in some regions. Third, the effectiveness (scoring) of response options might vary across regions. Along these lines, Nishii, Ployhart, Sacco, Wiechmann, and Rogg (2001) stated: "if a scoring key for a SJT is developed in one country and is based on certain cultural assumptions of appropriate or desirable behavior, then people from countries with different cultural assumptions may score lower on these tests. Yet these lower scores would not be indicative of what is considered appropriate or desirable response behavior in those countries". Fourth, the item-construct linkages might differ across regions. That is, a specific response option might be an indicator of a given construct in one region but an indicator of another construct in another region. For example, to decline a task assignment from a supervisor because of time constraints during a department meeting might indicate assertiveness or self-regulation in a culture low in power distance but might indicate impoliteness or rudeness in a culture high in power distance.

In short, these potential differences in the situations, response options, response option effectiveness, and item-construct linkages across geographical regions highlight that care should be taken to develop SJTs for measuring 21st Century Skills across regions. That is, strategies should be deployed for designing SJTs that alleviate these potential problems.

Strategies for SJT Design in a Cross-cultural Context: Emic, Etic, and Combined Emic-Etic Approach

In the search of strategies for dealing with potential threats to the cross-cultural transportability of SJTs, it is possible to borrow valuable insights from the large body of research in cross-cultural psychology. Generally, three possible approaches can be adopted for developing global (selection) instruments, namely an emic, an imposed etic, and a combined

emic-etic approach (Berry, 1969, 1990; Headland, Pike, & Harris, 1990; Leong, Leung, & Cheung, 2010; Morris, Leung, Ames, & Lickel, 1999; Pike, 1967; Yang, 2000).

An indigenous or emic approach posits that tests should be developed and validated with the own culture as a point-of-reference. In the context of SJTs, an example is the study of Chan and Schmitt (2002). They developed an SJT for civil service positions in Singapore. This implied that the job analysis, the collection of situations, the derivation of response alternatives, the development of the scoring key, and the validation took place in Singapore. Chan and Schmitt (2002) found that in Singapore the SJT was a valid predictor for overall performance and had incremental validity over cognitive ability, personality, and job experience. This corresponds to the meta-analytic validity research base in the United States (Christian et al., 2010; McDaniel et al., 2007; McDaniel et al., 2001).

In this example, the development of the SJT ensured that the job relevant scenarios were derived from input of local subject matter experts. However, there are also drawbacks in the emic approach. As an indigenous approach implicates the use of different instruments for different countries, it is a costly and time-consuming strategy. In addition, a challenge for the country-specific emic approach is to contribute to the cumulative knowledge in a specific domain that typically centers around generalizable concepts (Leong et al., 2010; Morris et al., 1999).

Contrary to the emic approach, the imposed etic approach assumes that the same instrument can be applied universally across different cultures (Berry, 1969; Church & Lonner, 1998). So, according to the imposed etic approach, a selection procedure developed in a given country can be exported for use in other countries when guidelines for test translation and adaptation are taken into consideration (International Test Commission, 2001). Hence, the imposed etic approach represents an efficient strategy for cross-cultural assessment. However, the imposed etic approach is also not without limitations. Even when

tests are appropriately translated and adapted, the test content of the transported instruments might reflect predominantly the culture from which the instrument is derived, thereby potentially omitting important emic aspects of the local culture (Cheung et al., 1996; Leong et al., 2010).

In light of these drawbacks, the effectiveness of the imposed etic approach for constructing international SJTs seems doubtful given the highly contextualized nature of SJT items. One study confirmed the problems inherent in using an imposed etic approach in contextualized instruments such as SJTs. Such and Schmidt (2004) validated an SJT in three countries. Results in a cross-validation sample showed that the SJT was valid in half of the countries, namely the United Kingdom and Australia. Conversely, it was not predictive in Mexico. These results suggest that effective behavior on the SJT was mainly determined in terms of what is considered effective behavior in two countries with a similar heritage (the United Kingdom and Australia).

Another study on the cross-cultural transportability of SJTs showed that an integrity SJT developed in the US was generally applicable to a Spanish population as well (Lievens, Corstjens et al., 2015). Most of the scenarios from the American SJT were rated to be realistic in the Spanish population, patterns of endorsements of various response options were mainly similar across cultures, the American scoring scheme correlated highly with Spanish scoring schemes and item construct linkages also appeared to be comparable, because correlations between self-reports and SJT scores were found to be similar across cultures. In sum, evidence for the imposed etic approach for constructing international SJTs is mixed.

Yet, the emic-etic distinction should not be seen as a dichotomy. Rather, it constitutes a continuum (Church, 2001; Morris et al., 1999; Sahoo, 1993). Therefore, it is possible to combine these cultural-general and cultural-specific approaches of international test development (Leong et al., 2010; Schmit, Kihm, & Robie, 2000), resulting in the combined

emic-etic approach. In such a combined emic and etic approach, the instrument is developed with cross-cultural input. In the personality domain, we are aware of two prior projects that successfully applied the combined emic-etic approach. First, in the development of the Chinese Personality Assessment Inventory (CPAI; Cheung, Cheung et al., 2008; Cheung, Fan, Cheung, & Leung, 2008; Cheung et al., 1996) descriptions of personality were extracted from multiple sources (e.g., proverbs, everyday life, etc.) to identify personality constructs relevant to the Chinese culture. These local expressions were then compared to translations of imported measures of similar constructs. Large-scale tests of the inventory in China showed that there was substantial overlap between the CPAI and the Big Five, although there were also unique features (i.e., the interpersonal relatedness factor). As a second illustration, Schmit et al. (2000) developed a global personality inventory. Hereby the behavioral indicators (items) of personality constructs that were written by worldwide panels of local experts varied, while the broader underlying constructs were similar across countries. Construct-related validity studies provided support for the same underlying structure of the global personality inventory across countries.

So, as a result of a combined emic-etic approach, both universal and indigenous constructs are incorporated: the inclusion of culture-specific concepts produces within-culture relevance, while the measurement of universal concepts allows cross-cultural comparisons (Cheung et al., 1996). The combined emic-etic approach also enables to expand the interpretation of indigenous constructs in a broader cultural context.

In sum, prior studies have developed and used SJTs in various geographical regions. However, many applications were within-country examinations that attest to an indigenous (culture specific/emic) approach. One study (Such & Schmidt, 2004) applied an imposed etic approach with the SJT not being valid in some countries. To avoid these problems, the combined emic-etic approach might serve as a potentially viable strategy for constructing

sample-based selection procedures such as SJTs for use in cross-cultural applications. So far, no empirical studies have used or tested this combined emic-etic approach in sample-based selection procedures such as SJTs. This study starts to fill this key research and practice gap by using a combined emic-etic approach for constructing an SJT for assessing 21st Century Skills across geographical regions.

Method

Development and Validation of a Global Competency Framework

Laureate International Universities developed and validated a comprehensive framework of competencies that are required by graduating students to be successful in the workplace across geographical regions, industries, and jobs. In line with the combined emic-etic approach, cross-regional input was gained across all developmental steps to ensure that the competency framework was relevant across regions and cultures.

The development of the competency framework was based on various sources of information. These included best practices in competency modeling (Campion et al., 2011; Kurz & Bartram, 2002), content of competency frameworks from academic institutions and professional companies (e.g., Getha-Taylor, Hummert, Nalbandian, & Silvia, 2013; Lee, 2009; Lunev, Petrova, & Zaripova, 2013), internal research conducted by several institutions in the Laureate network, and data from various research partners. A draft competency framework was developed by integrating information from these sources and utilizing competency names and definitions from the SHL Universal Competency Framework (Bartram, 2012).

To ensure that the draft competency framework comprehensively covered competencies that were applicable and important across geographical regions, industries, and jobs, it was reviewed, refined, and approved by various groups. These groups included a global advisory council, consisting of eighteen members from regions represented in

Laureate, two subject matter experts on competency modeling, and eighteen global focus groups that represented all regions, stakeholders (students/alumni, faculty/staff, academic leaders, and employers), and experts across disciplines. In total, the global focus group comprised of 86 participants.

Finally, two survey studies were conducted among Laureate's stakeholders across the network to evaluate and refine the competency framework. In Survey 1, 25,202 representatives across different stakeholders, roles, disciplines, and regions confirmed the importance of the competencies for entry-level professionals. In Survey 2, 10,420 of these representatives further reviewed and confirmed the individualist behaviors defined within each competency. The final competency framework includes 20 competencies. Further details about the competency framework, its development, and the global validation study are reported elsewhere (Strong, Burkholder, Solberg, Stellmack, & Presson, manuscript submitted for publication).

In this study, we focus on five core competencies that were identified in the global validation study as most important and critical for successful job performance of new professionals across geographical regions, industries, and jobs. These core competencies are achieving objectives, adapting to change, analyzing and solving problems, learning and self-development, and working well with others. The definitions of these competencies are provided in the Appendix Table A1.

SJT Item Design and Scoring

Analogous to the development of the competency framework, a combined emic-etic approach was applied to develop written SJT items with close-ended response format for the competencies. The development of the SJTs followed recommendations from Weekley, Ployhart, and Holtz (2006). We started with using the critical incident technique (Flanagan, 1954) to gain input for item development from subject matter experts. Given that the SJTs

should assess competencies required of graduating students to be successful in the workplace, students, faculty/staff, administrators, alumni, and advisory committee members of Laureate institutions as well as employers served as subject matter experts. Representatives from these groups were invited to fill in an online survey to describe specific situations for a chosen competency, in which one student performed exceptionally well and another student performed exceptionally poorly. In total, 1,749 critical incidents were gathered from 564 respondents.

Three experienced test construction consultants drafted initial items. They compiled, reviewed and synthesized the critical incidents. Per competency, critical incidents and related examples for excellent and poor performance were converted into item stems and response options. Per item stem, five response options were generated that aimed to measure different levels of proficiency for the same competency.

Item stems and response options were written in a way to be applicable across different regions, industries, and jobs. To verify this, two global focus group panels reviewed all items and determined the scoring key. The panels consisted of 21 and 22 participants, respectively. Both panels represented similar numbers of representatives from all geographical regions, functional roles (Laureate faculty/staff and employers), and employers from different industries. Panelists reviewed items with special focus on realism and face validity of depicted situations and response options within their geographical region and field of work. Potential issues were discussed and items were adapted, if necessary.

To set the scoring key per SJT, these panelists rated the effectiveness of each response option per item stem on a five-point scale (1 = very ineffective, 5 = very effective). In line with the consensus weighting method (see Chan & Schmitt, 1997), the average ratings were used to assign each response option a score of 1 through 5 points.

The items and related response options and scoring keys were further reviewed by assessment experts and employers. In total, twelve assessment experts (two per geographical region) with advanced degrees in Industrial/Organizational Psychology or a closely related discipline reviewed all items. Assessment experts provided feedback regarding item clarity or content from their own cultural perspective. Based upon this feedback, some items were slightly modified. Assessment experts also indicated whether each item appeared to tap into the respective competency. If at least half of the assessment experts indicated that an item did not appear to capture the targeted competency, the respective item was dropped. A final panel of fourteen employers reviewed all items. Again, this panel was formed by representatives from all global regions as well as from different industries and jobs.

After final minor item modifications, each of the competency specific SJTs constructed consisted of 21 items on average. Items had a behavioral tendency response instruction ("What would you do?"). For each item stem/scenario, participants were instructed to choose a response option they would most likely do and another response option they would least likely do. Participants could receive between 1 and 5 points for each choice. Therefore, scores could vary between 2 and 10 points per scenario.

All SJT items were translated from English into six additional languages. These additional languages were Latin American Spanish, European Spanish, Brazilian Portuguese, European Portuguese, French, and German. The rigorous translation process followed guidelines for translating tests (e.g., Van de Vijver, 2003), including repeated front and back translations by different translators.

Procedure and Sample

Laureate institutions invited their students to take part in this study to receive developmental feedback about their competency levels. The different SJTs were distributed across four different bundles that contained different competency specific SJTs. Students

were invited to complete one bundle but could complete additional bundles to receive developmental feedback about further competencies. Within each bundle, students completed a random set of eight scenarios per competency specific SJT. Finally, students responded to demographic questions.

To assure that only valid data were analyzed, we removed data for several reasons. In a limited number of 24 cases, students started the same bundle twice. To exclude biases due to retest effects regarding the same competencies or scenarios, we excluded responses from the second bundle completion. For the same reason, we removed responses of eight students from the second access to any SJT of the same competency. Given that we were interested in cross-regional comparisons, we took care that participants understood the test items well. Hence, we removed data for 87 students that indicated to be "not comfortable" with the language in which they completed the SJTs. Further, we removed students' responses per scenario if they were made in less than twelve seconds (internal test runs had shown it was impossible to choose both a best and worst response per scenario in less than twelve seconds). Remaining sample sizes for our five core competencies did not justify analyses for the geographical regions of Africa, Asia, Oceania, or the US. Therefore, we focused our analyses on students from Europe and Latin America.

After data cleaning, a total of 5,790 students (53% female) from twenty different institutions provided valid responses to the competency specific SJTs (mean age = 22.63, *SD* = 5.09); 64% of the students resided in Europe, 36% in Latin America. In total, students came from eighteen different countries. The majority of European students resided in Turkey (30%), Portugal (20%), or Spain (17%). The majority of Latin American students lived in Mexico (34%), Chile (22%), or Brazil (18%). Each student chose to complete the SJTs in one of seven available languages. The majority of students completed the SJTs in English (32%), Latin American Spanish (29%), or European Portuguese (13%); 74 % of all students

completed the SJTs in their dominant language; 72% of all students reported to be "very comfortable" with the language in which they completed the SJTs[2]. Students completed the SJTs either during their first (52%) or last year of study (48%) at the institution; 45% completed the SJTs in a proctored setting; 58% of students reported to have already gained some professional experience; 41% already completed an internship; 16% of all participants were graduate students. Students studied across thirteen different majors (31 % Business & Management, 15 % Engineering and Information Technology, 14% Health Sciences).

Results

Internal Consistency Reliabilities

We based our analyses on SJT scenario scores as sum scores for the best and worst choice per scenario. To calculate internal consistencies for each of the five SJTs, we used the full information maximum likelihood procedure and the ML estimator in Mplus Version 7.4 (Muthén & Muthén, 1998-2015) to estimate scenario scores from missing values. Then, we used intercorrelations between scenario scores to calculate Cronbach's alpha for our total sample. Internal consistencies of the five SJTs were moderate to acceptable for the total sample (.67-.78, see Table 1). Internal consistencies calculated separately for each region produced similar results (see Table 1).

[2] We re-ran our analyses once with only students included who did the SJTs in their dominant language and once only with students included who reported to be "very comfortable" with the test language. Given that results were similar and did not change conclusions, we report results for our complete sample only.

Table 1

Internal consistencies, means and standard deviations per geographical region by SJTs

	n	α	*M*	*SD*
Achieving Objectives (19 items)				
Europe and Latin America	3,666	.78	7.56	1.27
Europe	2,666	.79	7.57	1.23
Latin America	1,000	.78	7.53	1.38
Adapting to Change (20 items)				
Europe and Latin America	4,511	.69	7.58	1.17
Europe	3,586	.69	7.61	1.14
Latin America	925	.69	7.48	1.26
Analyzing & Solving Problems (19 items)				
Europe and Latin America	4,360	.67	7.55	1.11
Europe	3,100	.69	7.58	1.08
Latin America	1,260	.63	7.47	1.17
Learning & Self-Development (23 items)				
Europe and Latin America	3,892	.73	7.66	1.21
Europe	2,731	.73	7.65	1.17
Latin America	1,161	.75	7.68	1.30
Working Well with Others (20 items)				
Europe and Latin America	4,185	.76	7.85	1.15
Europe	3,200	.77	7.85	1.12
Latin America	985	.73	7.82	1.23

Measurement Invariance across Regions

To examine measurement invariance across regions for each of the five SJTs, we first sought to establish a baseline model for the total sample, then investigated model fit for the baseline model within each region, and afterwards ran increasingly restrictive multi-group confirmatory factor analyses (e.g., Byrne & Stewart, 2006; Byrne & Van de Vijver, 2010). We conducted these analyses in Mplus via the full information maximum likelihood procedure and the ML estimator.

To guide the examination of a baseline model for the total sample, we hypothesized that a one-factor model would explain scenario scores for each SJT. This hypothesis was based upon the fact that all scenarios and response options for a specific SJT were developed to tap into one respective competency. For all five SJTs, a one-factor model showed good model fit (see Table 2). Thus, a one-factor model was chosen as baseline model in all of the following steps.

Table 2

Goodness-of-fit indices for factor structure models (overall sample and within regions)

	n	χ^2 (*df*)	χ^2/df	CFI	RMSEA (90 % CI)	SRMR
		Achieving Objectives				
Europe and Latin America	3,666	293.63 (152)**	1.93	.908	.016 (.013-.019)	.047
Europe	2,666	294.67 (152)**	1.94	.885	.019 (.016-.022)	.055
Latin America	1,000	199.15 (152)**	1.31	.872	.018 (.010-.024)	.081
		Adapting to Change				
Europe and Latin America	4,511	254.00 (170)**	1.49	.921	.010 (.008-.013)	.041
Europe	3,586	264.45 (170)**	1.56	.896	.012 (.009-.015)	.046
Latin America	925	229.27 (170)**	1.35	.756	.019 (.012-.026)	.093
		Analyzing and Solving Problems				
Europe and Latin America	4,360	240.67 (152)**	1.58	.907	.012 (.009-.014)	.042
Europe	3,100	211.46 (152)**	1.39	.923	.011 (.007-.015)	.045
Latin America	1,260	175.53 (152)	1.15	.875	.011 (.000-.018)	.071
		Learning & Self-Development				
Europe and Latin America	3,892	305.26 (230)**	1.33	.908	.009 (.006-.012)	.051
Europe	2,731	288.66 (230)**	1.26	.901	.010 (.006-.013)	.058
Latin America	1,161	315.02 (230)**	1.37	.706	.018 (.013-.023)	.100
		Working Well with Others				
Europe and Latin America	4,185	240.54 (170)**	1.41	.949	.010 (.007-.013)	.040
Europe	3,200	209.81 (170)*	1.23	.964	.009 (.004-.012)	.043
Latin America	985	273.76 (170)**	1.61	.726	.025 (.019-.030)	.092

Note. * $p < .05$, ** $p < .01$

We then investigated model fit for this baseline model per region. For the SJTs of "achieving objectives" as well as "analyzing and solving problems", model fit for the baseline model within each region were at least acceptable. For the three remaining SJTs, the CFI value for the model fit within Latin America fell below the limit of acceptable model fit. Previous studies that investigated the factor structure of SJTs frequently found similar patterns and usually failed to find good model fit (with acceptable CFI values). To analyze measurement invariance, these studies then used the best fitting model as baseline model for the multi-group confirmatory factor analyses (e.g., Krumm et al., 2015; Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2018). In line with this approach, we kept the one-factor model as baseline model for our measurement invariance analyses.

To investigate measurement invariance, we sought to find evidence for configural and metric invariance for the baseline model across regions (see summary of Byrne & Van de Vijver, 2010). To investigate configural measurement invariance, we restricted the number of latent factors and the number of factor loadings to be equal across both regional groups. Configural measurement invariance therefore indicates that the same factorial structure explains the observed scores across regional groups. Second, we restricted the size of factor loadings to be equal across both regional groups to investigate metric measurement invariance. Metric measurement invariance thus suggests that observed scores are equally related to the assumed latent factor(s). In other words, metric measurement invariance indicates that the observed scores measure the latent factor(s) equally across (regional) groups (see, for example, Byrne & Stewart, 2006; Byrne & van de Vijver, 2010).

To examine configural and metric measurement invariance, we inspected model fit, and conducted nested model comparisons by using the chi-square difference test as well as the criterion proposed by Cheung and Rensvold (2002). These authors stated that measurement equivalence could be defended in practical terms, if increasingly restrictive confirmatory

factor analyses are associated with only marginal drops in CFI values (ΔCFI < .01; see also Byrne & Stewart, 2006). With exception of the SJT for "achieving objectives", chi-square difference tests were not significant for all five SJTs, which provides evidence for metric measurement invariance. In addition, drops in CFI values were marginal for all five SJTs (ΔCFI $\leq$ .008). Thus, we concluded that metric measurement equivalence could be established for all five SJTs (see Table 3). Importantly, this means that at a practical level differences in manifest mean scenario scores across regions can be compared.

Table 3

Tests of Measurement Invariance for One-Factor Model Underlying SJT Scores Across Participants from Europe and Latin America

Model	χ^2 (*df*)	χ^2/df	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA (90 % CI)	SRMR
			Achieving Objectives					
Equal number of factors	493.81 (304)**	1.62			.882		.018 (.015-.021)	.063
Equal factor loadings	523.03 (322)**	1.62	29.21*	18	.875	.007	.018 (.016-.021)	.068
			Adapting to Change					
Equal number of factors	493.72 (340)**	1.45			.866		.014 (.011-.017)	.059
Equal factor loadings	509.51 (359)**	1.42	15.79	19	.869	.003	.014 (.011-.016)	.061
			Analyzing and Solving Problems					
Equal number of factors	386.99 (304)**	1.27			.913		.011 (.007-.014)	.054
Equal factor loadings	409.39 (322)**	1.27	22.40	18	.909	.004	.011 (.008-.014)	.056
			Learning and Self-Development					
Equal number of factors	603.68 (460)**	1.31			.837		.013 (.010-.015)	.073
Equal factor loadings	632.84 (482)**	1.31	29.16	22	.829	.008	.013 (.010-.015)	.077
			Working Well with Others					
Equal number of factors	483.56 (340)**	1.42			.903		.014 (.011-.017)	.058
Equal factor loadings	507.38 (359)**	1.41	23.82	19	.900	.003	.014 (.011-.017)	.062

Note. * $p < .05$, ** $p < .01$

Discussion

Many educational and (non)profit organizations have investigated which skills or competencies are needed to face the challenges of the 21st Century (Binkley et al., 2012; Geisinger, 2016). Subsequently, researchers have started to investigate how such 21st Century Skills can be best measured (Kyllonen, 2012). One such key challenge deals with assessing 21st Century Skills without biases that may interfere with comparing results obtained across various geographical regions and cultures. This study advances our knowledge about appropriate assessment approaches for 21st Century Skills by outlining how the combined emic-etic approach enables developing SJTs that tap into 21st Century Skills across regional groups. To this end, we investigated measurement invariance across Europe and Latin America for five different SJTs that assessed a core competency for graduating students to be successful in entry-level jobs.

Our results showed that configural and metric measurement invariance could be established across Europe and Latin America for all five SJTs. Thus, the same factorial structure explained SJT scenario scores across these regional groups and SJT scenario scores measured the latent factor(s) equally across those regional groups (see, for example, Byrne & Stewart, 2006; Byrne & van de Vijver, 2010). In other words, participants from Europe and Latin America interpreted the SJT scenarios and response options in the same way and attributed the same meaning to them. This is a fundamental precondition to rule out measurement effects and to investigate mean differences across (regional) groups (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Our results advance knowledge about the use of SJTs across geographical regions and cultures. Given SJTs' highly contextualized nature, comparing SJT scores across regions and cultures is viewed as a crucial challenge (e.g., Lievens, 2006; Ployhart & Weekley, 2006). Previous cross-cultural investigations of SJTs also showed mixed results when the SJT

development followed an imposed etic approach and did not include cross-regional/cultural input across all steps of SJT development (Lievens, Corstjens, et al., 2015; Such & Schmidt, 2004). However, as we demonstrated, integrating subject matter experts from different regions and cultures during the definition of the construct of measurement, the sampling of critical incidents, scenario writing, generation of response options, and setting the scoring key provides the fundament for SJTs to work well and be transportable across regions/cultures.

Although a combined emic-etic approach is time and resource intensive, it seems to pay off in terms of the cross-cultural application of assessment methods. Our work therefore attests to the success of relying on a combined emic-etic approach and extends similarly positive findings from research on the cross-cultural transportability of personality inventories (Cheung, Cheung et al., 2008; Cheung, Fan et al., 2008; Cheung et al., 1996; Schmit et al., 2000). To the best of our knowledge, this study is the first to apply a combined emic-etic approach of SJT development and to investigate its effects on measurement invariance across geographical regions. Our general recommendation is that the combined emic-etic approach serves as a viable strategy to develop SJTs for assessing 21st Century Skills across geographical regions.

Some caveats are in order, though. First, traditional, written SJTs with close-ended response formats do not measure behavior related to 21st Century Skills. Instead, they capture people's procedural knowledge about engaging in behavior related to these skills (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo et al., 2006). Recent research explored SJTs with other stimulus and response formats such as constructed response multimedia tests. These tests present short video clip situations to participants who then have to display their behavioral response in front of a webcam. Evaluations of these constructed responses have been shown to be valid indicators of job and training performance (Cucina et al., 2015; De Soete, Lievens, Oostrom, & Westerveld, 2013; Herde & Lievens,

2018; Lievens, De Corte, & Westerveld, 2015; Lievens & Sackett, 2017; Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2018; Oostrom, Born, Serlie, & van der Molen, 2010, 2011). Although constructed response multimedia tests add costs to SJT development (i.e., design of video clips and evaluation of participants' behavioral responses), they might complement current approaches to the assessment of 21st Century Skills. Given their dynamic audiovisual stimulus format and their audiovisual constructed response format, constructed response multimedia tests are even more contextualized than written, close-ended SJTs. Future research should therefore investigate whether constructed response multimedia tests developed according to a combined emic-etic approach also produce scores of 21st Century Skills that can be compared across regions and cultures.

As another limitation, we had data for only two geographical regions (Europe and Latin America). That said, this sample incorporated participants from eighteen different countries, thereby attesting to a huge cultural diversity. Nonetheless, further empirical research is necessary to replicate our results and examine the comparability of scores derived from SJTs across other geographical regions and cultures.

Conclusion

In sum, this paper is the first to investigate the combined emic-etic approach to develop SJTs to obtain scores that can be compared across geographical regions and cultures. Our results established metric measurement invariance across five SJTs for participants from Europe and Latin America. Hence, this study attests to the potential of the combined emic-etic approach. We therefore encourage researchers and practitioners to adopt this approach in cross-cultural research and practice for assessing 21st Century Skills.

References

Ainley, J., Fraillon, J., Schulz, W., & Gebhardt, E. (2016). Conceptualizing and measuring computer and information literacy in cross-national contexts. *Applied Measurement in Education, 29*, 291-309. https://doi.org/10.1080/08957347.2016.1209205

Bartram, D. (2012). *The SHL universal competency framework* (SHL White Paper). Thames Ditton, UK: SHL Group.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4*, 119-128. https://doi.org/10.1080/00207596908247261

Berry, J. W. (1990). Imposed etics, emics, and derived etics: Their conceptual and operational status in cross-cultural psychology. In T. N. Headland, K. L. Pike, & M. Harris (Eds.), *Frontiers of anthropology, Vol. 7. Emics and etics: The insider/outsider debate* (pp. 84-99). Thousand Oaks, CA: Sage Publications.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). Dordrecht, Nederland: Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460-502. https://doi.org/10.1177/0013164410375112

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321. https://doi.org/10.1207/s15328007sem1302_7

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of

nonequivalence. *International Journal of Testing, 10,* 107-132. https://doi.org/10.1080/15305051003637306

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64,* 225-262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education, 29,* 250-264. https://doi.org/10.1080/08957347.2016.1209204

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

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233-254. https://doi.org/10.1207/S15327043HUP1503_01

Cheung, F. M., Cheung, S. F., Zhang, J., Leung, K., Leong, F., & Huiyeh, K. (2008). Relevance of openness as a personality dimension in Chinese culture: Aspects of its cultural relevance. *Journal of Cross-Cultural Psychology, 39,* 81-108. https://doi.org/10.1177/0022022107311968

Cheung, F. M., Fan, W., Cheung, S. F., & Leung, K. (2008). Standardization of the cross-cultural Chinese Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment. *Acta Psychologica Sinica, 40,* 839-852. https://doi.org/10.3724/SP.J.1041.2008.01639

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology, 27,* 181-199. https://doi.org/10.1177/0022022196272003

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. https://doi.org/10.1207/S15328007SEM0902_5

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

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69*, 979-1006. https://doi.org/10.1111/1467-6494.696172

Church, A. T., & Lonner, W. J. (1998). The cross-cultural perspective in the study of personality: Rationale and current research. *Journal of Cross-Cultural Psychology, 29*, 32-62. https://doi.org/10.1177/0022022198291003

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

Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education, 29*, 310-318. https://doi.org/10.1080/08957347.2016.1209210

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358. https://doi.org/10.1037/h0061470

Geisinger, K. F. (2016). 21st century skills: What are they and how do we assess them? *Applied Measurement in Education, 29*, 245-249. https://doi.org/10.1080/08957347.2016.1209207

Getha-Taylor, H., Hummert, R., Nalbandian, J., & Silvia, C. (2013). Competency model design and assessment: Findings and future directions. *Journal of Public Affairs Education, 19*, 141-171. https://doi.org/10.1080/15236803.2013.12001724

Greiff, S., & Kyllonen, P. (2016). Contemporary assessment challenges: The measurement of 21st century skills. *Applied Measurement in Education, 29*, 243-244. https://doi.org/10.1080/08957347.2016.1209209

Headland, T. N., Pike, K. L., & Harris, M. (1990). *Frontiers of anthropology, Vol. 7. Emics and etics: The insider/outsider debate*. Thousand Oaks, CA: Sage Publications.

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

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82*, 903-918. https://doi.org/10.1037//0022-3514.82.6.903

Herde, C. N., & Lievens, F. (2018). Multiple speed assessments: Theory, practice, & research evidence. *European Journal of Psychological Assessment.* Advance online article. https://doi.org/10.1027/1015-5759/a000512

Herde, C. N., Wüstenberg, S., & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education, 29*, 265-277. https://doi.org/10.1080/08957347.2016.1209208

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74*, 201-222. https://doi.org/10.1093/poq/nfq011

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309. https://doi.org/10.1177/0022022189203004

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*, 93-114. https://doi.org/10.1207/S15327574IJT0102_1

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264-277. https://doi.org/10.1177/0022022104272905

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

Kurz, R., & Bartram, D. (2002). Competency and individual performance: Modelling the world of work. In I. T. Robertson, M. Callinan, & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 227-255). Chichester, UK: Wiley.

Kyllonen, P. C. (2012). *Measurement of 21st century skills within the common core state standards*. Presented at the Invitational Research Symposium on Technology Enhanced Assessments (TEA). Washington, DC. Retrieved from https://www.ets.org/Media/Research/pdf/session5-kyllonen-paper-tea2012.pdf

Lee, Y. (2009). Competencies needed by Korean HRD master's graduates: A comparison between the ASTD WLP competency model and the Korean study. *Human Resource Development Quarterly, 20*, 107-133. https://doi.org/10.1002/hrdq.20010

Leong, F. T. L., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology, 16*, 590-597. https://doi.org/10.1037/a0020127

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 279-300). Mahwah, NJ: Erlbaum.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*, 269-276. https://doi.org/10.1080/15305058.2017.1309857

Lievens, F., Corstjens, J., Sorrel, M. Á., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? *International Journal of Selection and Assessment, 23*, 361-372. https://doi.org/10.1111/ijsa.12120

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., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3-22. https://doi.org/10.1017/iop.2015.71

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. (2018). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology, 104*, 715-726. http://dx.doi.org/10.1037/apl0000367

Lucas, B. (2016). A five-dimensional model of creativity and its assessment in schools. *Applied Measurement in Education, 29,* 278-290. https://doi.org/10.1080/08957347.2016.1209206

Lunev, A., Petrova, I., & Zaripova, V. (2013). Competency-based models of learning for engineers: A comparison. *European Journal of Engineering Education, 38,* 543-555. https://doi.org/10.1080/03043797.2013.824410

Lytle, A. L., Brett, J. M., Barsness, Z. I., Tinsley, C. H., & Janssens, M. (1995). A paradigm for confirmatory cross-cultural research in organizational behavior. *Research in Organizational Behavior, 17,* 167-214.

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

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740. https://doi.org/10.1037/0021-9010.86.4.730

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey & C. B. Walker (Eds.), *Personnel selection and classification* (pp. 193- 232). Hillsdale, NJ: Erlbaum.

Morris, M. W., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Integrating emic and etic insights about culture and justice judgment. *Academy of Management Review, 24*, 781-796. https://doi.org/10.5465/amr.1999.2553253

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

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

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

Nishii, L. H., Ployhart, R. E., Sacco, J. M., Wiechmann, D., & Rogg, K. L. (2001). *The influence of culture on situational judgment test responses*. Presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego, CA.

OECD. (2014). *PISA 2012 Technical Report*. Paris, France: OECD Publishing.

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

Pike, K. L. (1967). Etic and emic standpoints for the description of behavior. In K. L. Pike (Ed.), *Language in relation to a unified theory of the structure of human behavior* (pp. 37-72). Den Haag, Nederland: Mouton & Co.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 345-350). Mahwah, NJ: Erlbaum.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment test comparable? *Personnel Psychology*, *56*, 733-752. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x

Sahoo, F. M. (1993). Indigenisation of psychological measurement: Parameters and operationalisation. *Psychology and Developing Societies*, *5*, 1-13. https://doi.org/10.1177/097133369300500101

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153-193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508. https://doi.org/10.1037/0021-9010.89.3.497

Strong, M. H., Burkholder, G. J., Solberg, E. G., Stellmack, A., & Presson, W. D. (2019). *Development and validation of a global competency framework for preparing new graduates for early career professional roles*. Manuscript submitted for publication.

Such, M. J., & Schmidt, D. B. (2004). *Examining the effectiveness of empirical keying: A cross-cultural perspective.* Presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, IL.

Van de Vijver, F. J. R. (2003). Test adaptation/translation methods. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 960-964). Thousand Oaks, CA: Sage Publications.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. https://doi.org/10.1177/109442810031002

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 295-322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application* (pp. 157-182). Mahwah, NJ: Lawrence Erlbaum.

Yang, K.-S. (2000). Monocultural and cross-cultural indigenous approaches: The royal road to the development of a balanced global psychology. *Asian Journal of Social Psychology, 3*, 241-263. https://doi.org/10.1111/1467-839X.00067

Appendix

Table A1

Definitions of SJT Competencies

Competency	*Definition*
Achieving Objectives	Accepts or sets demanding individual goals. Meets individual goals and objectives. Takes initiative to seek additional responsibilities, as appropriate. Evaluates work outcomes to ensure quality standards are met.
Adapting to Change	Adjusts work style and interpersonal behavior to fit different situations and environments. Accepts and integrates new ideas and information on their merits. Supports and complies with change initiatives. Works effectively when faced with ambiguity.
Analyzing & Solving Problems	Critically evaluates information and its sources. Identifies gaps in information and seeks appropriate sources to close them. Synthesizes and integrates information into what is already known about a topic. Recognizes patterns in information to identify the bigger picture. Follows best practices and appropriately analyzes quantitative and qualitative data. Identifies and independently solves work problems, as appropriate. Considers multiple approaches when solving problems.
Learning & Self-Development	Identifies and addresses own knowledge gaps and training needs. Continually expands own knowledge and skills. Applies knowledge and training to professional contexts. Critically evaluates own strengths and weaknesses and pursues development. Seeks feedback and learns from successes and failures. Learns from others and seeks mentors.
Working Well with Others	Develops and maintains effective working relationships. Interacts effectively with people from different backgrounds. Listens to others and values and incorporates diverse viewpoints. Supports team decisions once they have been made. Adjusts own workload to help meet team commitments, as appropriate. Recognizes and demonstrates empathy for others' feelings, needs, and concerns. Appropriately resolves own work disagreements.

CHAPTER 3: MULTIPLE SPEED ASSESSMENTS: THEORY, PRACTICE, AND RESEARCH EVIDENCE[3]

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.

[3] This chapter is based on: Herde, C. N., & Lievens, F. (2018). Multiple Speed Assessments: Theory, practice, and research evidence. *European Journal of Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1027/1015-5759/a000512

Introduction

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), in recent times, various new selection approaches and technologies 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 5 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., carryover 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).

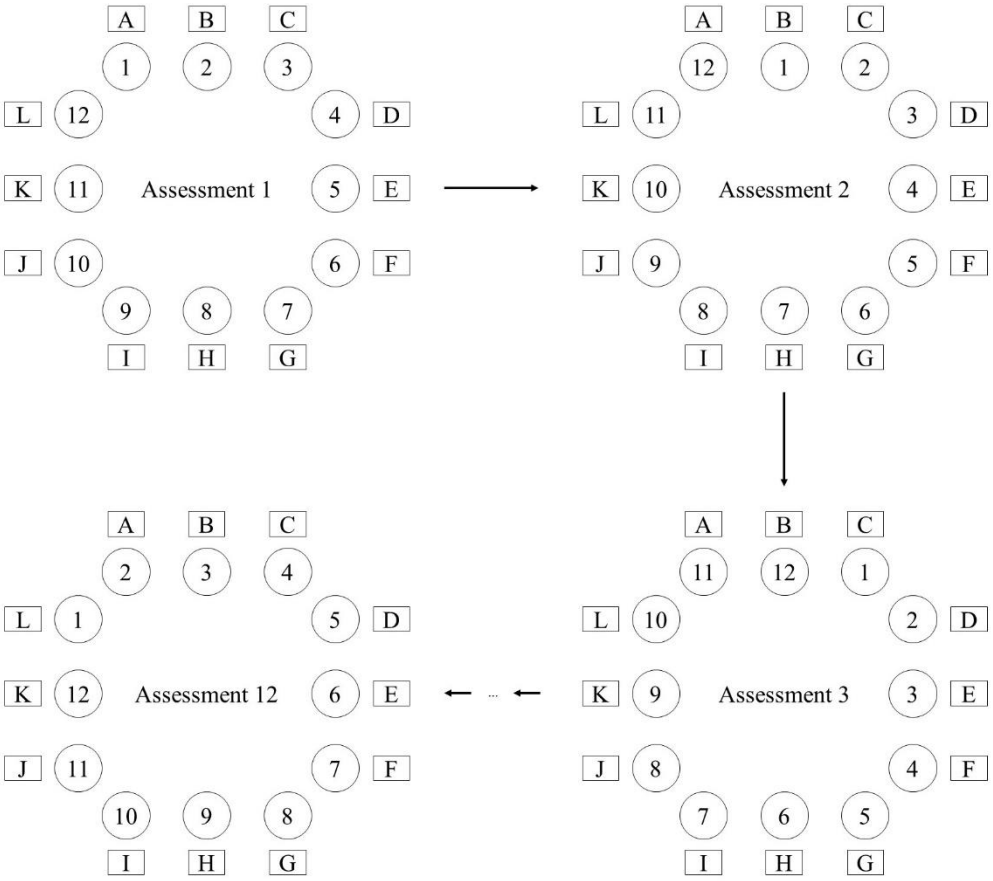

Figure 1. Schematic example of a Multiple Speed Assessment. In this example, 12 assessees (circles) simultaneously walk through a Multiple Speed Assessment that contains 12 simulations (rectangles). After each simulation, each participant goes on to a different simulation where they face again a role-player. Role-players may be seated on different tables or in different (virtual) rooms. This procedure repeats until all participants participated in all simulations.

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 prespecified 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.

Theoretical Fundaments of Multiple Speed Assessments

Zero/Minimal Acquaintance Paradigm

The "zero/minimal acquaintance" paradigm provides a first conceptual cornerstone for Multiple Speed Assessments. There exists a long-standing 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-ratings 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, 2018).

Moreover, Multiple Speed Assessments build upon evidence that the accuracy of judgments of multiple variables does 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 studies 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 run 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 contributes 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).

Prior Examples of Multiple Speed Assessments

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 makeup, 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 makeup 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[4] (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).

[4] Applicants also report being satisfied with a duration of 6 and 8 minutes (Cameron & MacKeigan, 2012).

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.

Purposes of Multiple Speed Assessments

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).

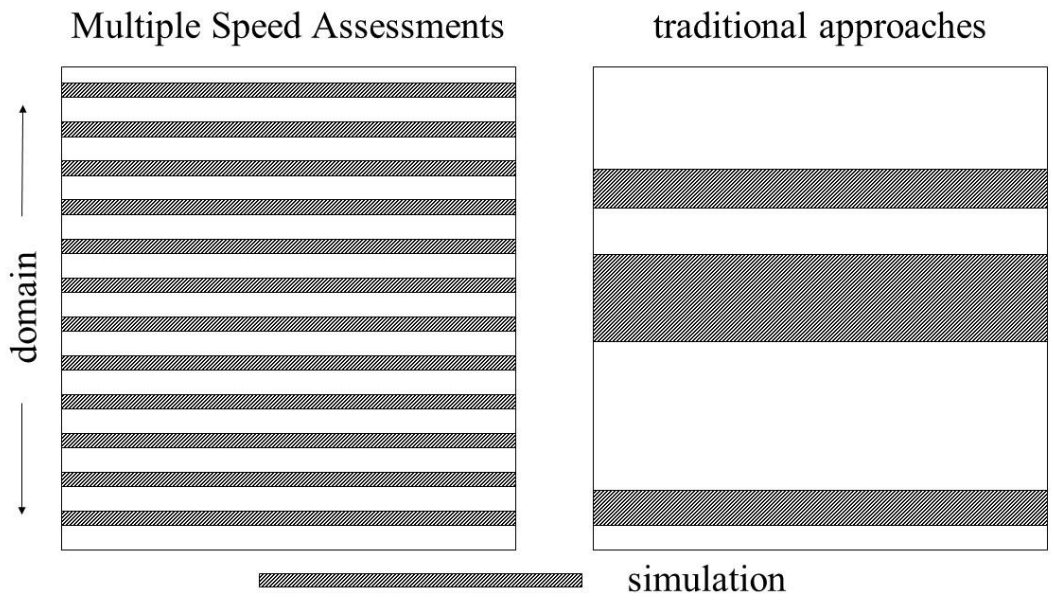

Figure 2. Schematic illustration of domain sampling by Multiple Speed Assessments compared to traditional approaches. The short duration of each simulation enables Multiple Speed Assessments to sample the domain more comprehensively than traditional approaches that build upon long-lasting, but fewer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

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[5] 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-reports 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 seems 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 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,

[5] 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).

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 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-reports and other reports of interpersonal adaptability? How do different performance trajectories across simulations relate to self-reports 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 the 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.

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)

Table 2 (Continued)

Summary of Empirical Evidence for Different Examples of Multiple Speed Assessment

	Constructed Response Multimedia Test	OSCE	MMI
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* ≤ r ≤ .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 ΔR^2 = .04* Supervisor ratings ΔR^2 = .00 (Oostrom et al., 2010) Selection decision ΔR^2 = .03* (Lievens et al., 2015) Training performance ΔR^2 = .03-.08* (Lievens et al., 2015)		In-programme and licensing examination performance (Knorr & Hissbach, 2014) Applied knowledge test ΔR^2 = .01 Clinical decision making skills examination ΔR^2 = .02* OSCE ΔR^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)

Table 2 (Continued)

Summary of Empirical Evidence for Different Examples of Multiple Speed Assessment

	Constructed Response Multimedia Test	OSCE	MMI
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) Ethnic majority vs. minority $d = .14$ (DeSoete et al., 2013) $d = .44^*$ (Lievens et al., 2015) White-Black $-.10 \leq d \leq .00$ White-Hispanic $.11^* \leq d \leq .22^*$ (Cucina et al., 2015)	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; average $d = .27^*$)	Majority of studies indicates equal scores across gender, age, or socio-economic subgroups (Rees et al., 2016)
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.

Conclusion

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. https://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. https://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, UK: 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. https://doi.org/10.1002/per.2119

Byham, W. (2016, October). *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. https://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

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*, 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Fetzer, M. (2015). Serious games for talent selection and development. *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. http://dx.doi.org/10.1037/0022-3514.60.5.773

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*, 144. https://doi.org/10.1186/s12909-015-0424-0

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. http://dx.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

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. https://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 situations.

Journal of Personality and Social Psychology, 103, 1007–1022. https://doi.org/10.1037/a0030385

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 J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (Vol. 28, pp. 99–152). Bingley, UK: 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. https://doi.org/10.1177/0149206314525207

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, NY: 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. https://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. http://dx.doi.org/10.1037/0021-9010.73.3.482

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*, 372-376. https://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.

CHAPTER 4: MULTIPLE SPEED ASSESSMENTS UNDER SCRUTINY: ARE THEIR RATINGS RELIABLE AND VALID?[6]

In recent times, shorter and faster assessments (e.g., "flash" interviews, "speed" role-plays) have emerged in selection practice. These shorter speed assessments raise questions whether they can serve as reliable and valid indicators of future performance. The objective of this paper is to scrutinize such a multiple speed assessment approach. We develop hypotheses on the basis of the minimal acquaintance/thin slices paradigm to test the reliability and validity of multiple, short interpersonal simulations that sample the leadership domain. Our sample consisted of 96 MBA students that participated in eighteen 3-minute interpersonal leadership role-plays. Results showed low single-rater reliabilities for role-play performance ratings. Acceptable reliability (>.70) was achieved if role-play performance ratings were aggregated across fourteen role-plays with two independent assessors or across nine role-plays with three independent assessors. Overall multiple speed assessment performance averaged across the eighteen role-plays revealed information about participants' cognitive ability, extraversion, and agreeableness. Further, overall multiple speed assessment performance predicted performance seven months later and added incremental validity above these traditional predictor constructs. Implications for selection theory, research, and practice are discussed.

[6] This chapter is based on: Herde, C.N., & Lievens, F. (2019). Multiple Speed Assessments under scrutiny: Are their ratings reliable and valid? In R. Hewett (Chair), *HR Assessments and Employee Responses.* Symposium conducted at the 79th Annual Meeting of the Academy of Management, Boston, MA, USA.

This paper has been judged by anonymous reviewers to be one of the best accepted papers in the conference program.

Introduction

"And so we're trapped in an economy that has become all about efficiency […]"

Haque (2016)

As the quote above suggests, an increasing number of phenomena in private and professional life accentuate the trend to maximize efficiency by making decisions on the basis of limited information. For example, many people use dating apps or participate in speed dates to screen potential candidates for a romantic evening. At work, speed networking has gained in popularity as people briefly and quickly meet up with possible suppliers, clients, or sparring partners (Bick, 2007). In a similar vein, investors try to make profitable investment decisions on the basis of short pitches of entrepreneurs.

The magic words "short" and "fast" have also entered the personnel selection arena (e.g., Pinsight, 2019). To respond to calls for short, fast-paced, and more engaging assessment processes that represent today's hectic and fragmented work life (Liff, 2017), selection practitioners have added new assessment approaches to their portfolio under the umbrella term of "multiple speed assessments" (Herde & Lievens, 2018). In multiple speed assessments, people participate in multiple short interpersonal/leadership simulations (typically about three minutes) that confront them with a variety of job situations and role-players. The evaluation process is streamlined, with assessors usually providing only a single evaluation of participants' performance per simulation. Examples of these speed assessments are "flash" interviews or fast-track interview sessions (Needleman, 2007), multiple short face-to-face simulations (Byham, 2016; Mendas, 2018), and brief webcam role-plays (Pinsight, 2018).

Some researchers have also promoted the use of shorter and faster assessments because they might be designed to more clearly capture aspects of the performance domain

(Brannick, 2008; Lievens, 2008). However, at the same time such multiple speed assessments might also contradict essential principles of test theory. That is, can someone really be reliably and validly assessed via a three-minute snapshot? So far, empirical research on (multiple) short simulations is virtually non-existent. Motowidlo, Hooper, and Jackson (2006) investigated relations between behavior in eight short role-plays, implicit trait policies as well as personality. However, little information was gained about the role-plays because they served as criterion measures. Recently, Ingold, Dönni, and Lievens (2018) investigated assessor ratings in four simulations based upon limited information (i.e., snap judgments). In this study, one group of assessors saw the full-length performance of participants, whereas another group of assessors 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. Although the Ingold et al. study does not really deal with multiple speed assessment, it does suggest that the reliability and validity of ratings made on the basis of limited information should not be taken for granted.

Given this lack of research, many questions about multiple speed assessments remain unanswered: What is the theoretical fundament that may support a selection approach that assesses participants' performance in multiple 3-minute simulations? Do ratings based upon such multiple short simulations reliably assess participants' performance? Do such ratings reveal information about participants' fundamental individual differences like cognitive ability and personality? Do ratings based upon multiple, short simulations predict participants' job-related performance and do they do so beyond traditional predictors? To the best of our knowledge, there are no empirical investigations that provide answers to these

pressing questions. Hence, the multiple speed assessment approach seems to be a practice that so far has moved ahead of research and rigorous empirical scrutiny.

Therefore, this paper aims to scrutinize a multiple speed assessment approach in terms of key psychometric criteria. To this end, we examine both the reliability and validity of assessor ratings based on multiple, short simulation performances in the leadership domain. This paper contributes to the personnel selection domain in various ways. At a theoretical level, we embed the multiple speed assessment approach into the zero acquaintance and thin slices paradigm in our hypothesis development. Further, we are the first to examine evidence related to the reliability and validity of speed assessments. At a practical level, we detail how one can vary design factors (i.e., number of simulations and assessors) to increase the reliability of multiple speed assessments. Our examinations of the predictive and incremental validity also provide information to decide about the viability of adopting a multiple speed assessment.

Study Background

The Minimal Acquaintance/"Thin slices" Paradigm

Research on the zero/minimal acquaintance and thin slices paradigm might provide a theoretical fundament for assessment approaches that build upon fast and short simulations. Research in personality and social psychology applied the zero acquaintance paradigm (Albright, Kenny, & Malloy, 1988), also referred to as minimal acquaintance/thin slices paradigm, to investigate how people form judgments about strangers. In the minimal acquaintance/thin slices paradigm, untrained people ("judges") are asked to rate others ("strangers") on the basis of only minimal information. Often, these minimal information situations represent short, dynamic excerpts ("thin slices") from the behavioural stream of strangers that last between several seconds to not more than five minutes (Ambady, Bernieri, & Richeson, 2000). In a seminal study, Borkenau, Mauer, Riemann, Spinath, and Angleitner

(2004) asked multiple judges to watch video tapes of strangers in various short unstructured situations (e.g., introducing themselves, telling a story, reading aloud newspaper headlines, convincing a neighbour to lower the radio volume) and rated them on personality and intelligence. This shows why this paradigm is relevant for shedding light on the multiple speed assessment approach that is now making inroads in selection practice.

Throughout the past years, many personality and social psychology studies applied the minimal acquaintance/thin slices paradigm to investigate judgments on a vast array of psychological constructs like personality, intelligence, self-esteem, or performance. In these studies, judges based their judgments upon limited, static or dynamic behavioral information (e.g., Ambady et al., 2000; Ambady & Rosenthal, 1992; Connelly & Ones, 2010). Across this body of research, the following three research streams can be distinguished that are insightful for building our hypotheses about the reliability and validity of multiple speed assessments.

Are Judgments Based Upon Minimal Information Reliable?

A first stream of research has examined whether judges can provide reliable judgments of strangers on the basis of only minimal information. The results of various meta-analyses of single-rater reliabilities indicate some evidence of correspondence of judgments from different judges, although large idiosyncratic judge effects remain. That is, a meta-analysis that summarized single-rater reliabilities of judgments on the Big Five based upon different types of minimal information revealed coefficients between $r = .23$ and $r = .40$ (Connelly & Ones, 2010). To obtain more reliable judgments about strangers based upon only minimal information, this meta-analysis also concluded to aggregate judgments of many judges (Connelly & Ones, 2010). Recently, Eisenkraft (2013) also compellingly demonstrated this: Judges watched muted short videos in which students were interviewed. Afterwards, judges provided their overall impressions of the students. Single-rater reliabilities were comparable

to past research on minimal acquaintance/thin slices judgments (ICC1 = .16). In contrast, the average judgment across all judges (41 judges) was much more reliable (ICC2 = .89).

Clearly, in multiple speed assessment, it is not feasible to rely on such a large number of assessors. Hence, the single-rater reliabilities of judgments in the reported meta-analysis should serve as benchmarks for the level of reliability that might be expected per simulation for single assessors in multiple speed assessment. That said, we do expect that these meta-analytic single-rater reliabilities set a lower bound of the reliability of single assessors' ratings in multiple speed assessment because multiple speed assessment builds in several critical safeguards to increase reliability. These safeguards are built in because the ratings in multiple speed assessment are of high-stakes for participants (feedback) and organizations (selection decisions) (cf. lab situations in minimal acquaintance studies). First, assessors in multiple speed assessments follow intensive trainings to limit observation errors (e.g., Byham, 1977) and adopt a common frame of reference for evaluating performance (Roch, Woehr, Mishra, & Kieszczynska, 2012). Research confirmed that such training approaches enhance interrater reliability and rating accuracy (Lievens, 2001; Roch et al., 2012). In contrast, judges in minimal acquaintance/thin slices studies usually do not follow any training. Second, assessors in multiple speed assessments use behavioral observation aids (e.g., BOS, BARS). Past research suggested the use of observational aids to further benefit reliability (Lievens, 1998). Conversely, many minimal acquaintance/thin slices studies focus on heuristic initial impressions and do not regularly employ detailed rating aids. Third, in multiple speed assessments multiple relevant behaviors are elicited from participants via situational cues (i.e., prompts; Schollaert & Lievens, 2011, 2012) to ensure structure (standardization) and facilitate evaluation, thereby increasing interrater reliability (Brannick, 2008; Lievens, Schollaert, & Keen, 2015). In contrast, in minimal acquaintance/thin slices research, strangers are typically placed in unstructured situations that only rarely elicit multiple relevant behaviors. Some

minimal acquaintance studies provided even only irrelevant stimuli about strangers (Kenny & West, 2008).

In conclusion, a multiple speed assessment approach shares various similarities with minimal acquaintance/thin slices situations. Yet, it is also different in other aspects because the stakes are higher and at least three safeguards (training, rating aids, prompts) are therefore implemented to ensure levels of interrater reliability of at least .70[7]. Hence, we propose: *Hypothesis 1 (H1): The interrater reliability of assessors' ratings in short interpersonal simulations will be at least .70.*

Do Judgments Based Upon Minimal Information Reveal Meaningful Personality and Ability Information?

A second research stream dealt with relationships between minimal acquaintance/thin slices judgments and traditional individual differences measures (e.g., cognitive ability measures or self- and close-acquaintance ratings of personality). A large body of research attested that judgments based upon minimal information reveal information about cognitive ability. A meta-analysis summarized that even little, static information like photographs enable making judgments that relate as close as r = .28 to strangers' scores on ability measures (Zebrowitz, Hall, Murphy, & Rhodes, 2002). This finding has been corroborated by studies that presented dynamic behavioral information to judges (Borkenau & Liebler, 1993; Borkenau et al., 2004; Carney, Colvin, & Hall, 2007; Murphy, 2007; Murphy, Hall, & Colvin, 2003; Reynolds & Gifford, 2001). In some of those studies, correlations with cognitive ability scores rose up to r = .43. Thus, even minimal information seems to reveal cues to allow judgments about cognitive ability.

[7] Although thresholds of reliability are dependent on the context and indicator of reliability, we follow rules of thumb from LeBreton and Senter (2007) to interpret interrater reliabilities below .51 as low, between .51 and .70 as moderate, and between .71 and .90 as high, and between .91 and 1.00 as very high.

Given the promising evidence from research on the minimal acquaintance/thin slices paradigm, we expect ratings in multiple speed assessments to reveal information about participants' cognitive ability. The multiple, short simulations in multiple speed assessments confront participants with various different problem situations that have to be swiftly solved. Therefore, they should provide excellent conditions to elicit behaviors related to efficient, accurate information processing and quickly adjusting to new situations, which are quintessential to cognitive ability (Kanfer & Ackerman, 1989). Examples of these behaviors might be asking targeted questions to further understand the problem or suggesting appropriate solutions. Thus, the short simulations in multiple speed assessments should enable trained assessors to make judgments related to cognitive ability.

Hypothesis 2 (H2): Assessors' ratings of participants in short interpersonal simulations will be significantly related to participants' cognitive ability.

Similar to the results about cognitive ability judgments, empirical research showed that judgments based upon minimal information reveal meaningful information about strangers' personality. Two meta-analyses (Connelly & Ones, 2010; Connolly, Kavanagh, & Viswesvaran, 2007) analyzed the convergence between judgments on Big Five traits based upon different types of minimal information: The mean correlations between self- and judges'-ratings reached values up to .29 (Connolly et al., 2007) and .22 (Connelly & Ones, 2010). Another meta-analysis on short, dynamic behavioral information found similar results (r = .20, Ambady et al., 2000).

Given the promising evidence from research on the minimal acquaintance/thin slices paradigm, we expect that ratings in multiple speed assessments reveal meaningful information especially about three Big Five traits. First, we expect the short interpersonal simulations to be prime vehicles for activating behavior related to interpersonal traits like extraversion and agreeableness (Leising & Bleidorn, 2011; McCrae & John, 1992). Related to extraversion, the

multiple role-plays likely elicit behavior that indicates whether participants enthusiastically approach others, are talkative, and enjoy interpersonal encounters versus whether they prefer to keep more of a distance from others and show a reserved body language (McCrae & John, 1992).

Second, the multiple role-plays confront participants with situations that activate behavior related to cooperation, negotiation, and interpersonal sensitivity on the part of leaders, which are indicative of the broader trait of agreeableness (McCrae & John, 1992). Consistent with these arguments, subject matter experts confirmed that role-play simulations activate individual differences in extraversion and agreeableness (Lievens, Chasteen, Day, & Christiansen, 2006). Thus, even in the short, fast-paced interpersonal simulations, we expect that enough relevant behavioral information should be available to trained assessors to make ratings related to participants' extraversion and agreeableness levels.

Third, the multiple interpersonal simulations confront participants with a large variety of unique and unusual problem situations that create prime conditions to elicit behavior related to openness. That is, the many different problem situations require them to investigate different angles of the problems, and/or to develop different and innovative solutions, all of which are prototypical for openness (McCrae & John, 1992; Mussel, 2013). Moreover, participants meet eighteen different characters and situations, prompting them to swiftly adapt to changing demands. In other words, we expect that sufficient relevant behavioral information should be available to trained assessors in multiple speed assessment to make ratings about participants' openness.

Hypothesis 3 (H3): Assessors' ratings of participants in short interpersonal simulations will be significantly related to participants' extraversion.

Hypothesis 4 (H4): Assessors' ratings of participants in short interpersonal simulations will be significantly related to participants' agreeableness.

Hypothesis 5 (H5): Assessors' ratings of participants in short interpersonal simulations will be significantly related to participants' openness.

Do Judgments Based Upon Minimal Information Predict Relevant Outcomes?

A third major research stream in the minimal acquaintance/thin slices paradigm 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). In this meta-analysis, the outcomes included, amongst others, deceptive and honest behavior, depression, or referral of alcoholic patients. Overall, thin slices judgments predicted these outcomes with an average of r = .39. Another meta-analysis conducted (Ambady et al., 2000) extended the available evidence for 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).

Although research on the predictive validity of judgments based upon minimal information is impressive, two important notes are in order. On one hand, many lab studies in social and personality psychology on the validity of minimal acquaintance/thin slices judgments typically report evidence for judgments averaged across multiple independent judges. For example, in the meta-analysis of Ambady and Rosenthal (1992), the median number of judges was 37, with a range between 2 and 446. Clearly, this aggregation across many independent judges reduces judges' idiosyncrasies and biases. In turn, as shown by Eisenkraft (2013), this increased reliability boosts the validity coefficients obtained. As noted, in an operational multiple speed assessment, such a large number of assessors is not feasible, which might reduce the validities obtained (depending on the reliability obtained, see H1).

On the other hand, research on minimal acquaintance/thin slices has advanced our understanding of what constitutes "good information" to make judgments about strangers.

Rather than increasing the observation time (e.g., beyond 1-2 minutes) and the quantity of information, research revealed the importance of obtaining multiple, *qualitatively different* pieces of information about strangers (Ambady et al., 2000; Ambady & Rosenthal, 1992; Back & Nestler, 2016; Carney et al., 2007; Connelly & Ones, 2010; Murphy, 2005; Murphy et al., 2015). This can be done by observing strangers across multiple, qualitatively different situations. This moderator matches well with the notion of increasing the point-to-point correspondence between predictor and criterion domain in validity research (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). Thus, both research on "good information" and the logic underlying behavioral sampling concur that more valid assessments of participants can be made when ratings are based on multiple, qualitatively different pieces of criterion-relevant information. This is exactly what occurs in multiple speed assessments because assessors evaluate participants in multiple, short situations that elicit behavior related to many qualitatively different parts of the criterion domain, thereby broadening the criterion coverage of the predictor and increasing the validity of the ratings obtained.

In sum, we expect that in multiple speed assessments these two aforementioned aspects (lower number of assessors but broader domain coverage) will balance each other out in terms of their effects on validity. Thus, we propose:

Hypothesis 6 (H6): Assessors' ratings of participants in multiple short simulations will significantly predict performance.

Methods

Sample

To gather data about a multiple speed assessment approach, we collaborated with a European business school that aimed to reinvigorate the assessment/admission procedure of their MBA program. Therefore, the multiple speed assessment was implemented for developmental purposes: That is, the entire MBA cohort of this business school participated

in our study to identify their strengths and weaknesses as leaders. The sample encompassed 96 participants (51% females, mean age = 23.63, *SD* = 1.85) from 19 different nations (67 % Belgian, 5 % Chinese, 4 % Romanian) who were studying to obtain a Master's degree in Marketing (51 %) or Financial Management. All had at least one year of working experience. The mean test-taking motivation of the participants was high: 3.96 (*SD* = 0.50, see below for the scale). Further anecdotal evidence supported that participants were motivated to perform well in the multiple speed assessment and to learn about their potential as leaders. For example, all participants wore business attire and appeared to be nervous.

Procedure

About one week prior to the multiple speed assessment, participants were invited to complete proctored computer-based tests which included measures of cognitive ability (a cognitive ability test and an in-basket) and a Big Five personality measure. The actual multiple speed assessment took place in a large hall and lasted 90 minutes. In the multiple speed assessment**,** participants completed 18 different role-play simulations in which they interacted with 18 different role-players. In the large hall, a circle was formed by desks. On each desk, one role-player was sitting. 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 stayed at the same desk and played the same role-play again[8]. Participants' performance during each of the role-plays was rated by the respective role-player (who thus also served as assessor, see below). In a later stage, two to three independent assessors also rated video recordings of the role-plays. Afterwards, participants received feedback reports

[8] To reduce role-player fatigue, role-players (a) played two different role-plays on each assessment day, (b) enjoyed scheduled breaks, and (c) were replaced by flying role-players in two to seven instances per day.

about their performance in the multiple speed assessment. Seven months after the speed assessment, the MBA supervisors (instructors) and peers rated the performance of the participants to provide us with criterion data.

Measures

The computer-based tests were developed by an international HR consultancy. All these instruments were validated and some of them were certified by the British Psychological Society (BPS). As we did not receive access to item-level data, we could not calculate internal consistency reliabilities.

Cognitive ability. To assess general cognitive ability, we used two different measurement approaches (Lievens & Reeve, 2012). First, we used a matrix-type figural reasoning test (Bogaert, Trbovic, & Van Keer, 2005). Various studies showed that matrix-type figural reasoning tests are good indicators of general cognitive 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 (Cronbach's alpha = .91), split-half reliability (Spearman-Brown formula = .94), and correlations with the Advanced Progressive Matrices (Raven, 1958) of r = .52.

Second, we used a computer-based in-basket as alternative measure of cognitive ability. In the in-basket (Volckaert & Dereuddre, 2013), participants assumed the role of a junior manager that had recently joined a new organization. They were assigned the role of a project coordinator of an event. In the in-basket, they were confronted with e-mails that provided information about the event and had to answer questions on how to cope with troublesome issues. For each question, participants had to indicate their agreement with several response options (1 = *strongly disagree*, 5 = *strongly agree*). The test manual supported the adequate psychometric properties of the in-basket in terms of internal consistency reliability, construct-related validity in terms of relations to cognitive ability (in

this study: $r = .21$, $p = .045$), and criterion-related validity in terms of relations to MBA study results (Volckaert & Dereuddre, 2013).

To gain an overall cognitive ability measure, we standardized the scores from the cognitive ability test as well as the in-basket and calculated a composite across both measures.

Big Five personality. We assessed personality with the Business Attitudes Questionnaire (Vrijdags, Bogaert, Tribovic, & Van Keer, 2014). This is a work-related personality questionnaire. Each item of the BAQ asks participants to indicate agreement with a 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 BPS. We used participants' summed scores on the Big Five scales. The test manual reports good psychometric properties in terms of internal consistency reliabilities ($.91 \leq \alpha \leq .94$), convergent validity with other contextualized and non-contextualized personality inventories, and criterion-related validity in terms of relations with job performance.

Multiple Speed Assessment. We developed eighteen different role-play simulations to sample relevant situations in the leadership domain. To derive the content of these role-plays, we drew from two sources. First, we built upon leadership theories such as the Multiple-Linkage Model (Yukl, 2010). Second, experienced consultants from the HR consultancy agency served as subject matter experts. They were qualified as experts in the leadership domain because they had provided solutions to select and develop successful leaders in multiple client projects. These two sources inspired the development of the situations of the 18 different role-plays to cover different parts of the leadership domain. For example, participants faced an employee who asked for help in an inter-employee conflict in one role-play. In another role-play, another employee criticized the participant for being too slow in decision-making.

All role-plays were integrated into an overarching background (i.e., the organization of the event, see in-basket). That is, each of the 18 role-plays confronted the participant as project coordinator with a different character from inside or outside the organization that mentioned a specific problem. This common overall background of all role-plays aimed to enhance realism and participants' immersion into the role-plays. Moreover, due to this overall background and the fact that participants had to process it by completing the in-basket, the amount of information given prior to each role-play was brief (at most one sentence).

A total of 30 role-players (80 % females) participated in the multiple speed assessment. These were either experienced consultants from the HR consultancy or graduate students from a large European university. As noted, each of these role-players were also live assessors. Apart from the live assessors (who were also role-players), in a later stage, recorded performances were rated by 20 other trained and paid assessors (60 % females, mean age = 21.90, *SD* = 4.20). These were recruited from a European university. All of them were studying to obtain a Bachelor's (60 %) or Master's degree in various fields of Psychology or Business Administration.

Assessor training for the consultants from the HR consultancy and students included core aspects of both behavior-driven (Byham, 1977) and frame-of-reference training (Roch et al., 2012). The training included lectures and exercises on observation, registration, classification, and evaluation of participants' performance. In addition, assessors were familiarized with the overall scoring procedure. Next, they practiced evaluating participant performances in the role-plays they specialized in. To this end, 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.

Consistent with role-player training guidelines (Byham, 1977; Lievens, Schollaert, & Keen, 2015), we started by giving role-players a short introduction to the overarching

background (i.e., the organization of the event). Role-players were then taught to use standardized prompts (Lievens, Schollaert, & Keen, 2015; Schollaert & Lievens, 2011, 2012) to structure the role-plays and elicit behavior. An example of a prompt 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 prompts by heart. Finally, role-players also practiced and received feedback about their role-playing behavior.

To ensure that role-players (assessors) focused their performance ratings on observable and relevant behaviors, we developed short checklists per role-play that listed behaviors indicative of effective performance. Per role-play, they also provided overall ratings of role-play performance (*1 = should clearly be improved: starters' level* to *9 = obviously strong: role model behavior*). Across role-plays and role-players/assessors, average internal consistency reliabilities for role-play performance ratings was .67 (*SD* = .21). Role-players' role-play performance ratings from the eighteen role-plays were averaged into a measure of overall multiple speed assessment performance.

For assessors that watched recorded performances in a later stage, the rating procedure was the same. Rewinding or pausing role-play conversations was prohibited.[9] To limit the influence of biases (e.g., order effects) in these assessor ratings, we took various precautions: (a) we distributed all records for the two role-plays per assessor across four blocks that each contained records of only one role-play, (b) we counterbalanced the appearance of records per role-play across assessors, and (c) we presented participants per role-play in a random order.

Control measures. We assessed participants' test taking motivation via a scale with four items from Arvey, Strickland, Drauden and Martin (1990) via a 5-point Likert scale (*1 = totally disagree; 5 = totally agree*; internal consistency reliability = .67). We also included participants' gender and age as control variables.

[9] For 20 percent of all conversations, cameras did not successfully record videos so that only audio records were available. In these cases, assessors used audio records to evaluate performance.

Criterion Measures. Each participant was rated by the instructors and peers (class mates) of the MBA program. Instructors rated various criteria that are related to the performance of leaders (task-oriented leadership, relation-oriented leadership, task adaptability, team member adaptivity, task performance and interpersonal contextual performance). Instructors provided the ratings via the relative percentile method (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Goffin, Jelley, Powell, & Johnston, 2009). In the relative percentile method, raters assign percentile scores to ratees. Goffin and colleagues developed this method to reduce rating inflation as the reference group to be used for the ratings consists of the average MBA student (i.e., a percentile score of 50). Prior research showed that the relative percentile method had higher criterion-related validity than conventional absolute rating formats (Goffin et al., 1996, 2009).

To investigate the factor structure of the instructor ratings, we conducted a series of confirmatory factor analyses in Mplus 7.4 (Muthén & Muthén, 1998-2015) by using the MLR estimator. The ratings from the relative percentile method were used as indicators. We compared a one-factor model (Model A), a model with the two correlated factors of task- and interpersonal-oriented performance (Model B), and a model with three correlated factors: task-, interpersonal-, and change-oriented performance (Johnson, 2001; Yukl, 1999). All three models showed good model fit, with exception of poor RMSEA values (Model A/B/C: χ^2(df) = 35.39(9)/34.03(8)/14.50(6), $p < .001/< .001/.0245$, CFI = .900/.901/.968, RMSEA (90% CI) = .176 (.117-.238)/.185 (.124-.251)/.122 (.041-.204), SRMR = .066/.061/.034). Information criteria indicated that Model C showed the best model fit (Model A/B/C: AIC: 5022.40/5020.49/5002.15). Next, we looked at parameter fit. Given that the latent factors for Model C showed high intercorrelations ($.71 \leq r \leq .99$, ps $< .001$), we opted for the most parsimonious solution and averaged ratings into an overall performance measure. Internal consistency reliabilities for this measure was .89.

Peers rated the same criterion dimensions as instructors via multi item scales (Griffin, Neal, & Parker, 2007; Motowidlo & Van Scotter, 1994; Williams & Anderson, 1991; Yukl, 1999), using a 5-point Likert scale (*1 = below average*; *5 = truly exceptional*). To reduce leniency in ratings, instructors assigned class mates as peers when they knew a participant well (e.g., due to project work). Peers knew the target participants between 6 and 279 months (M = 12.04, SD = 24.64). For their participation in the criterion study, participants received a coupon of 5 € and the chance to win another coupon of 100 € in a lottery. All but two participants were rated by at least one peer.[10] To investigate the factor structure of the peer ratings, we conducted the same confirmatory factor analyses as for the instructor ratings. For the peer ratings, we used mean scores across all items for each of the six scales as factor indicators. Model B and Model C showed good model fit, with the exception of poor RMSEA values (Model A/B/C: χ^2(df) = 165.17(9)/26.91(8)/32.10(6), p < .001/< .001/< .001, CFI = .579/.949/.930, RMSEA (90% CI) = .379 (.329-.430)/.140(.084-.200)/.190(.128-.256), SRMR = .109/.077/.055). Information criteria indicated that Model B showed the best fit to the data (Model A/B/C: AIC = 1508.76/1379.82/1380.77). Investigation of parameter fit indicated that the latent factors of the best fitting model (B) showed a high intercorrelation (r = .59, p < .001). Thus, in line with the instructor ratings, we averaged all peer ratings into an overall performance measure. Internal consistency reliability for this measure was .89.

Results

Are Ratings in Short and Fast-paced Simulations Reliable?

Our first hypothesis (H1) proposed that assessors' ratings in short and fast-paced simulations will reach levels of interrater reliability of at least .70. To investigate this hypothesis, we analyzed role-play performance ratings from all available assessors. Per role-play, participants were nested within role-players, but fully crossed with assessors. Therefore,

[10] We did not calculate interrater reliabilities for peer ratings of performance because we received a second peer rating for only 28 participants.

our design resembled an ill-structured measurement design (Putka, Le, McCloy, & Diaz, 2008). In ill-structured measurement designs, traditional indices of interrater reliability (e.g., intraclass correlations) generate biased estimates because they do not explicitly distinguish between the contribution of assessor main effects, assessor-participant interaction effects, and residual variance to observed score variance (Putka et al., 2008). Putka and colleagues (2008) therefore proposed to calculate interrater reliabilities with the $G(q,k)$ coefficient that describes the proportion of expected observed score variance that is attributable to true score variance.

Table 1 shows single-rater and average interrater reliabilities for role-play performance ratings. Single-rater reliabilities ($G[q,1]$) for role-play performance ratings were low to moderate (.18-.64, M = .38, SD = 0.12). Interrater reliabilities for role-play performance ratings averaged across all assessors ($G[q,k]$) were higher (.46-.85, M = .68, SD = 0.11). Thus, these findings lend partial support to H1. That is, H1 is supported only for role-play performance ratings averaged across multiple assessors in ten role-plays (see Table 1).

Table 1

Single-rater (G[q,1]) and Interrater Reliabilities for Role-Play Performance Ratings averaged across all Assessors (G[q,k])

Role-play	$\hat{k}$	*q*-multiplier *G*(*q*,*1*)	*q*-multiplier *G*(*q*,*k*)	*G*(*q*,*1*)	*G*(*q*,*k*)
1	4	.79	.04	.22	.63
2	3	.73	.07	.30	.63
3	4	.80	.05	.37	.73
4	3	.75	.09	.48	.75
5	3	.73	.06	.64	.85
6	3	.73	.07	.43	.71
7	3	.73	.07	.30	.61
8	3	.73	.07	.26	.57
9	3	.75	.09	.23	.50
10	3	.75	.09	.45	.74
11	3	.74	.07	.48	.77
12	3	.75	.08	.29	.62
13	3	.73	.06	.47	.77
14	4	.78	.03	.53	.83
15	4	.80	.05	.42	.79
16	3	.75	.08	.18	.46
17	3	.75	.08	.41	.70
18	3	.75	.08	.34	.61

Note. $\hat{k}$ = harmonic mean number of assessors per participant. The *q*-multiplier scales the amount of variance that can be attributed to assessor main effects. With increasing overlap between sets of assessors that rate each participant, *q* approaches 0. With decreasing overlap between sets of assessors that rate each participant, q approaches $1/\hat{k}$ (Putka et al., 2008).

An index of interrater reliability is insightful but in multiple speed assessments 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 (leniency/stringency) effects, etc. However, there are many more systematic and unsystematic sources of variance that are insightful to better understand assessors' role-play performance ratings. To decompose assessors' ratings into these different sources of variance, we conducted generalizability theory analyses (e.g., Brennan, 2001; Vispoel, Morris, & Kilinc, 2018). We fitted a linear random effects model with restricted maximum likelihood estimators (Putka et al., 2008) by using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) for R (R Core Team, 2015). We modeled participants, role-plays, and assessors as crossed-

random factors to examine the relative contribution of seven sources of variance to observed variance in assessors' role-play performance ratings.

Table 2 presents the substantive meaning of all seven variance components as well as the percentage of expected total variance in role-play performance ratings explained by the various variance components, the percentage of expected between-participant variance explained, and the percentage of expected reliable or unreliable variance that is explained. Results showed that the amount of reliable variance in between-participant variance of assessors' role-play performance ratings equaled 37 %. The largest source of reliable variance was attributable to participant × role-play interaction effects, explaining 22.6 % of between participant variance. This is in line with individuals varying their behavior across situations with different demands (Dalal, Bhave, & Fiset, 2014; Fournier, Moskowitz, & Zuroff, 2008; Shoda, Mischel, & Wright, 1994). It also shows that participants' performance differences across role-plays explained a larger portion of reliable variance than the participant main effect (differences in general performance across role-plays), that explained 14.7 % of between participant variance.

Table 2

Variance Component Estimates (%) for Role-Play Performance Ratings

Variance component	Substantive meaning	Total variance	Between-participant variance	Between-participant subtotal
Reliable variance				
σ² participant	Independent of assessors and role-plays, some participants receive higher ratings than others	13.5	14.7	39.4
σ² participant × role-play	Independent of assessors, some participants receive higher ratings in some role-plays than others	20.8	22.6	60.6
Subtotal		34.4	37.3	
Unreliable variance				
σ² assessor*	Independent of participants or role-plays, some assessors give higher ratings than others	8.8	9.5	15.2
σ² assessor × participant	Independent of role-plays, some assessors assign higher ratings to some participants than others	5.9	6.4	10.3
σ² assessor × role-play*	Independent of participants, some assessors give higher ratings in some role-plays than others	6.6	7.2	11.5
σ² 3-way interaction + residual	Depending on role-plays, some participants receive higher ratings from some assessors	36.5	39.6	63.1
Subtotal		57.8	62.7	
Other Components				
σ² role-play	Independent of participants or assessors, ratings are higher in some role-plays than in others	7.8		
Estimated *G*(*q*,*1*)			0.37	

Note. In line with prior studies (Jackson, Michaelides, Dewberry, & Kim, 2016; Putka & Hoffman, 2013), we intended to generalize role-play performance ratings across assessors and defined reliability in relative terms (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Therefore, we regarded sources of variance as reliable if they contributed to the similarity in participants' relative position compared to other participants based upon role-play performance ratings from different assessors. Conversely, sources of variance were considered as unreliable if they contributed to differences in participants' relative position compared to other participants based upon role-play performance ratings from different assessors (LeBreton & Senter, 2007; Putka & Sackett, 2010). Between-participant variance refers to the relative amount of variance that is explained in the total of reliable and unreliable variance by the respective variance component. Between-participant subtotal refers to the relative amount of variance that is explained in the total reliable or unreliable variance by the respective component. Estimated *G*(*q*,*1*) = reliable variance subtotal/(reliable variance subtotal + unreliable variance subtotal) = expected single-rater reliability for any role-play performance ratings.

* Components mirror unreliable variance because participants were not fully crossed with assessors. In line with prior studies (Jackson et al., 2016; Putka & Hoffman, 2013), these variance components were rescaled via the *q*-multiplier for *G*(*q*,*1*) of .98 because of the ill-structured measurement design.

At a practical level, the relative contribution of reliable and unreliable variance components to multiple speed assessment ratings from a generalizability theory analysis is useful because it informs how multiple speed assessments can be designed more efficiently to improve reliability. Given that multiple speed assessments require substantial personnel and time resources, organizations are interested to know how many simulations or assessors are necessary to have adequate reliability. Therefore, to investigate how to design such an "optimal" multiple speed assessment in terms of reliability, we used the estimated variance components from the linear random effects model to run a decision study. Decision studies follow the logic of the Spearman-Brown prophecy formula and show how reliability changes if the observations per facets are varied (Brennan, 2001; Vispoel et al., 2018). Here they can reveal how reliability changes by varying the number of role-plays/assessors per role-play.

In this analysis, we varied the number of assessors per role-play from one to three and the number of role-plays from one to twenty and calculated a generalizability coefficient for all these assessor-role-play combinations. Figure 1 summarizes the results: Interestingly, acceptable reliabilities (generalizability coefficients > .70)[11] resulted if role-play performance ratings from at least two independent assessors per role-play were aggregated across at least fourteen role-plays or if role-play performance ratings from three independent assessors per role-play were aggregated across at least nine role-plays.

[11] Again, we urge caution regarding general thresholds of reliability, but follow the rule of thumb to interpret reliabilities of > .70 as acceptable for newly developed measures (see, for example, LeBreton & Senter, 2007).

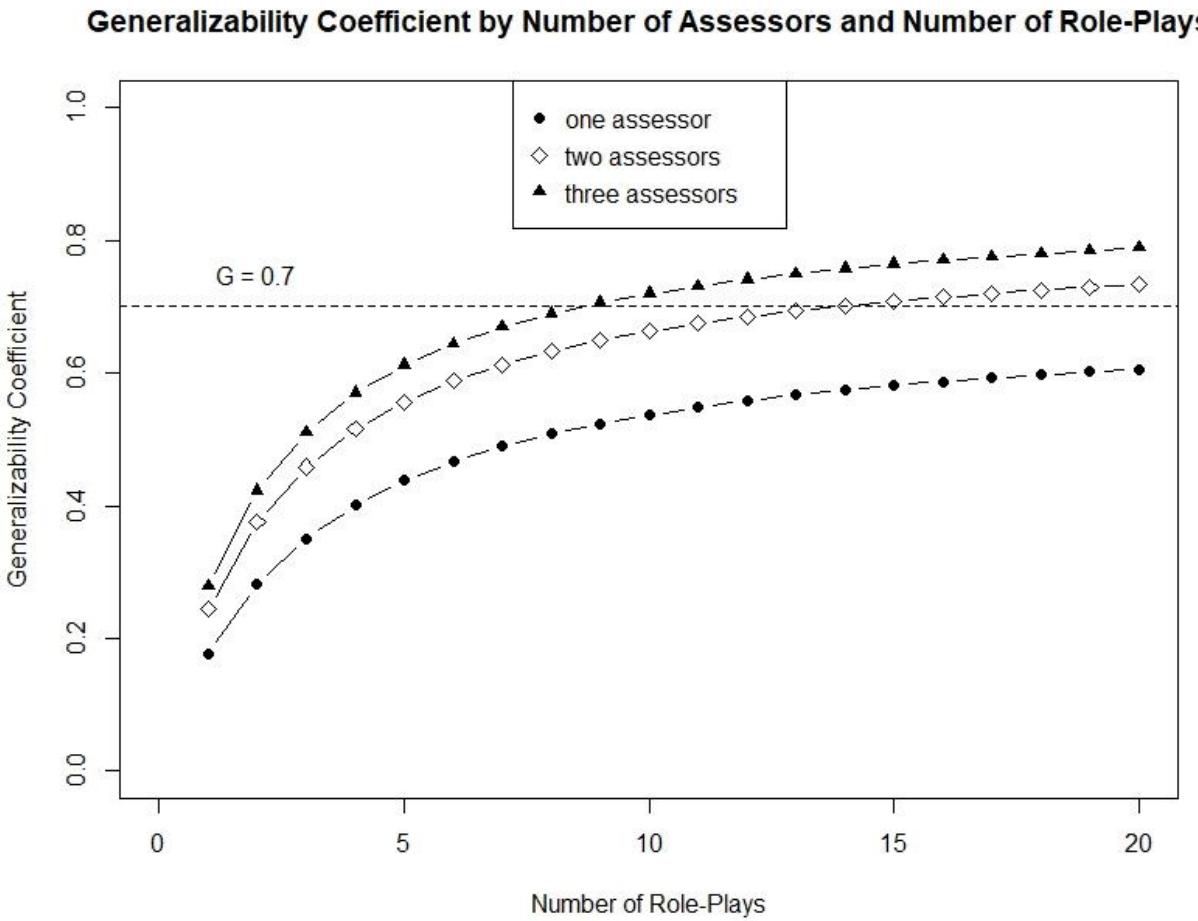

Figure 1. Generalizability Coefficient Depending on Number of Role-Plays and Assessors per Role-Play. Generalizability Coefficient =

$$\frac{\sigma^2 \text{participant}}{\sigma^2 \text{participant} + \left(\frac{\sigma^2\ \text{participant} \times \text{assessor}}{\text{number of assessors}} + \frac{\sigma^2\ \text{participant}\ \times\ \text{role} - \text{play}}{\text{number of role} - \text{plays}} + \frac{\sigma^2\ \text{participant}\ \times\ \text{assessor}\ \times\ \text{role} - \text{play} + \text{residual}}{\text{number of assessors}\ \times\ \text{number of role} - \text{plays}}\right)}$$

Do Ratings in Short and Fast-paced Simulations Capture Meaningful Information about Participants' Cognitive Ability and Personality?

Our next set of hypotheses (H2-H5) proposed that ratings in short and fast-paced simulations allow capturing substantive and meaningful information about participants' cognitive ability, extraversion, agreeableness, and openness. Thus, testing these hypotheses provides answers to the question what kind of fundamental individual differences constructs are represented in ratings in a multiple speed assessment about leadership.

To investigate these hypotheses, we examined the correlations between the composite measure of cognitive ability as well as self-reports of extraversion, agreeableness, and openness on the one hand with role-play performance ratings and overall multiple speed assessment performance on the other hand. As could be expected in light of the interrater reliability results, cognitive ability and personality measures showed highly variable correlations with role-play performance ratings from different role-plays (see Table 3) which resulted in low average correlations (cognitive ability: M = .18, range = .00-.35; extraversion: M = .19, range = .01-.37; agreeableness: M = .12, range = -.02-.31; openness: M = .06, range = -.13-.25).

However, when we changed the level of analysis and examined correlations to overall multiple speed assessment performance, there is support for our hypotheses (see Table 3). In line with H2-H4, overall multiple speed assessment performance correlated positively with participants' cognitive ability (r = .39, p < .001), extraversion (r = .38, p < .001), and agreeableness (r = .24, p = .02). Contrary to H5, openness did not significantly correlate with overall multiple speed assessment performance (r = .11, p = .305). As a possible explanation, each role-player observed and evaluated participants in only one role-play. To gain insights about participants' openness, role-players/assessors might need to observe participants in multiple role-plays to assess whether they approached the different role-plays from different

perspectives or developed different, innovative problem solutions across multiple role-plays. Emotional stability and conscientiousness also did not relate to overall multiple speed assessment performance (*p*s > .05, see Table 4).

Table 3

Relations between Role-play Performance Ratings, Overall Multiple Speed Assessment Performance, Cognitive Ability, Personality, and Criterion Performance

Role-play	*n*	*M*	*SD*	Cognitive ability	Extraversion	Agreeableness	Openness	Instructor rated performance	Peer rated performance
1	95	6.79	1.40	.31**	.20	.01	.02	.24*	.23*
2	94	6.42	1.65	.17	.15	.27**	-.13	.12	.05
3	96	5.81	1.42	.19	.30**	.16	.25*	.05	.11
4	92	5.16	1.55	.15	.27**	.08	.17	.19	.16
5	96	5.41	1.84	.26*	.17	-.01	-.01	.40**	.14
6	95	4.89	1.76	.20	.05	-.01	-.02	.27**	.25*
7	92	5.86	1.50	.16	.11	-.02	.09	.09	.08
8	94	6.21	1.46	.12	.16	.11	.02	.25*	-.02
9	94	5.36	1.33	.00	.01	.03	-.02	.03	.07
10	67	5.41	1.55	.06	.27*	.31*	.02	.23	.36**
11	94	5.68	1.38	.17	.28**	.26*	.17	.21*	.07
12	94	5.45	1.35	.08	.12	.10	.25*	.32**	.13
13	88	4.12	1.47	.13	.21	.09	.14	.16	.13
14	92	4.91	1.91	.26*	.37**	.12	.03	.23*	.06
15	91	5.41	1.35	.25*	.21*	.26*	-.01	.35**	.17
16	90	5.47	1.24	.30**	.17	.15	-.11	.15	.21*
17	95	5.02	1.40	.11	.10	.19	.06	.30**	.12
18	93	5.72	1.52	.35**	.23*	.10	.08	.24*	.15
***M*[a]**	**96**	**5.51**	**0.73**	**.39****	**.38****	**.24***	**.11**	**.44****	**.27****

Note. * $p < .05$, ** $p < .01$, [a] overall multiple speed assessment performance

Table 4

Means, Standard Deviations, and Intercorrelations between Study Variables

Variable	*N*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
Controls														
1 Gender	96	-	-	-										
2 Age	96	23.63	1.85	-.15	-									
3 Test motivation	49	3.96	0.50	-.08	.15	-								
Predictors														
4 Cognitive ability	95	0.00	0.78	.20	-.43**	.19	-							
5 Extraversion	95	90.82	12.54	.03	-.17	.18	.19	-						
6 Agreeableness	95	92.82	13.54	.04	-.27**	.25	.21*	.53**	-					
7 Openness	95	83.76	14.18	-.06	.02	.13	.08	.51**	.38**	-				
8 Emotional stability	95	86.55	12.39	-.22*	.01	.11	.05	.51**	.30**	.38**	-			
9 Conscientiousness	95	92.60	10.08	-.07	.06	-.11	.03	.07	.06	.27**	.13	-		
10 Overall multiple speed assessment performance	96	5.51	0.73	.00	-.38**	.19	.39**	.38**	.24*	.11	.05	-.06	-	
Criteria														
11 Instructor rated performance	95	56.68	19.92	-.15	-.25*	.10	.21*	.01	.15	-.07	-.09	-.04	.44**	-
12 Peer rated performance	94	2.86	0.68	.10	-.06	-.08	.14	-.05	.06	-.08	-.07	-.26*	.27**	.27**

Note. Gender is coded as follows: *male = 1, female = 2*. * $p < .05$, ** $p < .01$

Do Ratings in Short and Fast-paced Simulations Predict Criterion Performance?

H6 proposed that ratings based upon multiple short leadership simulations would predict criterion performance. The investigation of this hypothesis is crucial to find out whether multiple speed assessments can indeed validly predict criterion performance.

We investigated this hypothesis for role-play performance ratings and overall multiple speed assessment performance. Similar to our prior results, role-play performance ratings showed highly variable correlations to criterion performance which resulted in rather low average correlations (instructor-rated performance: M = .21, range = .03-.40; peer-rated performance: M = .14, range = -.02-.36, see Table 3). Again, the picture changes when we examined correlations between overall multiple speed assessment performance and criteria: In line with our hypothesis, overall multiple speed assessment performance significantly predicted performance rated by instructors (r = .44, p < .001) and peers (r = .27, p = .009). Therefore, H6 was supported for the overall multiple speed assessment.

Given that overall multiple speed assessment performance predicted criterion performance, we further investigated whether it adds incremental validity beyond traditional predictors. For organizations, these analyses shed light on the added value of multiple speed assessments: Do they pay off vis-à-vis instruments that tap into similar construct domains such as measures of cognitive ability, extraversion, agreeableness, and openness? To investigate whether overall multiple speed assessment performance ratings add incremental validity to predict criterion performance beyond traditional predictors, we ran separate multiple regression analyses to predict instructor and peer rated performance. In line with a theory-driven approach (Arthur & Villado, 2008), we included only predictors in our regressions that tap into similar construct domains as the multiple speed assessment. In model 1, we included gender and age as control variables. In model 2, we included measures of cognitive ability, extraversion, agreeableness, and openness. In block 3, overall multiple

speed assessment performance was added. Results showed that overall multiple speed assessment performance explained additional 14 % of variance in instructor rated performance ($p < .001$) and additional 9 % of variance in peer rated performance ($p = .005$) above all other predictors (see Table 5).

Table 5

Multiple Regressions to Predict Instructor and Peer Rated Performance

	Instructor rated performance								Peer rated performance							
Predictors	*b*	β	R^2	Δ R^2	*F*	*df*	*p*	Sig. *F* change	*b*	β	R^2	Δ R^2	*F*	*df*	*p*	Sig. *F* change
Model 1			.10	.10	4.85	2,91	.01	.01			.01	.01	0.51	2,91	.602	.602
Gender	-7.50	-.19							0.12	.09						
Age	-2.98**	-.28							-0.02	-.04						
Model 2			.14	.04	2.37	6,87	.036	.353			.04	.03	0.62	6,87	.717	.613
Gender	-8.72*	-.22							0.09	.07						
Age	-1.95	-.18							0.01	.03						
Cognitive ability	4.14	.16							0.12	.14						
Extraversion	-0.12	-.08							-0.01	-.10						
Agreeableness	0.23	.16							0.01	.11						
Openness	-0.17	-.12							0.00	-.08						
Model 3			.28	.14	4.71	7,86	< .001	< .001			.13	.09	1.78	7,86	.102	.005
Gender	-6.36	-.16							0.15	.11						
Age	-0.78	-.07							0.04	.11						
Cognitive ability	1.12	.04							0.04	.04						
Extraversion	-0.36	-.22							-0.01	-.21						
Agreeableness	0.25	.17							0.01	.12						
Openness	-0.10	-.07							0.00	-.04						
Overall multiple speed assessment performance	12.22**	.44							0.33**	.35						

Note. $N = 94$. Gender is coded as follows: *male = 1, female = 2*. * $p < .05$, ** $p < .01$

Discussion

Recently, multiple speed assessments have made rapid inroads into the selection arena. So far, however, the conceptual underpinning and empirical evidence related to these short, fast-paced assessment approaches has been lacking. Therefore, this study integrates this novel assessment approach into the minimal acquaintance/thin slices paradigm and is the first to present evidence-based answers to the key question: "Is there evidence for the reliability and validity of ratings made in multiple speed assessments?"

Main Conclusions

Our answer to our title question can be summarized as a nuanced "yes, but only under specific conditions". On one hand, results for individual speed role-plays were consistently disappointing. Single-rater reliabilities were low to at best moderate and echoed the reliabilities of judges in research in the minimal acquaintance/thin slices paradigm (Ambady et al., 2000; Connelly & Ones, 2010; Kenny, Albright, Malloy, & Kashy, 1994). In addition, role-play performance ratings showed variable and weak relations to participants' cognitive ability and relevant personality traits (extraversion, agreeableness, openness). Finally, evidence of the predictive validity of role-play performance ratings in single, short simulations was equally variable, showing an inconsistent pattern of relations with instructor rated and peer rated performance.

Thus, although role-players and assessors followed trainings, used behavioral observation aids, and relied upon behavior elicitation and evaluation via standardized prompts, there appears to remain too much error variance (idiosyncratic assessor effects) in their ratings for obtaining acceptable reliability and consistent relationships with relevant constructs and criteria. This was also evidenced by the large percentage of unreliable variance in their ratings. For decision-making purposes in selection contexts, higher amounts of reliable variance are desirable.

On the other hand, the picture of our results changed when ratings were aggregated across multiple assessors per simulation. In this case, composite interrater reliabilities were moderate to high. Similarly, when ratings were aggregated across all role-plays, we found substantial relations between overall multiple speed assessment performance and cognitive ability, extraversion, as well as agreeableness. Our predictive validity results mirrored this. When ratings were aggregated across all role-plays, the overall multiple speed assessment performance significantly predicted peer-rated as well as instructor-rated performance seven months after the multiple speed assessment. Further, this overall multiple speed assessment performance added a large amount of incremental variance of 14 % and 9 % in instructor rated and peer-rated performance, respectively, beyond relevant personality traits and cognitive ability.

Implications for Theory

In terms of theoretical implications, these results attest to the key role of the principle of aggregation in multiple speed assessments (Epstein, 1979). If ratings are aggregated across many different situations, reliability increases because such an aggregation process maximizes the amount of systematic variance in ratings that is shared across situations (Epstein, 1979; Kuncel & Sackett, 2014). In a similar way, aggregating across multiple different assessors reduces the relative amount of assessor-specific idiosyncrasies (Eisenkraft, 2013). Thus, although ratings from single assessors in single, short and fast-paced simulations show low reliability, aggregating their ratings across simulations and assessors in multiple speed assessments thus generates a reliable and valid indicator of performance.

As another theoretical implication, this study provides information about what constitutes “good information” for assessors when evaluating performance in multiple speed assessments. Inspired by the minimal acquaintance/thin slices paradigm, multiple short, but qualitatively different situations are required to make valid predictions about criterion

performance. When the various simulations invoke substantively different situational demands and do not serve as alternate measures of one another (see the large participants × role-play interaction), multiple speed assessments permit obtaining a broad domain coverage. Thus, by confronting participants with multiple, short simulations that elicit behavior relevant for a large set of qualitatively different parts of the criterion domain, multiple speed assessments seem to optimize the behavioral sampling and point-to-point correspondence between predictor and criterion (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Directions for Future Research

A first intriguing question is whether one 1-hour role-play or two 30-minute role-plays would obtain similar validities than the aggregated ratings across the 18 different 3-minute role-plays. Although this question awaits empirical examination, 18 different role-plays might have a conceptual advantage because they provide a much broader coverage of the criterion domain (see above). Utility analyses are needed to verify this further. Such utility analyses might also complement our incremental validity analyses and show whether and under which circumstances these validity benefits translate into positive return on investment.

Second, as revealed by our analyses, a participant × role-play interaction effect explained the majority of reliable variance in assessor ratings, which indicates that participants differ in their performance across the role-plays. This result echoes empirical evidence for intraindividual variability in behavior and performance across situations (e.g., Dalal et al., 2014; Fleeson, 2001; Fournier et al., 2008; Gibbons & Rupp, 2009; Judge, Simon, Hurst, & Kelley, 2014; Minbashian & Luppino, 2014; Moskowitz & Zuroff, 2004; Smith, Shoda, Cumming, & Smoll, 2009). Such intraindividual variability is not only due to random error, but suggests that individuals systematically construe situations in different ways, which leads to different behavioral choices and performance across situations (Lance, 2008; Mischel & Shoda, 1995). Hence, future studies should examine whether multiple speed assessments

can shed light onto people's variability across situations (Baard, Rench, & Kozlowski, 2014; Dalal et al., 2014; DeRue, Ashford, & Myers, 2012; Jundt, Shoss, & Huang, 2015). By using taxonomies for constructing the situations (Rauthmann et al., 2014), situation-behavior linkages might also be uncovered. Recently, Lievens et al. (2018) assessed such intraindividual variability via variability of responses across items of a Situational Judgment Test (SJT). Intraindividual variability across SJT items converged with self-ratings of functional flexibility, added incremental validity to the prediction of performance beyond individuals' mean scores across situations, and predicted actual intraindividual variability across ten days. In a similar way, intraindividual variability could be efficiently assessed in a multiple speed assessment approach across the multiple simulations.

Third, future research should examine the developmental applications of multiple speed assessments. For example, a coach could observe how participants behave across several simulations and provide developmental feedback. Then, the coach could check whether the developmental feedback is successfully applied in the following simulations.

Limitations

First, the multiple speed assessments in this study included only role-play simulations. In principle, other types of simulations can be implemented into multiple speed assessments, such as short fact-finding-exercises, presentations, etc. The multiple speed assessment was also set up in a "brick-and-mortar" fashion. We cannot extend our conclusions to other speed assessment formats such as web-based SJTs (also known as constructed reponse multimedia tests; e.g., Cucina, Su, Busciglio, Harris Thomas, & Thompson Peyton, 2015; Herde & Lievens, 2018; Lievens, De Corte, & Westerveld, 2015).

Second, domains other than leadership might be sampled in multiple speed assessments, such as the interpersonal, integrity, or decision-making domains. Multiple speed assessments might also focus on specific job families such as business consultants, customer

service personnel, or call center agents. Extrapolating on our results, we expect that multiple speed assessments will produce good validity results at the aggregate level as long as the multiple simulations sample and cover a diverse set of different situations related to the targeted performance domain so that the point-to-point correspondence between predictor and criterion domain is ensured (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Implications for Practice

In today's selection practice, multiple speed assessments exist in various formats. Our results send a strong warning to organizations that equate speed assessment with one short simulation or use a limited number of speed assessments. A key conclusion is that selection practices should not be degraded to such single, short simulations with ratings of single assessors. Conversely, a large set of short and different simulations need to be deployed and ratings need to be aggregated across simulations and assessors to be reliable and show consistent validity evidence. In this study, speed assessments "work" only if ratings are aggregated across multiple simulations and assessors. Only then, speed assessments can provide reliable and valid insights about performance that add incremental validity above cognitive ability and personality. Given these results, organizations may decide to adopt multiple speed assessments – under these conditions – in their selection systems to complement existing methods, although other considerations (subgroup differences, costs, etc.) should also be considered.

This study also yields guidelines in terms of the number of assessors and simulations needed to obtain reliable speed assessments of performance. To deploy available human resources and logistics more efficiently, our decision study revealed that aggregating performance ratings across fourteen short role-plays with each two independent assessors or across nine role-plays with each three independent assessors produced reliabilities of at least $G = .70$.

Finally, we emphasize caution in following up on these recommendations. This is only the first empirical study that investigates the reliability and validity of short and fast-paced simulations. Clearly, further replications and studies are needed.

Conclusion

This study investigated the theoretical underpinnings and empirical evidence behind the recent trend of speed and flash assessments. We found that assessor ratings from short and fast-paced simulations were reliable and valid indicators of performance, but *only if* these ratings were aggregated across a large set of diverse situations and multiple assessors. In other words, there do not seem to exist shortcuts in psychometrics: Speed assessment approaches truly need to be conceptualized as *multiple* speed assessments.

References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55*, 387–395. http://dx.doi.org/10.1037/0022-3514.55.3.387

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

Arthur, 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, MA: Cambridge University Press.

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

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

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

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

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

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

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

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

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., 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

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

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*, 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

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

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

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal*, *50*, 327–347. https://doi.org/10.5465/amj.2007.24634438

Haque, U. (2016, March 1). Our economy is obsessed with efficiency and terrible at everything else. *Harvard Business Review*. Retrieved from

https://hbr.org/2016/03/our-economy-is-obsessed-with-efficiency-and-terrible-at-everything-else

Herde, C. N., & Lievens, F. (2018). Multiple Speed Assessments: Theory, practice, & research evidence. *European Journal of Psychological Assessment,* Advance online article. https://doi.org/10.1027/1015-5759/a000512

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. http://dx.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

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

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology*, *86*, 984–996. http://dx.doi.org/10.1037/0021-9010.86.5.984

Judge, T. A., Simon, L. S., Hurst, C., & Kelley, K. (2014). What I experienced yesterday is who I am today: Relationship of work motivations and behaviors to within-individual variation in the five-factor model of personality. *Journal of Applied Psychology, 99,* 199–221. https://doi.org/10.1037/a0034485

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

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690. http://dx.doi.org/10.1037/0021-9010.74.4.657

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116*, 245–258. http://dx.doi.org/10.1037/0033-2909.116.2.245

Kenny, D. A., & West, T. V. (2008). Zero acquaintance: Definitions, statistical model, findings, and process. In J. Skowronski & N. Ambady (Eds.), *First impressions* (pp. 129–146). New York, NY: Guilford.

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

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

LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642

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. (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. (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. (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., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, *91*, 247–258. https://doi.org/10.1037/0021-9010.91.2.247

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., & Reeve, C. L. (2012). Where I–O Psychology should really (re)start its investigation of intelligence constructs and their measurement. *Industrial and Organizational Psychology*, *5*, 153–158. https://doi.org/10.1111/j.1754-9434.2012.01421.x

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

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.

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

Mendas. (2018, October 19). Mini Assessment Centres. Retrieved from http://mendas.com/what-we-do/tools/mini-assessment-centres.php

Minbashian, A., & Luppino, D. (2014). Short-term and long-term within-person variability in performance: An integrative model. *Journal of Applied Psychology*, *99*, 898–914. https://doi.org/10.1037/a0037402

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

Moskowitz, D. S., & Zuroff, D. C. (2004). Flux, pulse, and spin: Dynamic additions to the personality lexicon. *Journal of Personality and Social Psychology*, *86*, 880–893. https://doi.org/10.1037/0022-3514.86.6.880

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

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., … 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

Mussel, P. (2013). Intellect: A theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology*, *104*, 885–906. https://doi.org/10.1037/a0031918

Muthén, L. K., & Muthén, B. O. (1998). *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

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

Pinsight (2019, June 19). Shorter & Faster: Recent Trend in Assessment Centers. Retrieved from https://www.pinsight.com/resources/2019/3/5/recent-trend-in-assessment-centers

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

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9–49). New York, NY: Routledge.

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

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

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

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

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

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

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

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

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

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372-376. http://dx.doi.org/10.1037/h0026244

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

Yukl, G. (1999). An evaluative essay on current conceptions of effective leadership. *European Journal of Work and Organizational Psychology*, *8*, 33–48. https://doi.org/10.1080/135943299398429

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

CHAPTER 5: A CLOSER LOOK AT INTRAINDIVIDUAL VARIABILITY IN INTERPERSONAL BEHAVIOR AND INTERPERSONAL DYNAMICS IN HIGH-FIDELITY SIMULATIONS[12]

High-fidelity simulations such as assessment center role-plays offer a lot of opportunities to gain insights into people's behavioral repertoire in job-related situations because these simulations allow to observe how participants interact with other human actors. However, in the past, this potential was not fully exploited because interpersonal behavior and dynamics between participants and other human actors were typically described with single-point estimates. Thus, we still lack information about how interpersonal behavior of human actors in high-fidelity simulations varies *within* simulations and how possible intraindividual variability in interpersonal behavior translates into interpersonal dynamics between human actors. Building upon interpersonal theory, this study applies a continuous assessment of interpersonal behavior and dynamics between 96 participants and role-players across four role-plays. Results showed that participants and role-players show significant intraindividual variability in interpersonal behavior within all four role-plays. Further, in line with interpersonal theory, participants and role-players adapted their interpersonal behavior to each other in line with the principles of complementarity. That is, more affiliative behavior elicited more affiliative behavior and more dominant behavior elicited more submissive behavior. However, this pattern was only consistently found if interpersonal behavior was analyzed at the continuous behavioral level. Complementarity across the four role-plays predicted instructor and peer ratings of interpersonal adaptability and task-related performance. Implications for theory and practice are discussed.

[12] Portions of this paper were presented in: Herde, C.N., & Lievens, F. (2018, April). Moment-to-moment Interpersonal Behavior in AC Exercises: Some Unexploited Potential? In P.V. Ingold & B. J. Hoffman (Chair*), The AC, You, and Me: Insights From an Interpersonal Perspective.* Symposium conducted at the 33rd Annual Conference of the Society of Industrial and Organizational Psychology (SIOP), Chicago, IL, USA.

Introduction

High-fidelity simulations are popular procedures to assess participants' behavioral repertoire in assessment and development (Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Gebert, 2003; Krause, Rossberger, Dowdeswell, Venter, & Joubert, 2011; Krause & Thornton, 2009; Thornton & Rupp, 2006). Examples of high-fidelity simulations are assessment center exercises such as role-plays or leaderless group discussions. One advantage of these simulations is the potential to observe actual behavior of participants. Especially, high-fidelity simulations enable to gain insights about how participants interact with other human actors, such as other participants or role-players. This provides an excellent opportunity to observe interpersonal dynamics between participants and other human actors as they unfold across time (Lievens & Klimoski, 2001).

Surprisingly, however, we still lack a sound understanding of such interpersonal dynamics. That is, in a comprehensive review of past research on high-fidelity simulations, Kleinmann and Ingold (2019) pointed out that the vast majority of studies investigated phenomena related to either participants or assessors in isolation[13]. Strikingly, almost 20 years ago, Lievens and Klimoski (2001) drew the same conclusion, showing that studies on interpersonal *dynamics* between participants and assessors are scarce. As a rare exception, Oliver, Hausdorf, Lievens, and Conlon (2016) investigated interpersonal dynamics between participants and role-players. They demonstrated that both interpersonal behavior of role-players and task-related situational demands influence participants' interpersonal behavior. Further, they examined what kind of interpersonal behavior of participants benefits

[13] We acknowledge studies that investigated phenomena like impression management (e.g., McFarland, Ryan, & Kriska, 2003; McFarland, Yun, Harold, Viera, & Moore, 2005), participants' ability to identify criteria (e.g., Jansen, Lievens, & Kleinmann, 2011; Jansen et al., 2013; Kleinmann, 1993; Kleinmann et al., 2011), or how standardized role-player actions activate participants' trait relevant behavior (Lievens, Schollaert, & Keen, 2015; Schollaert & Lievens, 2011, 2012). Although these studies might capture dynamics between participants and assessors/role-players, they do not cover *fundamental interpersonal* dynamics between participants and assessors.

performance ratings across various role-plays as well as different interpersonal and task demands.

Although Oliver et al. (2016) provided first insights into interpersonal dynamics in high-fidelity simulations, this study applied only single-point estimates of interpersonal behavior. Single-point estimates of interpersonal behavior, however, do not account for the substantial intraindividual variability *within* interpersonal interactions that has been found in social and clinical psychological research (e.g., Markey, Lowmaster, & Eichler, 2010; Sadler, Ethier, Gunn, Duong, & Woody, 2009; Tracey, 2004). Further, research has shown that the nature and size of interpersonal dynamics as well as their relations to outcomes differ by the level of analyses (Tracey, 2004). Accurate insights about the true nature of interpersonal dynamics in high-fidelity simulations can thus only be unveiled via *continuous* assessments that capture intraindividual variability in interpersonal behavior (see also Gabriel, Diefendorff, Bennett, & Sloan, 2017; Jebb & Tay, 2017).

Therefore, many questions about the interpersonal dynamics in high-fidelity simulations still remain that can only be answered with continuous assessments of interpersonal behavior. For example, do participants and other human actors like role-players show intraindividual variability in interpersonal behavior within simulations? Do momentary interpersonal dynamics between human actors in simulations follow any systematic pattern? Do momentary interpersonal dynamics in simulations predict performance ratings in these simulations? Do momentary interpersonal dynamics in simulations even predict job-related performance outside of the simulation context?

To answer these pressing questions, this study investigates interpersonal dynamics at the continuous behavioral level in high-fidelity simulations as well as their relation to ratings of simulation performance and job-related performance criteria. We draw from Interpersonal Theory and the principles of complementarity (e.g., Carson, 1969; Kiesler, 1983) to derive

hypotheses about interpersonal behavior and dynamics within high-fidelity simulations between participants and role-players. Afterwards, we (a) investigate interpersonal dynamics between participants and role-players at the overall level, (b) examine whether there is intraindividual variability in interpersonal behavior within simulations, (c) investigate interpersonal dynamics between participants and role-players at the momentary level, and (d) examine whether interpersonal dynamics within simulations predict performance ratings in simulations as well as job-related performance criteria.

This paper contributes to a better understanding of the interpersonal micro cosmos and dynamics in high-fidelity simulations in various ways. At a theoretical level, we use Interpersonal Theory to introduce the concepts of variability and complementarity in the literature on high fidelity simulations. Accordingly, we better connect this literature to Interpersonal Theory. Further, we advance Interpersonal Theory by testing whether its assumptions are still valid in stronger situations (i.e., assessments that are often used in selection and development settings). Moreover, we outline the importance and added value of continuous assessments of interpersonal behavior and dynamics in high-fidelity simulations. At a practical level, our results provide implications for practitioners to utilize continuous assessments of interpersonal behavior and dynamics. Finally, we propose that interpersonal dynamics observed in high-fidelity simulations might provide a new angle to assess interpersonal adaptability for assessment, selection and development purposes.

Study Background

Interpersonal Theory and the Interpersonal Circumplex Model

To better understand momentary interpersonal behavior and dynamics within high-fidelity simulations, we draw upon Interpersonal Theory and the Interpersonal Circumplex Model (e.g., Carson, 1969; Kiesler, 1983) and apply it to high-fidelity simulations. Interpersonal Theory and the Interpersonal Circumplex Model propose that interpersonal

behavior can be sufficiently described in terms of a Cartesian plane that builds upon two orthogonal and bipolar axes. The x-axis of the Interpersonal Circumplex Model describes the dimension of affiliation. This dimension spans up a continuum between behavior expressing warmth or friendliness on the positive or right part of the continuum, and behavior expressing coldness or hostility on the negative or left part of the continuum. The y-axis of the Interpersonal Circumplex Model describes the dimension of dominance. This dimension spans up a continuum between behavior expressing dominance or leading others on the positive or upper part of the continuum, and behavior expressing submissiveness or following others on the negative or lower part of the continuum (e.g., Carson, 1969; Kiesler, 1983).

Substantial empirical evidence lends support to the Interpersonal Circumplex Model. There is evidence that interpersonal behavior can indeed be sufficiently described via the two orthogonal dimensions of affiliation and dominance, and that the organization of different forms of interpersonal behavior along these two dimensions follows a circular structure (e.g., Ansell, Kurtz, & Markey, 2008; Leising & Bleidorn, 2011; Markey, Funder, & Ozer, 2003; Moskowitz, 1994; Strong et al., 1988; Tracey, 1994; Tracey, Ryan, & Jaschik-Herman, 2001).

The beauty of the Interpersonal Circumplex Model is due to its parsimony to describe the interpersonal behavior of two individuals who are interacting with each other with the very same, two dimensions. So, the Interpersonal Circle serves to describe both an actor's behavior as well as the interpersonal situation (= the interaction partner) that the actor is facing with the dimensions of affiliation and dominance. Thereby, Interpersonal Theory aims to describe interpersonal behavior of two actors who interact with each other, but also to predict and explain interpersonal dynamics between these two actors (e.g., Carson, 1969; Kiesler, 1983).

The Principles of Complementarity

Interpersonal Theory proposes that interpersonal dynamics between two human actors follow the two principles of complementarity (e.g., Carson, 1969; Kiesler, 1983). The two principles of complementarity assume that the two interaction partners' levels of displayed affiliation and dominance will mutually influence each other. The first principle is correspondence in affiliation. Take, for example, an interaction between two people called Ann and Lisa. Interpersonal Theory assumes that the more friendly and warm behavior Ann is displaying overall, the more likely will Lisa correspond with more friendly and warm behavior overall. The second principle of complementarity is reciprocity in dominance. Interpersonal Theory assumes that the more dominant and leading behavior Ann is displaying overall, the more likely will Lisa respond with submissive and following behavior overall. Importantly, however, Interpersonal Theory assumes no interactions between the two dimensions of affiliation and dominance. So, the theory states that the two principles of complementarity coexist as two independent phenomena (Carson, 1969; Kiesler, 1983; see also Sadler & Woody, 2003).

Empirical research supports that dyadic interactions follow the two principles of complementarity (for an overview, see Sadler, Ethier, & Woody, 2011). Across a diverse set of studies, individuals showed corresponding levels of affiliation and reciprocal levels of dominance. For instance, complementarity was established in lab settings involving unstructured discussions and negotiations (Markey et al., 2010; Markey et al., 2003; Sadler et al., 2009; Sadler & Woody, 2003; Strong et al., 1988), collaborative tasks/games (Markey et al., 2003), or murder mysteries (Locke & Sadler, 2007). In addition, complementarity was found outside of lab settings, such as between romantic partners or close friends (Tracey et al., 2001), therapists and clients (Tracey, Albright, & Sherry, 1999), college roommates (Ansell et al., 2008; Markey & Kurtz, 2006), and in naturally occurring interpersonal

interactions, including interactions at work (e.g., Fournier, Moskowitz, & Zuroff, 2008; Yao & Moskowitz, 2015).

Complementarity in High-fidelity Simulations

Although past research in social, personality, and clinical psychology showed strong evidence for complementarity in dyadic interpersonal interactions, this evidence cannot be straightforwardly applied to high-fidelity simulations for selection and development purposes. That is, Interpersonal Theory proposes the principles of complementarity to be strongest in naturally occuring, unstructured interactions (e.g., Kiesler, 1983). High-fidelity simulations for selection and development purposes, however, are often interpreted as strong situations that restrict the range of behavior that participants will show (see Meyer, Dalal, & Hermida, 2010). In particular, Trait Activation Theory (Lievens, Tett, & Schleicher, 2009; Tett & Burnett, 2003) proposes that participants' behavior is not only influenced by the interpersonal demands of work-related situations that are sampled during high-fidelity simulations (i.e., role-players' interpersonal behavior), but also by task and organizational situational demands (see also Moskowitz, Ho, & Turcotte-Tremblay, 2007; Oliver et al., 2016).

For dominance, there is reason to expect that the interpersonal dynamics between participants and role-players in high-fidelity simulations follow the principle of reciprocity. That is because high-fidelity simulations usually confront participants with problems and require them to propose solutions to solve these problems. We concur with the arguments mentioned by Oliver et al. (2016) that task and interpersonal situational demands will likely influence participants' expressions of dominance in the same way. That is, to solve a problem presented by a dominant role-player who actively leads the discussion, participants might overall best act submissively and listen closely to gather all available information about the problem situation. To solve a problem presented by a submissive role-player, however, who overall acts passively and does not disclose information herself, participants might overall

best lead the discussion with targeted questions to gain relevant information. In line with these arguments, Oliver et al. (2016) demonstrated a negative relation between role-players overall expression of dominance and the amount of directive communication on behalf of participants. As further support, Moskowitz et al. (2007) found even stronger evidence for reciprocity in dominance in work settings compared to nonwork settings. Thus, we expect that interpersonal dynamics between role-players and participants in high-fidelity simulations can be described with the principle of reciprocity in dominance.

Hypothesis 1: Role-players' and participants overall expression of dominance in high-fidelity simulations will be negatively related to each other.

Regarding affiliation, organizational or social norms and the expectation of participants to be evaluated in these situations might limit the expression of behavior indicative of low affiliation or coldness. Irrespective of the role-player's overall expression of affiliation, participants might thus show affiliative behavior overall. In line with this argument, Moskowitz et al. (2007) found weaker evidence for correspondence in affiliation in work settings compared to nonwork settings. Oliver et al. (2016) made a similar point and predicted that the strong task and organizational situational features of high-fidelity simulations for selection and development purposes might inhibit the applicability of the principle of correspondence in affiliation to participants' interpersonal behavior in high-fidelity simulations. That is, they predicted that participants would overall respond to role-players who express coldness with even higher levels of affiliation. Oliver et al. (2016) argued that this pattern might serve to build up a relationship that ultimately benefits to solve the problem that is discussed during the high-fidelity simulation. This hypothesis was supported by a negative relation between role-players' overall expression of affiliation and participants' overall relationship-building behavior across multiple high-fidelity simulations (Oliver et al., 2016). To provide further insights into this issue, we investigate whether assessees and role-

players overall follow the principle of correspondence in affiliation in high-fidelity simulations.

Research Question 1: Do participants and role-players overall show corresponding levels of affiliation in high-fidelity simulations?

Interpersonal Behavior and Dynamics at the Momentary Level

The level of measurement of interpersonal behavior. To discern predictable patterns in people's interpersonal behavior, past research in social and clinical psychology revealed that an appropriate understanding of interpersonal behavior and dynamics requires careful attention to the level of measurement. That is, interpersonal behavior can be described at least on one of the following levels (e.g., Sadler et al., 2011; Tracey, 2004): (a) An overall tendency of interpersonal behavior aggregated across time, situations or interaction partners (i.e., an overall interpersonal style), (b) an overall tendency of interpersonal behavior during a specific situation, such as an interaction with an interaction partner, and (c) the concrete interpersonal behavior during a situation on a momentary basis across time (Sadler et al., 2011; Tracey, 2004).

Although analyses of interpersonal behavior at all these levels can provide valuable insights, analyses of overall interpersonal styles and overall tendencies during a specific interaction cannot capture potential variability on a continuous moment-to-moment level across time (e.g., Pincus et al., 2014). This is crucial, however, because studies that investigated interpersonal behavior at the continuous level across various contexts have consistently demonstrated substantial intraindividual variability on this continuous, momentary level (e.g., Hopwood et al., 2018; Markey et al., 2010; Pennings et al., 2014; Sadler et al., 2009; Sadler, Woody, McDonald, Lizdek, & Little, 2015; Strong et al., 1988; Tracey, 1994; Tracey, 2004). Thus, an appropriate understanding of interpersonal behavior

and interpersonal dynamics requires the assessment of interpersonal behavior as it continuously occurs and unfolds across time in a specific interaction.

In high-fidelity simulations, it has been widely demonstrated and accepted that participants vary their behavior *across* different simulations (e.g., Gibbons & Rupp, 2009; Jackson, Michaelides, Dewberry, & Kim, 2016; Lance, 2008; Lievens, 2009; Putka & Hoffman, 2013). However, the question whether participants do vary their behavior also *within* simulations has received almost no attention (see also Brannick, 2008). Answers to the question of intraindividual variability in interpersonal behavior of assessees and role-players within high-fidelity simulations is crucial, however, to gain an accurate understanding of interpersonal behavior in high-fidelity simulations and to understand how the interpersonal behavior of participants and role-players translates into interpersonal dynamics. As one rare exception, Brannick, Michaels, and Baker (1989) report low intercorrelations between responses to different sets of in-basket items. However, this study does not reveal information about intraindividual variability in participants' interpersonal behavior during actual interactions with other human beings.

Moreover, intraindividual variability in behavior of role-players has been neglected in the same way. Role-players might receive instructions to display specific interpersonal behaviors and prompts (Lievens, Schollaert, & Keen, 2015; Schollaert & Lievens, 2011, 2012). It is however unclear, whether these general instructions aim to standardize role-players' interpersonal across simulations and participants. They might not constrain variability within simulations where it is expected that role-players adapt to the specific utterances of participants. In a lab experiment in which two students discussed with each other, Tracey (2004) found that even confederates who were instructed to display a specific interpersonal disposition towards their interaction partner showed intraindividual variability in interpersonal behavior. Hence, we propose:

Hypothesis 2: Participants and role-players vary their interpersonal behavior within high-fidelity simulations.

The level of measurement of complementarity. In line with the notion of interpersonal behavior at different levels, past research has also investigated the phenomenon of complementarity at various levels of analyses. For example, complementarity can be operationalized at the level of (a) interpersonal styles, (b) aggregated behavior during a specific interaction with a specific interaction partner, or even (c) at the level of behavioral interchanges within a specific interaction that describes how interaction partners mutually influence each other's behavior on a momentary basis across time (Sadler et al., 2011; Tracey, 2004). Although evidence for complementarity has been found across all these levels, evidence has been strongest for complementarity at the momentary interchange level (see Sadler et al., 2011 for an overview). In fact, in its essence, Interpersonal Theory focused its principles of complementarity on the momentary level of behavioral interchanges between individuals (Carson, 1969; Kiesler, 1983). Although operationalizations of complementarity at different levels of analysis were demonstrated to be interrelated (Tracey, 2004), complementarity at more aggregated levels of analyses (i.e., interpersonal styles and aggregated behavior within situations) might paint only a blurry picture of the true interpersonal dynamics that unfold across time within a given situation. That is because complementarity at higher levels do not account for how specific behavioral acts between interpersonal agents match the principles of complementarity at a specific moment in time (Sadler et al., 2011; Tracey, 2004). In line with this reasoning, Tracey (2004) demonstrated that momentary complementarity is closer related to interaction outcomes than complementarity of interpersonal styles or complementarity of interpersonal behavior aggregated across time in a given situation. Complementarity operationalized at the

momentary level might thus be best suited to capture the true interpersonal dynamics between interaction partners in a given situation (Sadler et al., 2011; Tracey, 2004).

Hence, a sound understanding of interpersonal dynamics between participants and role-players in high-fidelity simulations requires to analyze the occurrence of the principles of complementarity at the momentary level. Given that the only published study on interpersonal dynamics in high-fidelity simulations (Oliver et al., 2016) assessed interpersonal behavior only with single-point estimates, we lack knowledge about how role-players and participants mutually influence each other's interpersonal behavior on a momentary basis across time.

Although complementarity at the momentary level provides a more accurate estimate of interpersonal dynamics, we still expect that the given interpersonal, task, and organizational situational demands in high-fidelity simulations influence participants' interpersonal behavior in the same way as on the more aggregated level of analysis. In particular, we expect participants will consistently display behaviors of high affiliation, irrespective of any unfriendly or cold behavioral expressions on behalf of role-players. Thus, we do not expect to find evidence for correspondence in affiliation between participants and role-players at the momentary level.

Research question 2: Do participants and role-players show correspondence in affiliation in high-fidelity simulations at the momentary level?

However, we do expect that interpersonal and task demands in high-fidelity simulations will work in synchrony to let participants act more dominantly (e.g., ask targeted questions, propose problem solutions) if role-players act more submissively (e.g., do not actively reveal relevant information about the problem situation, ask for help). Vice versa, we expect participants to act more submissively (e.g., listen closely), if role-players act more dominantly (e.g., actively revealing relevant information about the problem situation, make a

demand). Hence, we expect evidence for reciprocity in dominance between participants and role-players at the momentary level.

Hypothesis 3: Participants and role-players show reciprocity in dominance in high-fidelity simulations at the momentary level.

The Predictive Value of Complementarity

Interpersonal Theory proposes that complementarity is an indicator of satisfying and successful interpersonal interactions (Carson, 1969; Kiesler, 1983). In particular, showing complementarity means that both interaction partners accept their own status within the hierarchy. This is meant to serve a process of self-validation and a feeling of security. However, when the principles of complementarity are not followed, interaction partners might perceive interpersonal anxiety and experience a less effective collaboration instead (Sullivan, 1953; Wiggins, 1980). In other words, interactions that follow the principles of complementarity satisfy the dominant and affiliative needs of both interaction partners in the given interpersonal situation (e.g., Carson, 1969; Kiesler, 1983).

Past empirical research in social and clinical psychology provided evidence for the benefit of complementarity (see Sadler et al., 2011 for an overview). For example, complementarity was found to be positively related to how interaction partners evaluate their interaction with each other in terms of measures such as satisfaction (Dryer & Horowitz, 1997; Locke & Sadler, 2007; Shechtman & Horowitz, 2006; Tracey, 2004), liking of each other (e.g., Markey et al., 2010; Nowicki & Manheim, 1991; O'Connor & Dyce, 1997; Tiedens & Fragale, 2003), or relationship quality (e.g., Ansell et al., 2008; Markey & Markey, 2007; Yaughn & Nowicki, 1999). Finally, complementarity also predicted performance in basic lab tasks (Estroff & Nowicki, 1992; Markey et al., 2010; Smelser, 1961).

Complementarity and performance in high-fidelity simulations. Despite the evidence that complementarity predicts important outcomes such as satisfaction in

interactions, liking of interaction partners or performance in simple tasks, it is not clear whether this equally applies to performance in high-fidelity simulations that capture job-relevant situations. For dominance, one might expect a positive relationship between complementarity and performance of participants in high-fidelity simulations. That is, for participants, acting more dominantly (e.g., actively ask for relevant information, propose a solution for a problem) when the role-player acts more submissively (e.g., not actively disclosing relevant information, asking for help) and acting more submissively (e.g., listen closely to gather relevant information) when the role-player acts more dominantly (e.g., leading the discussion and actively disclosing relevant information) should benefit to solve the problem in a given simulation. However, Oliver et al. (2016) found that the positive relationship between participants' use of directive communication and performance in high-fidelity simulations was not significant if participants interacted with a role-player who expressed low levels of control. Instead, Oliver et al. (2016) only found a positive relation between participants' use of directive communication and performance if participants had to interact with a role-player who expressed high levels of control. Oliver et al. (2016) explained this puzzling result by the fact that role-players who express high levels of control might not necessarily address relevant information to solve the problem. Instead, it is of special relevance for participants to then turn the conversation to the core of the problem. Although this provides a convincing explanation, an alternative explanation might be that the analyses from Oliver et al. (2016) did not investigate complementarity at the momentary level which paints a more accurate picture of the interpersonal dynamics. Thus, more research is necessary to gain knowledge about the relation between reciprocity in dominance and performance in high-fidelity simulations.

Research question 3: Does reciprocity in dominance predict participants' performance rated by role-players?

Regarding affiliation, task and organizational demands might create situations that punish participants for correspondence in affiliation because social and organizational norms usually appreciate friendly behavior. Hence, participants who respond to role-player expressions of unfriendliness with equal unfriendliness might be discredited and receive lower performance evaluations. One might therefore expect a negative relation between correspondence in affiliation between participants and role-players on one hand and role-play performance ratings on the other hand (see Oliver et al., 2016 for similar arguments). In line with this reasoning, Oliver et al. (2016) found a significant positive relationship between relationship-building behavior and performance of participants confronted with an unfriendly role-player but no significant relationship between relationship-building behavior and performance of participants confronted with a friendly role-player. However, Oliver et al. (2016) only investigated single-point estimates of aggregated interpersonal behavior within high-fidelity simulations. Hence, we need to investigate whether complementarity measured at the momentary level might lead to different conclusions.

Research question 4: Does correspondence in affiliation predict participants' performance rated by role-players?

An important caveat is in order, though. That is, in high-fidelity simulations, performance of participants is often rated by the role-players who interact with the participants. Hence, ratings made by role-players might be at least partly influenced by the fact that complementarity predicts individuals' satisfaction with an interaction and liking of the interaction partner. Interpersonal Theory predicts that such positive relationship outcomes might be obtained even if the principle of correspondence in affiliation is achieved by two interaction partners who show equal levels of unfriendly behavior, because this might serve as a form of self-validation and reduces interpersonal anxiety (Carson, 1969; Kiesler, 1983; Sullivan, 1953; Wiggins, 1980). Although complementarity might not necessarily be a true

determinant of high performance, role-players might then assign higher performance ratings to participants that they like more or that contribute to satisfying interactions. In this way, complementarity might generate a bias in performance ratings in high-fidelity simulations that has not been explored so far. In fact, Oliver et al. (2016) limited their investigations of the relation between interpersonal behavior or interpersonal dynamics and performance to performance ratings delivered by role-players who interacted with the participant. To investigate this potential bias, we gather additional performance ratings from assessors who do not involve in the interaction with the participant and solely focus on observing behavior and rating performance. In this way, assessors are not part of the interpersonal micro cosmos and are thus not influenced by the degree of complementarity that might occur between participants and role-players.

Research question 5: Does correspondence in affiliation predict participants' performance rated by assessors?

Research question 6: Does reciprocity in dominance predict participants' performance rated by assessors?

Complementarity and job-related performance. Apart from studies that investigated the relation between complementarity and social or clinical outcomes, research also demonstrated the benefit of complementarity to job- and organizational-related outcomes. One stream of research investigated the effects of the composition of work teams in terms of their individuals' interpersonal style. In line with the principle of reciprocity in dominance, teams with balanced amounts of highly extraverted individuals showed higher social cohesion (Barrick, Stewart, Neubert, & Mount, 1998), task focus and performance (Barry & Stewart, 1997). In addition, individuals reported higher attraction to their work teams if the individual's level of extraversion was dissimilar to the average team level of extraversion. Higher attraction to work teams was, in turn, an important predictor of

individuals' interindividual and task performance (Kristof-Brown, Barrick, & Stevens, 2005). In line with the principle of correspondence in affiliation, variance in team members' level of agreeableness was found to be positively related to team conflict, but negatively related to social cohesion, communication, and workload sharing (Barrick et al., 1998).

Another stream of research scrutinized the principle of reciprocity in dominance to supervisors and their employees. As support for the principle of reciprocity in dominance, employees reported higher levels of satisfaction with supervisors higher than themselves in control (Glomb & Welsh, 2005). Further, employees' level of proactivity was found to moderate the effect of leaders' extraversion on team unit performance. In particular, teams with employees low in proactivity were found to produce higher performance with more extraverted leaders. However, teams with employees high in proactivity were found to produce lower performance with more extraverted leaders (Grant, Gino, & Hofmann, 2011).

Given this evidence for the benefit of complementarity at the workplace, the level of complementarity shown in high-fidelity simulations might predict job-related or organizational criteria. Clearly, complementarity is a dyadic phenomenon. That is, the degree of complementarity that a specific dyad shows is dependent on both interaction partners' interpersonal behavior. Applied to the context of high-fidelity simulations, this means that the degree of complementarity in a given simulation is not only dependent on the interpersonal behavior of the participant and how the participant adapts the interpersonal behavior to the role-player. Instead, the degree of complementarity is also dependent on the level of adaptation that the role-player shows to the interpersonal behavior of the participant. However, if participants are able to establish interactions with high degrees of complementarity across multiple human actors across different situations, this might be seen as a stable, individual difference of the participant to adapt interpersonal behavior to different interaction partners (see, for example, Oliver & Lievens, 2014; Pincus et al., 2014; Sadler et

al., 2011 for similar arguments). Hence, observing the degree of complementarity that participants establish with different interaction partners across different high-fidelity simulations might predict job-related outcomes.

Intriguingly, the degree of complementarity that participants establish across multiple simulations with different interaction partners (i.e., role-players) matches the notion of interpersonal adaptability. Interpersonal adaptability (e.g., Oliver & Lievens, 2014; Ployhart & Bliese, 2006) is regarded as a crucial individual difference to successfully master the challenges of today's world of work in which individuals need to effectively collaborate with different human actors, such as employees, coworkers, supervisors, clients, or suppliers (see, for example, Griffin, Neal, & Parker, 2007; Pulakos, Arad, Donovan, & Plamondon, 2000). The degree of complementarity that participants establish across different situations (i.e., high-fidelity simulations) with different interaction partners (i.e., role-players) might thus provide a new angle to the measurement of interpersonal adaptability (Oliver & Lievens, 2014; see also Pincus et al., 2014; Sadler et al., 2011 for similar arguments).

Hypothesis 4: Participants averaged degree of complementarity across multiple high-fidelity simulations with different role-players predicts participants' interpersonal adaptability.

Further, the degree of complementarity that participants establish across multiple simulations with different interaction partners (i.e., role-players) will likely serve individuals to show high task performance. That is because complementarity is conceptualized to build the interpersonal fundament for successful cooperation (Carson, 1969; Kiesler, 1983; Wiggins, 1980). In any context in which individuals frequently interact with other human actors, complementarity might thus serve to facilitate interpersonal encounters and the fulfillment of tasks. Especially complementarity in dominance has been proposed to be beneficial for task performance because reciprocity in dominance is meant to indicate a

common agreement upon status and assignment of roles at any particular moment (Carson, 1969; Kiesler, 1983) that should serve efficient performance of tasks (see, for example, Bendersky & Hays, 2012; Locke & Sadler, 2007). In line with past research that associated the principles of complementarity with higher job-related and organizational performance (Barry & Stewart, 1997; Grant et al., 2011), we propose:

Hypothesis 5: Participants averaged degree of complementarity across multiple high-fidelity simulations with different role-players predicts participants' task performance.

Methods

To investigate our research questions and test our hypotheses, we captured participants' and role-players' interpersonal behavior in four short role-plays that confronted them with different interpersonal demands and sampled relevant situations in the broad interpersonal leadership domain.

Sample and Procedure

The entire MBA cohort of a European business school participated in our study to identify their strengths and weaknesses as leaders. The sample encompassed 96 participants (51% females, mean age = 23.63, *SD* = 1.85) from 19 different nations (67 % Belgian, 5 % Chinese, 4 % Romanian) who were studying to obtain a Master's degree in Marketing (51 %) or Financial Management. All had at least one year of working experience. The mean test-taking motivation of the participants was high: 3.96 (*SD* = 0.50, see below for the scale). Further anecdotal evidence supported that participants were motivated to perform well in the role-plays and to learn about their potential as leaders. For example, all participants wore business attire and appeared to be nervous.

About one week prior to participating in the role-plays, participants completed proctored computer-based tests which included a Big Five personality measure and other measures that are not relevant for the current study. Participants completed four different role-

plays in which they interacted with four different role-players. The role-plays took place in a large hall and lasted three minutes each. In the large hall, a circle was formed by desks. On each desk, one role-player was sitting. Role-players typically stayed at the same desk and played the same role-play again[14]. Participants' performance during each of the role-plays was rated by the respective role-player (who thus also served as assessor, see below). In a later stage, two to three independent assessors also rated participants' performance via video recordings of the role-plays. In addition, the interpersonal behavior of participants and role-players during the role-plays was rated by a distinct sample of coders who also watched the video recordings of the role-plays. Afterwards, participants received feedback reports about their performance in the role-plays. To provide criterion data, the MBA supervisors (instructors) and peers rated the job-related performance dimensions seven months after the role-plays.

Measures

Big Five personality. We assessed personality with the Business Attitudes Questionnaire (Vrijdags, Bogaert, Tribovic, & Van Keer, 2014). This is a work-related personality questionnaire that was developed and validated by an international HR consultancy. It was certified by the British Psychological Society. Each item of the BAQ asks participants to indicate agreement with a statement on a 5-point Likert scale (*1 = totally disagree; 5 = totally agree*). It comprises a total of 150 items. We used participants' summed scores on the Big Five dimensions. As we did not receive access to item-level data, we could not calculate internal consistency reliabilities. The test manual reports good psychometric properties in terms of internal consistency reliabilities ($.91 \leq \alpha \leq .94$), convergent validity

[14] To reduce role-player fatigue, role-players (a) played the same role-play not longer than four hours on each assessment day, (b) enjoyed scheduled breaks, and (c) were replaced by flying role-players in two to four instances.

with other contextualized and non-contextualized personality inventories, and criterion-related validity in terms of relations with job performance.

Role-plays. We developed four different role-play simulations that aimed to sample different interpersonal demands as well as different relevant situations in the leadership domain. To sample different interpersonal demands, we decided to develop one role-play for each of the four quadrants of the Interpersonal Circumplex Model (see Table 1).

To derive the content of the role-plays regarding different relevant situations in the leadership domain, we drew from two sources. First, we drew from leadership theories such as the Multiple-Linkage Model (Yukl, 2010). Second, experienced consultants from the HR consultancy firm served as subject matter experts. They were qualified as experts in the leadership domain because they had selected and developed successful leaders in multiple client projects.

Table 1

Overview of Content and Interpersonal Demand of Role-Plays

	Low affiliation	High affiliation
High dominance	*Role-play 2*: Role-player criticizes participant for slow decision-making.	*Role-play 4*: Friendly role-player proposes to change the core activity of the charity event and intends to find an ideal solution for all stakeholders.
Low dominance	*Role-play 1*: Role-player mentions a popular sport event on the same day as the charity event, but is not motivated to come up with a constructive solution.	*Role-play 3*: Friendly role-player asks for advice to maximize charity earnings during the event and cooperates effectively.

Pre-study. Prior to the data collection, we verified whether the four role-plays indeed sampled four different interpersonal demands in line with the four quadrants of the Interpersonal Circumplex. To do so, two female graduate students in industrial/organizational psychology with expertise in Interpersonal Theory and the Interpersonal Circumplex Model rated the role-players' overall expression of affiliation and dominance as depicted in the role-player instructions. The students made their ratings via the Interpersonal Grid (Moskowitz & Zuroff, 2005). The Interpersonal Grid projects the Interpersonal Circumplex to a Cartesian plane consisting of 11x11 boxes and several adjectives that describe the different sections of the Interpersonal Circumplex. Raters then tick a box that best matches the interpersonal behavior of a target person, leading to a score for affiliation and for dominance that can vary between 1 and 11.

Ratings of the role-player instructions showed moderate to excellent interrater reliabilities (ICC[2,1] = .96 for affiliation; ICC[2,1] = .65 for dominance) and approved the prescribed overall interpersonal behavior for role-players. Across raters, both role-plays with

low prescribed affiliation received low affiliation ratings of 3.5 and both role-plays with high prescribed affiliation received high affiliation ratings of 8.50 (role-play 3) and 9.00 (role-play 4). Both role-plays with low prescribed dominance received moderate dominance ratings of 5.5 (role-play 1) and 6 (role-play 3) and both role-plays with prescribed high dominance received high dominance ratings of 10 (role-play 2) and 7.5 (role-play 4). These results show that we succeeded with our intended manipulation of overall role-players' behavior for 3 out of the 4 role-plays. However, we exert caution for role-play 3 because the score for role-players' overall dominance slightly exceeded the score that might be seen as low or moderate in dominance.

A total of 13 role-players (12 females) acted in the four role-plays. These were seven experienced consultants from the HR consultancy and six graduate students from a large European university. As noted, each of these role-players were also live assessors. Apart from the live assessors (who were also role-players), in a later stage, recorded performances were rated by 8 other trained and paid assessors (4 females, mean age = 23.13, *SD* = 6.49). These were recruited from a European university. All of them were studying to obtain a Bachelor's (50 %) or Master's degree in various fields of Psychology or Business Administration.

Assessor training for the consultants from the HR consultancy and students included core aspects of both behavior-driven (Byham, 1977) and frame-of-reference training (Roch, Woehr, Mishra, & Kieszczynska, 2012). The training included lectures and exercises on observation, registration, classification, and evaluation of participants' performance. In addition, assessors were familiarized with the overall scoring procedure. Next, they practiced evaluating participants in the role-plays they specialized in. To this end, they first watched videotapes of role-plays and then independently provided evaluations. Assessors then met to reach consensus. This procedure was repeated for a total of three practice tapes.

Consistent with role-player training guidelines (Byham, 1977; Lievens, Schollaert, & Keen, 2015), role-players were taught to use prompts (Lievens, Schollaert, & Keen, 2015; Schollaert & Lievens, 2011, 2012) to structure the role-plays and elicit behavior. An example of a prompt was "If things continue this way, we will have to take the matter up with management". Role-players learned the prompts by heart. Finally, role-players also practiced and received feedback about their role-playing behavior.

To ensure that role-players (assessors) focused their performance ratings on observable and relevant behaviors, we developed short checklists per role-play that listed behaviors indicative of effective performance. Per role-play, they also provided overall ratings of role-play performance (*1 = should clearly be improved: starters' level* to *9 = obviously strong: role model behavior*). Across role-plays and role-players/assessors, average internal consistency reliabilities for role-play performance ratings was .70 (*SD* = .20).

The rating procedure was the same for assessors that watched recorded performances in a later stage. Rewinding or pausing role-play conversations was prohibited.[15] To limit biasing effects (e.g., order effects) in these assessor ratings, we took various precautions: (a) we counterbalanced the appearance of records per role-play across assessors, (b) we presented participants per role-play in a random order, and (c) for one assessor who made ratings of participants in two role-plays, we distributed all records for the two role-plays per assessor across four blocks that each contained records of only one role-play.

To provide estimates of interrater reliabilities of role-play performance ratings, we calculated G(*q*,*k*) coefficients. Per role-play, participants were nested within role-players, but fully crossed with assessors. Therefore, our design resembled an ill-structured measurement design (Putka, Le, McCloy, & Diaz, 2008). In ill-structured measurement designs, traditional indices of interrater reliability (e.g., intraclass correlations) generate biased estimates because

[15] For 20 percent of all conversations, cameras did not successfully record videos so that only audio records were available. In these cases, assessors used audio records to evaluate performance.

they do not explicitly distinguish between the contribution of assessor main effects, assessor-participant interaction effects, and residual variance to observed score variance (Putka et al., 2008). Putka and colleagues (2008) therefore proposed to calculate interrater reliabilities with the $G(q,k)$ coefficient that describes the proportion of expected observed score variance that is attributable to true score variance. Table 2 shows single-rater reliabilities ($G[q,1]$) as well as interrater reliabilities for performance ratings averaged across all assessors and interrater reliabilities for performance ratings averaged across role-players and all assessors ($G[q,k]$). Single-rater reliabilities that serve as another estimator of the reliability of role-player performance ratings were low to moderate (.23-.64, M = .41, SD = 0.20). Interrater reliabilities for performance ratings averaged across all assessors (without role-players included) were low to high (.45-.78, M = .65, SD = 0.16).

Table 2

Single-rater (G[q,1]) and Interrater Reliabilities for Role-play Performance Ratings averaged across all Assessors (G[q,k]) as well as Role-players and all Assessors

	All assessors					*Role-players and all assessors*				
Role-play	$\hat{k}$	*q-multiplier* *G(q,1)*	*q*-multiplier *G(q,k)*	*G*(*q*,*1*)	*G*(*q*,*k*)	$\hat{k}$	*q-multiplier* *G(q,1)*	*q*-multiplier *G(q,k)*	*G*(*q*,*1*)	*G*(*q*,*k*)
1	2	.50	.00	.64	.78	3	.73	.06	.64	.85
2	2	.50	.00	.35	.60	3	.73	.06	.26	.57
3	2	.50	.00	.29	.45	3	.75	.09	.23	.50
4	3	.67	.00	.54	.78	4	.78	.03	.53	.83

Note. $\hat{k}$ = harmonic mean number of assessors per participant. The *q*-multiplier scales the amount of variance that can be attributed to assessor main effects. With increasing overlap between sets of assessors that rate each participant, *q* approaches 0. With decreasing overlap between sets of assessors that rate each participant, *q* approaches $1/\hat{k}$ (Putka et al., 2008).

Interpersonal behavior via Social Behavior Inventory. Role-players filled in a short set of items from the Social Behavior Inventory (SBI; Moskowitz, 1994) to measure participants' overall interpersonal behavior. The SBI consists of four sets of items that correspond to each of the four end poles of the Interpersonal Circumplex Model (i.e., agreeableness, quarrelsomeness, dominance, and submissiveness). In prior empirical studies, the SBI has been shown to generate reliable and valid measures of affiliation and dominance (for an overview, see Moskowitz & Sadikaj, 2011). To limit cognitive load (related to simultaneously enacting and rating), we selected only three items from the SBI item pool per dimension (pole) per role-play that were most applicable to the role-play's content. In the role-plays, namely, they indicated whether the interpersonal behavior described by these items was shown by the participant ("+" [1]), not shown ("-" [0]), or not applicable. Table 3 shows example items as well as internal consistency reliabilities for mean scores across all composites per dimension. In line with past research (e.g., Moskowitz, 1994; Moskowitz & Zuroff, 2005), overall levels of participants' affiliation within a role-play were computed by subtracting the mean score of quarrelsomeness across prompts from the mean score of agreeableness across prompts. Overall levels of participants' dominance within a role-play were computed by subtracting the mean score of submissiveness across prompts from the mean score of dominance across prompts.

Table 3

Example Items and Internal Consistency Reliabilities (α) for the SBI

	Example item	α across prompts
Agreeableness	"*Smiled and laughed with the other*"	*M* = .80, *SD* = 0.11
Quarrelsomeness	"*Criticized the other*"	*M* = .69, *SD* = 0.16
Dominance	*"Tried to get the other to do something else"*	*M* = .72, *SD* = 0.05
Submissiveness	*"Gave in"*	*M* = .85, *SD* = 0.11

Note. Internal consistency reliabilities reported in this table are averaged across prompts and role-plays.

Interpersonal behavior via Continuous Assessment of Interpersonal Dynamics. Apart from assessing overall interpersonal behavior via the SBI in each of the four role-plays, we also used the Continuous Assessment of Interpersonal Dynamics (CAID; Sadler et al., 2009) to measure participants' and role-players' momentary interpersonal behavior per role-play. The CAID measures interpersonal behavior in line with the Interpersonal Circumplex Model via coders who use a joystick connected to a computer and joystick monitoring software. The x-axis of the joystick thereby represents the affiliation axis of the Interpersonal Circumplex Model on a scale between -1000 (extreme expression of unfriendliness) and +1000 (extreme expression of friendliness). The y-axis of the joystick thereby represents the dominance axis of the Interpersonal Circumplex Model, on a scale between -1000 (extreme expression of submissiveness) and +1000 (extreme expression of dominance).

Coders who use the CAID to measure interpersonal behavior watch video records of a dyadic interaction and focus on one target person only. Coders then monitor the interpersonal behavior of the target person, thereby moving the joystick to the position within the Interpersonal Circumplex Model that represents the current level of affiliation and dominance. The joystick monitoring software displays the Interpersonal Circumplex as well as the current position of the joystick on the computer screen and continuously writes data points for both affiliation and dominance every half of a second (Sadler et al., 2009, 2015; Thomas, Hopwood, Woody, Ethier, & Sadler, 2014; Tracey, Bludworth, & Glidden-Tracey, 2012). These time series can then be used to explore intraindividual variability in interpersonal behavior. Further, all data points per time series can be aggregated to gain an indicator of overall interpersonal behavior within a given interaction.

Past research showed that the CAID provides reliable and valid indicators for both overall interpersonal behavior as well as intraindividual variability. Interrater reliabilities for overall interpersonal behavior varied between modest (e.g., intraclass correlations of .58 and

.61 reported by Markey et al., 2010) and excellent (e.g., intraclass correlations of .88 reported by Tracey et al., 2012). Interrater reliabilities for momentary changes of targets' interpersonal behavior were moderate. For example, Markey et al. (2010) reported cross-correlations between time series of different coders between .60 and .65. Despite these promising results, it is strongly recommended to aggregate across multiple independent coders to gain even more reliable estimates (Sadler et al., 2009). In terms of validity, Sadler et al. (2009) demonstrated convergent validity evidence of overall interpersonal behavior derived from the CAID in terms of strong positive correlations to corresponding dimensions of the Social Behavior Inventory and discriminant validity evidence in terms of nonsignificant correlations to noncorresponding dimensional correlations. Further, several studies revealed evidence for momentary complementarity via CAID (e.g., Hopwood et al., 2018; Markey et al., 2010; Sadler et al., 2009) as well as significant relations between complementarity and relevant outcome variables (e.g., Markey et al., 2010; Tracey et al., 2012).

A total of 17 students from a European university served as coders and received compensation for their codings (16 females, mean age = 21.67, *SD* = 1.35). All of them were Bachelor's (59 %) or Master's students in Psychology. Coders received a training of about eight hours from the first author to appropriately use the CAID. The training integrated elements from frame-of-reference training (Roch et al., 2012) and elements used in past studies with the CAID (e.g., Markey et al., 2010; Sadler et al., 2009, 2015; Sadler & Woody, 2016).

First, coders received an in-depth introduction to the Interpersonal Circumplex Model. Coders were also introduced to the background of the role-plays. Finally, coders were familiarized with the BARS from Oliver (2012) that aims to describe the overall level of expressed interpersonal behavior on two separate 9-point scales with verbal anchors for different degrees of affiliation and dominance. Next, coders practiced rating overall level of

expressed affiliation and dominance of target persons in role-plays via this BARS via at least six video tapes of role-plays not used in this study and then independently provided evaluations. Coders then met to discuss and reach consensus. This procedure aimed to develop a common frame of reference of different levels of affiliation and dominance (see Roch et al., 2012).

Second, coders learned to apply the CAID (see Sadler & Woody 2016). Coders followed a short lecture about the principle to continuously code the interpersonal behavior of targets via a joystick. Next, the trainer mentioned adjectives indicative of various positions of the Interpersonal Circumplex Model and coders had to move the joystick to the corresponding position. Afterwards, coders watched how the trainer coded a target in a practice tape. Finally, coders practiced the CAID in at least seven role-play tapes that were not used in this study. The trainer monitored the coders' practice codings, resolved any problems, and presented means, standard deviations, cross-correlations as well as plots of the codings to give normative feedback.

All tapes from the four role-plays of the study were distributed to four to five coders. Coders who were assigned to a specific tape coded both the participant and the role-player in two distinct runs. To limit biasing effects (e.g., order effects), the order of tapes was randomized. About 2 % of codings had to be dropped due to technical errors. After dropping codings with errors, the average harmonic mean of coders per target was 3.92.[16]

[16] Similar to past research that applied the CAID, we used stationary video cameras with a fixed camera angle. This led to the case that targets sometimes moved out of the angle and were only partially visible on the video tapes. In line with past studies (Sadler et al., 2009, 2015), coders were instructed to use all available information about targets' interpersonal behavior at any particular moment. Even when targets were sometimes not fully visible, they often presented auditory cues (e.g., sighs, murmurs, or "uh-huhs") or showed body movements (e.g., leaning into the camera) that indicated interpersonal behavior. When targets were not speaking or presenting any auditory cues, the nonverbal behavior did not indicate any change in behavior or the target was not fully visible, coders were instructed to keep the joystick in the same position until any new information were available that indicated a change in interpersonal behavior that would require to adjust the position of the joystick (e.g., Sadler et al., 2009, 2015). The same instructions were given to coders regarding the coding of audio files for interactions that were not successfully recorded by the video cameras. Analyses on a prior data set indicated that codings of videos and codings of audios produce similar results regarding overall interpersonal behavior, intraindividual variability in behavior, and complementarity.

We calculated interrater reliabilities of the CAID ratings for both (a) overall behavior aggregated across time within a role-play, and (b) the continuous behavior interchange level within a role-play. In line with recommendations of past research (e.g., Sadler et al., 2009; Thomas et al., 2014), we deleted the first ten data points (i.e., first five seconds) of all coders' time series. This accounts for the fact that coders need to "settle" into the interaction watched. To calculate interrater reliability at the overall behavior level, we computed *G(q,k)* coefficients based upon the average interpersonal behavior per time series from all coders. Average *G(q,k)* across all four role-plays and targets (i.e., participants and role-players) was .50 for affiliation (*SD* = 0.12) and .46 for dominance (*SD* = 0.19). To calculate interrater reliability at the continuous behavior level, we computed cross-correlations between the time series across coders. Average cross-correlations across all four role-plays and targets was .22 for affiliation (*SD* = 0.08) and .36 for dominance (*SD* = 0.14).

Control measures. We assessed participants' test taking motivation via a scale with four items from Arvey, Strickland, Drauden and Martin (1990) via a 5-point Likert scale (*1 = totally disagree; 5 = totally agree*; internal consistency reliability = .67). We also included participants' gender and age as control variables. That is because gender is theoretically assumed to relate to the two dimensions of affiliation and dominance (e.g., Gurtman & Lee, 2009). Further, age might relate to role-play and job-related performance because older people might have gathered more job-experience in relatively young samples (McEvoy & Cascio, 1989).

Criterion measures. Each participant was rated by the instructors and peers (class mates) of the MBA program. Instructors rated various criteria that are related to adapting interpersonal behavior to different interaction partners (team member adaptivity, interpersonal adaptability) and task performance of leaders (task-oriented leadership, in-role behavior). Instructors provided the ratings via the relative percentile method (Goffin, Gellatly,

Paunonen, Jackson, & Meyer, 1996; Goffin, Jelley, Powell, & Johnston, 2009). In the relative percentile method, raters assign percentile scores to ratees. Goffin and colleagues developed this method to reduce rating inflation as the reference group to be used for the ratings consists of the average MBA student (i.e., a percentile score of 50). Prior research showed that the relative percentile method had higher criterion-related validity than conventional absolute rating formats (Goffin et al., 1996, 2009).

To investigate the factor structure of the instructor ratings, we conducted a series of confirmatory factor analyses in Mplus 7.4 (Muthén & Muthén, 1998-2015) by using the ML estimator. The ratings from the relative percentile method were used as indicators. We compared a one-factor model (Model A) with a model with two correlated factors of task performance and interpersonal adaptability (Model B). Only Model B showed a good model fit with exception of a poor RMSEA value (Model A/B: χ^2(df) = 35.17(2)/1.62(1), $p < .001$/ $p = .203$, CFI = .769/.996, RMSEA (90% CI) = .418 (.304-.544)/.081 (.000-.299), SRMR = .087/.014). Information criteria indicated that Model B did indeed show a better fit than Model A (Model A/B: AIC: 3413.94/3382.39).

Peers rated the same criterion dimensions as instructors via multi item scales (Griffin et al., 2007; Ployhart & Bliese, 2006; Williams & Anderson, 1991; Yukl, 1999), using a 5-point Likert scale (*1 = below average*; *5 = truly exceptional*). Table 4 shows example items and internal consistency reliabilities of the scales' ratings. To reduce leniency in ratings, instructors assigned class mates as peers when they knew a participant well (e.g., due to project work). Peers' acquaintance with the target participants varied between 6 and 279 months ($M = 12.04$, $SD = 24.64$). For their participation in the criterion study, participants received a coupon of 5 € and the chance to win another coupon of 100 € in a lottery. All but two participants were rated by at least one peer.[17] To investigate the factor structure of the

[17] We did not calculate interrater reliabilities for peer ratings of performance because we received a second peer rating for only 28 participants.

peer ratings, we conducted the same confirmatory factor analyses as for the instructor ratings. For the peer ratings, we used mean scores across all items for each of the four scales as factor indicators. Again, only Model B showed a good model fit with exception of a poor RMSEA value (Model A/B: χ^2(df) = 53.19(2)/2.73(1), $p < .001/p = .098$, CFI = .788/.993, RMSEA (90% CI) = .460 (.358-.571)/.120 (.000-.300), SRMR = .096/.013). Information criteria indicated that Model B did indeed show a better fit than Model A (Model A/B: AIC: 1062.50/1014.04).

Table 4

Used Scales for Peer Ratings

Criterion	Measures	Example Items	Internal consistency
Task performance	*In-role behavior* 7 items adapted from Williams and Anderson (1991)	Adequately completes assigned duties.	.94
	Task-oriented leadership: 6 items from Yukl (1999)	Clearly explains what results are expected for a task or project.	.92
Interpersonal adaptability	*Interpersonal adaptability:* 7 items from Ployhart & Bliese (2006)	Adapts his/her behavior to get along with others.	.93
	Team member adaptivity: 3 items from Griffin et al. (2007)	Responds constructively to changes in the way the team works.	.86

Note. We did not calculate inter-rater reliabilities, because we only received a second peer rating for 27 participants.

Results

Prior to all calculations that involved CAID codings, we aggregated time series across all coders who coded the same target.

Manipulation Check

We used the CAID codings as another manipulation check to investigate whether the four role-plays indeed capture different overall interpersonal demands. To do so, we investigated means of role-players' average interpersonal behavior across the time series per role-play. For affiliation, overall role-players' interpersonal behavior in the four role-plays matched our intended manipulation. That is, role-players overall showed negative, low values for affiliation in role-play 1 (*M* = -47.20, *SD* = 104.07, *95%CI* = [-68.29; -26.12]) and role-play 2 (*M* = -120.52, *SD* = 147.81, *95%CI* = [-150.63; -90.41]), but positive, high values for affiliation in role-play 3 (*M* = 183.10, *SD* = 98.05, *95%CI* = [163.02; 203.18]) and role-play 4 (*M* = 93.73, *SD* = 95.27, *95%CI* = [74.33; 113.14]). Based upon the 95% confidence intervals, the four role-plays confronted participants with distinct levels of affiliation.

For dominance, role-players overall showed positive values across all four role-plays. However, as indicated by the 95% confidence intervals, participants were confronted with three distinct overall levels of dominance. In line with the intended manipulation, role-play 1 showed a low positive expression of dominance (*M* = 47.21, *SD* = 129.00, *95%CI* = [21.07; 73.34]), role-play 2 showed the highest expression of dominance (*M* = 236.68, *SD* = 104.51, *95%CI* = [215.39; 257.97]) and role-play 3 (*M* = 146.06, *SD* = 91.69, *95%CI* = [127.28; 164.85]), and 4 (*M* = 162.54, *SD* = 117.05, *95%CI* = [138.69; 186.38]) confronted participants with an expression of dominance that lies in-between of the other two role-plays.

Although the results for dominance indicate that we did not sample the four quadrants of the Interpersonal Circumplex, the results do however imply that the four role-plays confronted participants with different overall interpersonal demands.

Construct Validation of CAID Codings

To test the construct-related validity of CAID codings for participants, we related participants' overall interpersonal behavior as indicated by the CAID codings to role-players' ratings of participants overall interpersonal behavior via the SBI. For affiliation, both measures showed significant positive correlations in role-play 2 ($r = .35$, $p < .001$) and role-play 3 ($r = .24$, $p = .047$), but nonsignificant positive correlations in role-play 1 ($r = .15$, $p = .140$) and role-play 4 ($r = .13$, $p = .214$). At the level of aggregated overall interpersonal behavior averaged across all four role-plays, both measures correlated significantly at $r = .41$ ($p < .001$).

For dominance, both measures showed significant positive correlations in role-play 1 ($r = .50$, $p < .001$), role-play 2 ($r = .32$, $p = .002$), and role-play 4 ($r = .42$, $p < .001$), but a nonsignificant positive correlation in role-play 3 ($r = .06$, $p = .551$). At the level of aggregated overall interpersonal behavior averaged across all four role-plays, both measures correlated significantly at $r = .45$ ($p < .001$). Thus, overall, we found supporting evidence for the construct-related validity of our CAID codings.

Complementarity in High-fidelity Simulations

Hypothesis 1 and research question 1dealt with the fact whether participants and role-players follow the principles of complementarity at the overall level within high-fidelity simulations. To investigate this, we calculated correlations between participants' and role-players' overall expressions of dominance and affiliation indicated by averaged time series provided by the CAID per role-play.

For dominance, we found a significant positive correlation between participants' and role-players' overall interpersonal behavior in role-play 2 ($r = .21$, $p = .042$), indicating that higher expressions of dominance from role-players tend to provoke higher expressions of dominance from participants and vice versa. In all other role-plays, role-players' and

participants' overall level of dominance was not significantly correlated ($.00 \leq r \leq .09$, all *p*s > .05). Thus, we did not find evidence for Hypothesis 1 that role-participants and role-players show reciprocity in dominance at the overall level.

For affiliation, we found significant positive correlations between participants' and role-players' overall interpersonal behavior in role-play 1 ($r = .21$, $p = .042$), role-play 3 ($r = .47$, $p < .001$) and role-play 4 ($r = .24$, $p = .019$), but not in role-play 2 ($r = .10$, $p = .356$). In three out of four role-plays, participants and role-players thus adapted their interpersonal behavior to each other at the overall level in line with the principle of correspondence in affiliation.

Interpersonal Behavior and Dynamics at the Momentary Level

Interpersonal behavior at the momentary level. Hypothesis 2 proposed that participants and role-players vary their interpersonal behavior within high-fidelity simulations. To investigate this, we calculated the standard deviation across all data points per time series for participants and role-players. Further, we calculated one-sample *t*-tests to explore whether the mean standard deviation across all participants and role-players significantly deviate from zero. For participants, average intraindividual variability across all four role-plays as indicated by the standard deviation across all data points per time series per role-play was 82.06 ($SD = 6.64$) for affiliation and 167.55 ($SD = 4.10$) for dominance. For role-players, average intraindividual variability was 99.64 ($SD = 28.13$) for affiliation and 144.18 ($SD = 3.24$) for dominance. For both participants and role-players, significant one-sample *t*-tests showed that the mean intraindividual variability significantly deviated from zero for both affiliation and dominance in all four role-plays (all *p*s < .001, see Table 5).

Table 5

Intraindividual Variabilities in Interpersonal Behavior for Participants and Role-players per Role-play

	Affiliation participants					*Affiliation role-players*				
Role-play	*M*	*SD*	*t(df)*	*p*	*95%CI*	*M*	*SD*	*t(df)*	*p*	*95%CI*
1	86.25	31.51	26.82(95)	< .001	[79.87; 92.63]	118.78	40.42	28.79(95)	< .001	[110.59; 126.97]
2	86.90	36.04	23.50(94)	< .001	[79.55; 94.24]	127.75	34.23	36.38(94)	< .001	[120.78; 134.72]
3	72.51	23.55	29.85(93)	< .001	[67.69; 77.33]	68.90	27.11	24.64(93)	< .001	[63.35; 74.46]
4	82.58	37.71	21.35(94)	< .001	[74.90; 90.26]	83.11	37.78	21.44(94)	< .001	[75.41; 90.80]
	Dominance participants					Dominance role-players				
Role-play	*M*	*SD*	*t(df)*	*p*	*95%CI*	*M*	*SD*	*t(df)*	*p*	*95%CI*
1	164.07	38.55	41.70(95)	< .001	[156.26; 171.88]	144.00	33.54	42.07(95)	< .001	[137.21; 150.80]
2	163.93	43.05	37.12(94)	< .001	[155.16; 172.70]	140.10	46.32	29.48(94)	< .001	[130.66; 149.54]
3	170.85	40.10	41.31(93)	< .001	[162.64; 179.07]	144.63	44.99	31.17(93)	< .001	[135.42; 153.84]
4	171.35	40.82	40.91(94)	< .001	[163.03; 179.66]	147.99	44.67	32.29(94)	< .001	[138.89; 157.09]

Note. Intraindividual variabilities = standard deviations across all data points per time series.

Interpersonal complementarity at the momentary level. Hypothesis 3 and research question 2 dealt with the fact whether participants and role-players adapt their interpersonal behavior to each other in line with the principle of complementarity at the momentary level. To investigate this, we examined cross-correlations between the time series of participants and role-players within each role-play[18]. Further, we used one-sample *t*-tests to explore whether the cross-correlations significantly deviate from zero.

For both affiliation and dominance, we found consistent evidence for the principles of complementarity at the momentary level. For dominance, we found negative correlations between the time series of participants and role-players in all four role-plays (M = -0.56, SD = 0.10). This indicates that more submissive behavior of role-players evokes more dominant behavior from participants, and vice versa. For affiliation, we found positive correlations between the time series of participants and role-players in all four role-plays (M = 0.24. SD = 0.05). This indicates that more friendly behavior of role-players evokes more friendly behavior from participants, and vice versa. One-sample *t*-tests showed that all average cross-correlations significantly deviated from zero (see Table 6). Thus, there is evidence that participants and role-players follow the principles of correspondence in affiliation and reciprocity in dominance at the momentary level, which provides support for Hypothesis 3.

[18] Momentary complementarity has been calculated as Pearson cross-correlation per dyad. Cross-correlations per dyad have then been standardized using Fisher's *z* transformation for all further analyses. Results reported for momentary complementarity thus follow the metric of Fisher's *z* (see Hopwood et al., 2018).

Table 6

Momentary Complementarity in Affiliation and Dominance between Participants and Role-players per Role-play

	Affiliation					*Dominance*				
Role-play	*M*	*SD*	*t(df)*	*p*	*95%CI*	*M*	*SD*	*t(df)*	*p*	*95%CI*
1	.17	0.44	3.84(95)	< .001	[0.08; 0.26]	-.44	0.42	-10.09(95)	< .001	[-0.52; -0.35]
2	.21	0.37	5.62(94)	< .001	[0.14; 0.29]	-.51	0.44	-11.33(94)	< .001	[-0.60; -0.42]
3	.28	0.33	8.42(93)	< .001	[0.22; 0.35]	-.67	0.38	-17.10(93)	< .001	[-0.75; -0.59]
4	.28	0.35	7.64(94)	< .001	[0.20; 0.35]	-.61	0.47	-12.71(94)	< .001	[-0.71; -0.52]

Note. Momentary complementarity has been calculated as Pearson cross-correlation per dyad. Cross-correlations per dyad have then been standardized using Fisher's *z* transformation for all further analyses. Results reported for momentary complementarity follow the metric of Fisher's *z* (see Hopwood et al., 2018).

Complementarity and Performance in High-fidelity Simulations

Research questions 3-6 dealt with the relation between complementarity and performance ratings made by role-players as well as independent assessors in high-fidelity simulations. To investigate this, we examined relations between complementarity at the level of overall behavior aggregated within each role-play (i.e., overall complementarity) as well as at the level of momentary complementarity and role-play performance ratings. Further, we investigated relations between complementarity and performance ratings within each role-play as well as the same relations aggregated across all four role-plays.

To examine relations between overall complementarity and role-play performance ratings, we first calculated a deviation score per dyad and dimension that indicates the total deviation from perfect complementarity (see Ansell et al., 2008). For affiliation, we calculated overall complementarity as:

Overall complementarity in affiliation =

$$\sqrt{(overall\ affiliation\ participant - overall\ affiliation\ role - player)^2}$$

For dominance, we calculated overall complementarity as:

Overall complementarity in dominance =

$$\sqrt{(overall\ dominance\ participant + \ overall\ dominance\ role - player)^2}$$

We then calculated Pearson correlations between these deviation scores and role-play performance ratings made by role-players as well as independent assessors. For affiliation, overall complementarity predicted role-player ratings only in role-play 2 (r = .33, p = .001). Given that our score for overall complementarity mirrors a dyad's deviation from perfect complementarity, the positive relation to role-players' performance ratings indicates that role-players gave lower performance ratings if the interaction with the participant overall followed the principle of correspondence in affiliation. Assessor ratings of performance in each of the four single role-plays was unrelated to overall complementarity in affiliation. For dominance,

overall complementarity predicted both role-player (r = .23, p = .024) as well as assessor ratings (r = .26, p = .011), but only in role-play 1. Again, the positive relation indicates that participants receive lower performance ratings if the interaction with the role-player overall followed the principle of reciprocity in dominance.

We further investigated relations between overall complementarity and performance ratings averaged across all four role-plays. For affiliation, overall complementarity predicted both role-player (r = .33, p < .001) and assessor ratings of performance (r = .20, p = .049). For dominance, overall complementarity predicted only assessor (r = .23, p = .023) but not role-player ratings (r = .03, p = .752). In sum, higher overall complementarity appears to be related to lower role-play performance ratings (see Table 7).

To examine relations between momentary complementarity and role-play performance ratings, we calculated Pearson correlations between the Fisher's z standardized cross-correlations of time series per dimension on one hand, and role-play performance ratings on the other hand. For affiliation, momentary complementarity was only related to role-player ratings of performance in role-play 3 (r = .23, p = .027). Given that higher cross-correlations between participants' and role-players' time series for affiliation indicate higher complementarity, higher complementarity was related to higher performance ratings made by role-players. However, assessor ratings of performance were unrelated to momentary complementarity in affiliation in each of the four single role-plays. For dominance, momentary complementarity was only related to role-player ratings of performance in role-play 3 (r = -.23, p = .027). Given that more negative cross-correlations between participants' and role-players' time series for dominance indicate higher complementarity, higher complementarity was related to higher performance ratings made by role-players. In line with the result for affiliation, assessor ratings of performance were unrelated to momentary complementarity in dominance in each of the four role-plays.

We further investigated relations between momentary complementarity and performance ratings averaged across all four role-plays. For affiliation, momentary complementarity marginally predicted both role-player (r = .18, p = .075) and assessor ratings of performance (r = .18, p = .084). For dominance, momentary complementarity in dominance predicted only role-player (r = -.26, p = .001) but not assessor ratings of performance (r = -.13, p = .207). In sum, there appears to be a trend that higher momentary complementarity is related to higher performance ratings. However, the relations for affiliation are only marginally significant and the relations for dominance are only shown for role-player ratings.

Table 7

Descriptives and Correlations for Variables Averaged Across Role-plays

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Control variables																						
1 Gender	1.51																					
2 Age	23.62	1.85	-.15																			
3 Test motivation	3.96	0.50	-.08	.15																		
Personality																						
4 Altruism	92.82	13.54	.04	-.27**	.25																	
5 Extraversion	90.82	12.54	.03	-.17	.17	.53**																
Interpersonal behavior																						
6 Affiliation SBI	0.36	0.16	.02	-.12	-.09	.10	.04															
7 Dominance SBI	0.25	0.19	-.14	-.21*	.09	.15	.16	.15														
8 Affiliation CAID	171.24	70.67	.14	-.03	-.06	.11	.15	.41**	.12													
9 Dominance CAID	188.19	85.70	-.21*	-.23*	.11	.11	.30**	.02	.45**	.06												
10 Intraindividual variability affiliation	82.11	18.08	-.01	.04	.17	.10	.16	-.25*	.07	-.30**	.33**											
11 Intraindividual variability dominance	167.99	21.42	.28**	.05	.14	-.08	.01	-.27**	.07	-.08	-.08	-.03										
Interpersonal dynamics																						
12 OC affiliation	172.15	64.04	.12	-.15	-.06	.18	.17	.27**	-.02	.55**	.18	-.03	-.15									
13 OC dominance	336.29	122.63	-.22*	-.04	.16	.00	.20	-.19	.12	-.10	.78**	.45**	-.21*	.03								
14 MC affiliation	0.24	0.20	-.21*	-.09	-.05	.13	.19	-.03	.14	-.25*	.14	.09	-.09	-.28**	.08							
15 MC dominance	-0.55	0.29	-.14	.11	.17	-.17	-.05	-.21*	-.18	-.33**	.22*	.40**	-.18	-.14	.62**	.07						
Role-play performance																						
16 Role-player ratings	5.46	1.08	-.07	-.28**	.09	.11	.30**	.34**	.42**	.31**	.44**	-.12	-.02	.33**	.03	.18	-.26**					
17 Assessor ratings	5.62	0.80	.02	-.38**	.15	.16	.31**	.18	.52**	.28**	.54**	-.13	.09	.20*	.23*	.18	-.13	.65**				
Instructor ratings																						
18 Interpersonal adaptability	56.48	22.38	-.27**	-.11	.03	.10	-.03	.15	.17	.20	.36**	-.09	-.16	.02	.25*	.20*	-.03	.31**	.33**			
19 Task performance	60.17	22.46	-.08	-.34**	.13	.17	.01	.12	.10	.29**	.21*	-.17	.01	.26*	.03	-.12	-.23*	.34**	.44**	.50**		
Peer rating																						
20 Interpersonal adaptability	2.82	0.76	.00	-.10	-.04	.04	-.06	.11	.29**	.28**	.06	-.18	-.03	.06	.01	-.07	-.12	.14	.29**	.16	.21*	
21 Task performance	2.84	0.78	.19	-.05	-.05	-.03	-.09	.10	.15	.14	.11	-.09	-.01	.08	.07	-.22*	-.07	.08	.33**	.12	.29**	.52**

Note. $N = 96$. Gender is coded as follows: male = 1, female = 2. Interpersonal behavior refers to participants. OC: Overall complementarity = a dyad's total deviation from perfect complementarity. MC: Momentary complementarity. * $p < .05$, ** $p < .01$.

To investigate whether complementarity adds incremental validity to the prediction of role-play performance ratings beyond other predictors that tap into the interpersonal domain, we ran multiple regressions to predict role-player as well as assessor rated performance averaged across all four role-plays. Both multiple regressions followed the same procedure: In step 1, we included gender and age as controls. In step 2, we added self-reported extraversion and agreeableness to the model[19] as well as participants' overall affiliation and dominance in the role-plays. To limit multicollinearity, we used SBI ratings made by role-players for these variables. In step 3, we added overall complementarity in affiliation and dominance. In step 4, we added momentary complementarity in affiliation and dominance.

Results for role-player ratings of performance showed that including overall complementarity in the model explained additional 5 % of variance (p = .011), and including momentary complementarity in the model explained further 4 % of variance (p = .020) in role-player ratings beyond all other control variables and interpersonal predictors (see Table 8). Results for assessor ratings of performance showed that including overall complementarity in the model explained additional 4 % of variance (p = .033) in assessor ratings beyond all other control variables and interpersonal variables. Including momentary complementarity in the model did not increase explained variance significantly, however (p = .113; see Table 9).

[19] We added these measures because the Interpersonal Circumplex Model overlaps with the two interpersonal dimensions of extraversion and agreeableness of the Five-Factor Model of personality (e.g., McCrae & Costa, 1989, 1995). In particular, the Interpersonal Circumplex can be interpreted as a rotated version of a Cartesian plane that is built upon extraversion and agreeableness. In the language of the Interpersonal Circumplex Model, extraversion can be described as expression of dominance and warmth. In turn, agreeableness can be described as an expression of submissiveness and warmth (e.g., Markey & Markey, 2006; McCrae & Costa, 1989; see also Leising & Bleidorn, 2011).

Table 8

Multiple Regression to Predict Role-Player Rated Role-Play Performance

	R^2 *adj*	Δ R^2 *adj.*	*F(df)*	Sig. *F* change	*B*	*SE B*	β	*p*
Step 1	.074		4.73(2,92)					.011
Constant					9.95	1.47		< .001
Gender					-0.26	0.22	-0.12	.227
Age					-0.17	0.06	-0.30	.004
Step 2	.305	.231	7.88(6,88)	< .001				
Constant					6.10	1.73		< .001
Gender					-0.14	0.19	-0.06	.479
Age					-0.11	0.05	-0.19	.043
Agreeableness					-0.01	0.01	-0.18	.088
Extraversion					0.03	0.01	0.31	.003
Affiliation					1.82	0.58	0.28	.002
Dominance					1.69	0.51	0.30	.001
Step 3	.359	.054	7.59(8,86)	.011				< .001
Constant					5.79	1.71		.001
Gender					-0.20	0.19	-0.09	.294
Age					-0.10	0.05	-0.17	.063
Agreeableness					-0.02	0.01	-0.21	.039
Extraversion					0.03	0.01	0.29	.005
Affiliation					1.28	0.59	0.19	.033
Dominance					1.90	0.50	0.34	< .001
Overall complementarity affiliation					0.00	0.00	0.27	.003
Overall complementarity dominance					0.00	0.00	-0.06	0.516
Step 4	.402	.043	7.32(10,84)	.020				< .001
Constant					4.81	1.69		.006
Gender					-0.13	0.19	-0.06	.496
Age					-0.08	0.05	-0.14	.109
Agreeableness					-0.02	0.01	-0.24	.018
Extraversion					0.02	0.01	0.24	.017
Affiliation					1.24	0.57	0.19	.032
Dominance					1.56	0.50	0.28	.003
Overall complementarity affiliation					0.01	0.00	0.30	.001
Overall complementarity dominance					0.00	0.00	0.07	.518
Momentary complementarity affiliation					1.08	0.48	0.20	.027
Momentary complementarity dominance					-0.79	0.42	-0.21	.063

Note. $N = 95$. Gender is coded as follows: male = 1, female = 2. Overall complementarity = a dyad's total deviation from perfect complementarity.

Table 9

Multiple Regression to Predict Assessor Rated Role-Play Performance

	R^2 *adj*	ΔR^2 *adj.*	*F(df)*	Sig. *F* change	*B*	*SE B*	β	*p*
Step 1	.123		7.76(2,92)					< .001
Constant					9.62	1.06		< .001
Gender					-0.07	0.16	-0.05	.645
Age					-0.16	0.04	-0.38	< .001
Step 2	.361	.238	9.86(6,88)	< .001				< .001
Constant					6.74	1.22		< .001
Gender					0.06	0.14	0.04	.642
Age					-0.11	0.04	-0.26	.005
Agreeableness					-0.01	0.01	-0.12	.220
Extraversion					0.02	0.01	0.26	.011
Affiliation					0.41	0.41	0.08	.318
Dominance					1.82	0.36	0.44	< .001
Step 3	.396	.035	8.71(8,86)	.033				< .001
Constant					5.97	1.23		< .001
Gender					0.11	0.14	0.07	.407
Age					-0.10	0.04	-0.23	.010
Agreeableness					-0.01	0.01	-0.11	.250
Extraversion					0.01	0.01	0.20	.050
Affiliation					0.40	0.43	0.08	.349
Dominance					1.86	0.36	0.45	< .001
Overall complementarity affiliation					0.00	0.00	0.15	.089
Overall complementarity dominance					0.00	0.00	0.17	.061
Step 4	.413	.017	7.61(10,84)	.113				< .001
Constant					5.42	1.24		< .001
Gender					0.15	0.14	0.09	.280
Age					-0.09	0.04	-0.21	.018
Agreeableness					-0.01	0.01	-0.13	.180
Extraversion					0.01	0.01	0.16	.113
Affiliation					0.38	0.42	0.08	.365
Dominance					1.66	0.37	0.40	< .001
Overall complementarity affiliation					0.00	0.00	0.16	.075
Overall complementarity dominance					0.00	0.00	0.28	.016
Momentary complementarity affiliation					0.53	0.35	0.13	.138
Momentary complementarity dominance					-0.48	0.31	-0.17	.120

Note. $N = 95$. Gender is coded as follows: male = 1, female = 2. Overall complementarity = a dyad's total deviation from perfect complementarity.

Complementarity and Job-related Performance

Hypothesis 4 and 5 proposed that participants' averaged degree of complementarity across multiple high-fidelity simulations with different role-players predict participants' interpersonal adaptability and task performance. To investigate this, we explored relations between both overall and momentary complementarity on one hand and both instructor as well as peer ratings of job-related performance on the other hand.

For interpersonal adaptability, instructor ratings were significantly predicted by overall complementarity in dominance (r = .25, p = .013) as well as by momentary complementarity in affiliation (r = .20, p = .048). The direction of the correlations indicate that reciprocity in dominance at the overall level predicted lower ratings of interpersonal adaptability whereas correspondence in affiliation at the momentary level predicted higher ratings of interpersonal adaptability. Thus, we found mixed support for hypothesis 4.

For task performance, instructor ratings were significantly predicted by overall complementarity in affiliation (r = .26, p = .011) as well as momentary complementarity in dominance (r = -.23, p = .028). Further, peer ratings were significantly predicted by momentary complementarity in affiliation (r = -.22, p = .035). The direction of the correlations indicate that correspondence in affiliation predicts lower ratings of task performance whereas reciprocity in dominance predicts higher ratings of task performance. Thus, overall we found mixed evidence for hypothesis 5.

To investigate whether complementarity adds incremental validity to the prediction of job-related performance ratings beyond other predictors that tap into the interpersonal domain, we ran multiple regressions to predict instructor rated interpersonal adaptability, instructor rated task performance and peer rated task performance. All multiple regressions followed the same procedure: In step 1, we included gender and age as controls. In step 2, we added self-reported extraversion and agreeableness as well as participants' overall affiliation

and dominance in the role-plays. To limit multicollinearity, we used SBI ratings made by role-players for these variables. In step 3, we added overall complementarity in affiliation and dominance. In step 4, we added momentary complementarity in affiliation and dominance.

Results for instructor ratings of interpersonal adaptability showed that including overall complementarity in the model explained additional 3 % of variance, but this was only marginally significant (p = .093). Including momentary complementarity in the model explained further 5 % of variance (p = .027) in instructor rated interpersonal adaptability beyond all control variables and other interpersonal predictors.

Results for instructor ratings of task performance showed that including overall complementarity in the model explained additional 4 % of variance, but this was only marginally significant (p = .065). Including momentary complementarity in the model explained further 5 % of variance (p = .027) in instructor rated task performance beyond all control variables and other interpersonal predictors.

Results for peer ratings of task performance, however, showed that neither adding overall complementarity (p = .237), nor adding momentary complementarity (p = .171) to the regression model added incremental validity beyond all control variables and other interpersonal predictors.

Table 10

Multiple Regression to Predict Instructor Rated Interpersonal Adaptability

	R^2 *adj*	ΔR^2 *adj.*	*F(df)*	Sig. *F* change	*B*	*SE B*	β	*p*
Step 1	.081		5.08(2,91)					.008
Constant					121.60	30.44		< .001
Gender					-13.45	4.50	-0.30	.004
Age					-1.91	1.21	-0.16	.118
Step 2	.078	-.003	2.32(6,87)	.446				.040
Constant					102.76	21.84		.016
Gender					-12.85	4.62	-0.29	.007
Age					-1.39	1.29	-0.12	.285
Agreeableness					0.20	0.20	0.12	.320
Extraversion					-0.23	0.22	-0.13	.284
Affiliation					17.16	13.95	0.12	.222
Dominance					8.16	12.33	0.07	.510
Step 3	.108	.030	2.40(8,85)	.093				.022
Constant					76.69	43.15		.079
Gender					-9.68	4.77	-0.22	.045
Age					-1.07	1.29	-0.09	.408
Agreeableness					0.25	0.20	0.15	.205
Extraversion					-0.31	0.22	-0.17	.156
Affiliation					24.33	14.56	0.18	.098
Dominance					7.13	12.24	0.06	.562
Overall complementarity affiliation					-0.01	0.04	-0.03	.803
Overall complementarity dominance					0.05	0.02	0.24	.030
Step 4	0.162	.054	2.78(10,83)	.027				.005
Constant					54.68	42.58		.203
Gender					-8.60	4.70	-0.19	.071
Age					-0.76	1.25	-0.06	.545
Agreeableness					0.21	0.19	0.13	.285
Extraversion					-0.42	0.21	-0.23	.054
Affiliation					23.77	14.11	0.17	.096
Dominance					-2.56	12.50	-0.02	.835
Overall complementarity affiliation					-0.01	0.04	-0.02	.884
Overall complementarity dominance					0.08	0.03	0.41	.002
Momentary complementarity affiliation					20.18	11.82	0.18	.091
Momentary complementarity dominance					-23.28	10.45	-0.29	.029

Note. N = 94. Gender is coded as follows: male = 1, female = 2. Overall complementarity = a dyad's total deviation from perfect complementarity.

Table 11

Multiple Regression to Predict Instructor Rated Task Performance

	R^2 *adj*	ΔR^2 *adj.*	*F(df)*	Sig. *F* change	*B*	*SE B*	β	*p*
Step 1	.112		6.86(2,91)					.002
Constant					171.20	30.03		< .001
Gender					-5.79	4.44	-0.13	.195
Age					-4.33	1.19	-0.36	< .001
Step 2	.095	-.017	2.62(6,87)	.691				.022
Constant					160.62	41.61		< .001
Gender					-5.85	4.59	-0.13	.206
Age					-4.04	1.29	-0.34	.002
Agreeableness					0.23	0.20	0.14	.253
Extraversion					-0.23	0.22	-0.13	.287
Affiliation					9.93	13.88	0.07	.476
Dominance					-0.23	12.26	0.00	.985
Step 3	.131	.036	2.75(8,85)	.065				.009
Constant					149.1	42.74		< .001
Gender					-6.41	4.72	-0.14	.178
Age					-3.74	1.27	-0.31	.004
Agreeableness					0.19	0.20	0.12	.331
Extraversion					-0.28	0.21	-0.15	.199
Affiliation					1.22	14.41	0.01	.933
Dominance					3.49	12.12	0.03	.774
Overall complementarity affiliation					0.09	0.04	0.25	.020
Overall complementarity dominance					0.00	0.02	0.00	.979
Step 4	.184	.053	3.10(10,83)	.027				.002
Constant					141.20	42.17		.001
Gender					-7.94	4.66	-0.18	.092
Age					-3.83	1.24	-0.32	.003
Agreeableness					0.18	0.19	0.11	.353
Extraversion					-0.29	0.21	-0.16	.167
Affiliation					2.05	13.98	0.01	.884
Dominance					-4.65	12.38	-0.04	.708
Overall complementarity affiliation					0.06	0.04	0.16	.131
Overall complementarity dominance					0.04	0.02	0.19	.146
Momentary complementarity affiliation					-15.13	11.70	-0.13	.200
Momentary complementarity dominance					-24.30	10.35	-0.30	.021

Note. N = 94. Gender is coded as follows: male = 1, female = 2. Overall complementarity = a dyad's total deviation from perfect complementarity.

Table 12

Multiple Regression to Predict Peer Rated Task Performance

	R^2 *adj*	ΔR^2 *adj.*	*F(df)*	Sig. *F* change	*B*	*SE B*	β	*p*
Step 1	.016		1.77(2,91)					.176
Constant					2.61	1.09		.019
Gender					0.29	0.16	0.19	.073
Age					-0.01	0.04	-0.02	.841
Step 2	.027	.011	1.43(6,87)	.297				.213
Constant					2.57	1.48		.087
Gender					0.35	0.16	0.23	.035
Age					0.01	0.05	0.01	.906
Agreeableness					0.00	0.01	-0.01	.942
Extraversion					-0.01	0.01	-0.11	.346
Affiliation					0.35	0.49	0.07	.481
Dominance					0.78	0.44	0.19	.078
Step 3	.037	.010	1.45(8,85)	.237				.188
Constant					1.78	1.55		.254
Gender					0.42	0.17	0.27	.016
Age					0.02	0.05	0.04	.712
Agreeableness					0.00	0.01	0.01	.960
Extraversion					-0.01	0.01	-0.15	.212
Affiliation					0.42	0.52	0.09	.423
Dominance					0.80	0.44	0.20	.072
Overall complementarity affiliation					0.00	0.00	0.07	.523
Overall complementarity dominance					0.00	0.00	0.17	.127
Step 4	.055	.018	1.54(10,83)	.171				.139
Constant					1.91	1.56		.224
Gender					0.36	0.17	0.24	.038
Age					0.01	0.04	0.02	.833
Agreeableness					0.00	0.01	0.01	.921
Extraversion					-0.01	0.01	-0.13	.280
Affiliation					0.45	0.51	0.10	.385
Dominance					0.73	0.46	0.18	.114
Overall complementarity affiliation					0.00	0.00	0.00	.992
Overall complementarity dominance					0.00	0.00	0.24	.088
Momentary complementarity affiliation					-0.73	0.43	-0.19	.094
Momentary complementarity dominance					-0.29	0.38	-0.11	.443

Note. $N = 94$. Gender is coded as follows: male = 1, female = 2. Overall complementarity = a dyad's total deviation from perfect complementarity.

Discussion

High-fidelity simulations offer great potential for research and practice to observe interpersonal behavior of participants and interpersonal dynamics between participants and other human actors, such as role-players. Unfortunately, however, our knowledge about interpersonal dynamics in high-fidelity simulations is still scarce because past studies relied on single-point estimates of interpersonal behavior which exclude appropriate investigations of dynamics as they unfold across time in a given simulation. This study overcomes this limitation by applying a continuous assessment of interpersonal dynamics to investigate the interpersonal behavior of participants and role-players as well as their interpersonal dynamics in four distinct high-fidelity simulations. Results show that (a) participants and role-players show intraindividual variability in their interpersonal behavior, (b) intraindividual variability in interpersonal behavior of participants and role-players is entrained in a way that momentary interpersonal dynamics between role-players and participants can be described with the principles of complementarity, and (c) participants' degree of complementarity with different interaction partners across different high-fidelity simulations predicts job-related performance in terms of interpersonal adaptability and task performance seven months after the high-fidelity simulations.

Implications for Theory

As a first contribution, our results show that both participants and role-players show intraindividual variability in interpersonal behavior within high-fidelity simulations across time. This is in line with past research in social and clinical psychology that acknowledged and found evidence for continuous intraindividual variability in interpersonal behavior within a given situation across time (e.g., Hopwood et al., 2018; Markey et al., 2010; Pennings et al., 2014; Sadler et al., 2009, 2015; Strong et al., 1988; Tracey, 1994; Tracey, 2004). Despite the fact that this was already established across different contexts in social and clinical

psychological (lab) settings, intraindividual variability within a given situation across time is remarkable in high-fidelity simulations for selection and development purposes for both participants and role-players. For participants, this is remarkable for at least two reasons. First, high-fidelity simulations might be construed as posing high situational strength upon participants to generally show high expressions of affiliation (see, for example, Meyer et al., 2010; Moskowitz et al., 2007; Oliver et al., 2016). Despite this generally accepted assumption, we find evidence that participants do nonetheless show variations in their level of affiliation across time within four different high-fidelity simulations. The same results were obtained for dominance: Our results show that within each of the four high-fidelity simulations, participants vary in their level of dominance across time. Second, our results complement and extend the recently formed agreement that participants vary their behavior *across* different high-fidelity simulations (e.g., Gibbons & Rupp, 2009; Jackson et al., 2016; Lance, 2008; Lievens, 2009; Putka & Hoffman, 2013). Our results thus build upon the acknowledgement of intraindividual variability but extend it to the perspective of variability *within* high-fidelity simulations. For role-players, intraindividual variability in interpersonal behavior is also remarkable. That is because role-players often receive instructions to display specific interpersonal behavior and prompts in high-fidelity simulations (Lievens et al., 2015; Schollaert & Lievens, 2011, 2012). As shown by our results however, role-players are not stationary in their interpersonal behavior within a high-fidelity simulation, but do vary their expressions of affiliation and dominance across time within a given simulation.

As a second contribution, we add knowledge about the nature of interpersonal dynamics between participants and role-players in high-fidelity simulations. Our results show that such interpersonal dynamics can be described via the principles of complementarity, but only if they are investigated at the appropriate level of momentary behavior via continuous assessments. That is, in sum, we found limited evidence for the principles of complementarity

as possible describing heuristic for interpersonal dynamics between participants and role-players in high-fidelity simulations when interpersonal dynamics are examined at the overall level that aggregates the continuous stream of interpersonal behavior into one overall score per dimension.

For affiliation, correspondence in affiliation explained these interpersonal dynamics in three but not all four role-plays. Such inconsistent evidence for the principle of correspondence in affiliation is in line with past empirical evidence for lower correspondence in affiliation in work than in nonwork settings (Moskowitz et al., 2007) and results from Oliver et al. (2016) who showed that participants invest even more in relationship-building behavior when faced with unfriendly role-players. As a possible explanation, we concur with others that high-fidelity simulations introduce strong task and organizational demands that might reduce the tendency to show behavior that expresses low affiliation independent of interindividual differences or the interpersonal behavior of the role-player that participants interact with (see, for example, Moskowitz et al., 2007; Oliver et al., 2016). In turn, participants might rather show interpersonal behavior that indicates high expressions of affiliation overall. This explanation is supported by the fact that participants overall expression of affiliation was positive in all four role-plays and that no support for correspondence in affiliation was found in one of the role-plays with a role-player who was instructed to act unfriendly overall.

For dominance, reciprocity in dominance was found in none of the four role-plays. Instead, participants' and role-players' overall degree of dominance was unrelated to each other in three role-plays and positively related in one role-play. In this role-play, more dominant expressions of role-players provoked also more dominant expressions from participants. Such a lack of evidence for complementarity is in contrast to results from Moskowitz et al. (2007) who found even more evidence for reciprocity in dominance in work

than in nonwork settings. As a possible explanation, real-life work settings might usually involve pre-established relationships with more clearly assigned roles (Moskowitz et al., 2007), that might be related to more reciprocity in dominance than work settings that are captured in high-fidelity simulations. That is, in high-fidelity simulations, participants are in the spotlight and are asked to solve the problems they are confronted with. Even though this might also involve periods of more submissive behaviors, such as listening closely to relevant information revealed by role-players who actively lead the discussion, overall, participants might be expected to express more dominant behaviors such as actively asking targeted questions, expressing personal opinions, and making requests or suggestions to solve the problem. As tentative support for this argument, participants showed positive values for dominance across both measures (SBI and CAID) across all four role-plays with exception of one role-play in which the SBI score indicates a rather neutral overall behavior in terms of dominance. Hence, irrespective of the overall expression of dominance of role-players, these aspects might explain why participants show overall rather dominant behavior in high-fidelity simulations and why we do not find evidence for reciprocity in dominance at the overall level.

Our results are therefore in contrast with results from Oliver et al. (2016) who found a negative relation between role-players' expression of affiliation and participants' relationship-building behavior across multiple high-fidelity simulations as well as between role-players' expression of dominance and participants' directive communication. These contradicting results might be explained by at least two methodological reasons. First, Oliver et al. (2016) investigated the principles of complementarity not *within* high-fidelity simulations, but *across* multiple simulations. Oliver et al. (2016) therefore examined whether participants' behavior across different high-fidelity simulations with different role-players followed the principles of complementarity. It thus differs from our approach to investigate whether the interpersonal dynamics between participants and role-players can be describe with the principles of

complementarity *within* each, single high-fidelity simulation. Second, Oliver et al. (2016) examined participants' behavior not in terms of the broad dimensions of affiliation and dominance, but in terms of the more specific dimensions of relationship-building behavior and directive communication. For example, it is thus possible that participants' interpersonal behavior in terms of these more specific dimensions followed the principle of reciprocity in dominance, but that other components of dominance that have been integrated into our broader measurements did not consistently follow the principle of reciprocity in dominance at the overall level. In sum, at the overall level, no consistent pattern emerged that served to describe interpersonal dynamics between participants and role-players in high-fidelity simulations.[20]

However, this picture changed when we acknowledged the intraindividual variability in interpersonal behavior at the momentary level of both participants and role-players. That is, we found significant positive cross-correlations between participants' and role-players momentary affiliation and significant negative cross-correlations between participants' and role-players' momentary dominance. Thus, when zooming into the momentary level of interpersonal behavior, the principles of correspondence in affiliation and reciprocity in dominance consistently describe the interpersonal dynamics between participants and role-players in high-fidelity simulations.

Our results therefore extend the evidence for complementarity found in lab settings and more natural interactions (Sadler et al., 2011) to high-fidelity simulations. Despite the strong task and organizational norms that might limit extreme forms of low affiliation in

[20] One might wonder whether range-restriction on behalf of role-players' interpersonal behavior across interactions with different participants might explain why we do not find (consistent) evidence for complementarity at the overall level. That is, role-players have been instructed and trained to display a consistent interpersonal disposition across interactions with different participants. However, standard deviations for overall interpersonal behavior of role-players appear to be comparable to the ones for participants. This means that role-players differ their interpersonal behavior across interactions with different participants similarly to the variation of interpersonal behavior of different participants. Thus, range-restriction in interpersonal behavior on behalf of role-players does not appear to explain the lack of consistent evidence for complementarity at the overall level.

work-related contexts (see, for example, Moskowitz et al., 2007; Oliver et al., 2016) or that might trigger participants to show overall more dominant behavior, we found evidence that role-players and participants mutually adapt their momentary level of interpersonal behavior to each other. Hence, the principles of complementarity appear to be powerful heuristics for continuous interpersonal dynamics across time even in high-fidelity simulations that sample work-related situations that are shaped by strong task and organizational situational demands.

As a third contribution, our results outline the relevance of interpersonal dynamics within high-fidelity simulations because interpersonal dynamics in terms of the principles of complementarity were found to predict performance ratings in high-fidelity simulations as well as job-related performance outside of the context of high-fidelity simulations.

For relations between complementarity and role-play performance, there is a caveat in order: For both overall complementarity and momentary complementarity, relations to performance ratings in single role-plays were inconsistent or absent. It thus appears like situation-specific task demands might influence whether complementarity is related to performance ratings or not. Further, we must again refer to the fact that complementarity is a dyadic phenomenon and is thus both influenced by the participants and role-players. Therefore, whether complementarity emerges in a single simulation, is not only due to the participant's behavior, but also due to the role-player's behavior. This might at least partially explain why complementarity and performance ratings are inconsistently or unrelated at the level of single high-fidelity simulations.

However, if some participants show higher degrees of complementarity than other participants across multiple high-fidelity simulations in which they interact with multiple different role-players, such an average degree of complementarity can more clearly be attributed to an individual's tendency to adapt interpersonal behavior to different interaction partners and create interactions that follow the principles of complementarity (see, for

example, Oliver & Lievens, 2014; Pincus et al., 2014; Sadler et al., 2011 for similar arguments). In line with these arguments, complementarity averaged across multiple high-fidelity simulations showed more consistent results.

For overall complementarity, participants who show higher degrees of complementarity in affiliation and dominance do receive lower performance ratings. For dominance, as mentioned above, high-fidelity simulations might overall demand participants to show more dominant behavior irrespective of role-players' overall expressions of dominance. Thus, participants who overall follow the principle of reciprocity in dominance across different high-fidelity simulations, show overall expressions of more submissive behavior when faced with role-players who show overall expressions of more dominant behavior. Hence, they might receive less favorably performance ratings because they do not show a sufficient amount of dominant behaviors, such as actively asking targeted questions, expressing personal opinions, and making requests or suggestions that benefit to solve the problems in the high-fidelity simulations. This matches the result of Oliver et al. (2016) who showed a positive relation between directive communication on behalf of participants and performance ratings in high-fidelity simulations with dominant role-players.

For affiliation, this is in line with our expectation that participants who show correspondence across multiple different high-fidelity simulations with role-players varying in their expressions of overall affiliation are discredited because they would correspond to overall low expressions of affiliation equally with expressions of overall low affiliation. In contrast to natural interactions outside of the work context, in which correspondence in affiliation appears to serve a form of self-validation, even in situations with correspondence in unfriendly behavior (e.g., Carson, 1969; Kiesler, 1983; Sadler et al., 2011), correspondence in affiliation that involves expressions of low affiliation is usually not appreciated in work settings (see Moskowitz et al., 2007; Oliver et al., 2016).

For momentary complementarity, the picture changes again. There appears a trend that both complementarity in affiliation and dominance are related to higher performance ratings. Note that this trend was only marginally significant for affiliation and for dominance it was only visible for role-player ratings. We can thus not exclude the possibility that at least the relation between momentary complementarity in dominance and role-play performance ratings are due to a bias. That is, it has been well established that complementarity benefits satisfaction with interactions and liking of interaction partners (see, Sadler et al., 2011). Role-players might therefore implicitly reward participants for momentary complementary in high-fidelity simulations although momentary complementarity might not tap into the essentials of performance in the high-fidelity simulations.

Further, complementarity adds to our understanding of performance ratings from role-players beyond other interpersonal predictors like self-rated personality and overall affiliation and dominance. However, this is only true for role-player rated performance. This strengthens the view that role-player performance ratings are at least partly biased by the rewarding and satisfying nature of complementarity that is not related to true performance in high-fidelity simulations.

In line with our hypotheses, complementarity served to predict job-related performance outside of the setting of high-fidelity simulations. That is, in line with our hypothesis, momentary complementarity in affiliation positively predicted interpersonal adaptability rated by instructors seven months after the high-fidelity simulations. Momentary complementarity further added incremental validity in the prediction of interpersonal adaptability beyond controls and other interpersonal predictors. This supports our reasoning that the principles of complementarity might serve as a new angle to measure interpersonal adaptability.

Despite the well-acknowledged relevance of interpersonal adaptability (e.g., Griffin et al., 2007; Pulakos et al., 2000), the measurement of interpersonal adaptability is still in its infancy (see Oliver & Lievens, 2014). That is because interpersonal adaptability is usually assessed via self- or other-report questionnaires that assess a general tendency to adapt one's behavior to interaction partners (Charbonnier-Voirin & Roussel, 2012; Griffin et al., 2007; Ployhart & Bliese, 2006; Pulakos et al., 2000). However, we agree with Oliver and Lievens (2014) that the nature of interpersonal adaptability might best be assessed via dynamic, situational measures that confront test-takers with different interpersonal situations and assess how test-takers behave in each of these situations. Our results support the idea that Interpersonal Theory might serve as an excellent theoretical fundament for such measurement approaches, because it both provides a theoretical framework of interpersonal situations (i.e., interaction partners' levels of affiliation and dominance) as well as theoretically and empirically supported principles (i.e., the principles of complementarity) which interpersonal behavior might best match the given situation (Oliver & Lievens, 2014; see also Pincus et al., 2014; Sadler et al., 2011 for similar arguments).

Overall complementarity in dominance was, however, found to be negatively related to instructor ratings of interpersonal adaptability. This might further indicate that complementarity at the overall level might be a biased estimate of the true interpersonal dynamics. On top of that and in line with our arguments from above, overall complementarity in high-fidelity simulations might point towards not appropriately adapting one's interpersonal behavior to the combination of interpersonal, task, and organizational demands and norms.

As further support for our hypotheses, momentary complementarity in dominance positively predicted instructor rated task performance. Momentary complementarity did further add incremental variance in the prediction of instructor rated task performance beyond

control variables and other interpersonal predictors. This result is in line with the notion that reciprocity in dominance implies a common agreement upon each individual's status and assignment of roles within interactions (Carson, 1969; Kiesler, 1983) that is, in turn, meant to facilitate efficient task performance (see, for example, Bendersky & Hays, 2012; Locke & Sadler, 2007).

Complementarity in affiliation at the overall and momentary level was in turn negatively related to task performance rated by instructors and peers. Based upon our results, it might be that correspondence in affiliation is more relevant for more interpersonally driven criteria such as interpersonal adaptability. For task-related criteria, striving for correspondence in affiliation might sometimes distract from successfully fulfilling the current task duties.

As a fourth contribution, our study further attests to the relevance of continuous assessments (see Gabriel et al., 2017; Jebb & Tay, 2017). That is, in our study, we could show that intraindividual variability is evident in high-fidelity simulations at the continuous level. Further, we were able to demonstrate that this intraindividual variability at the continuous level is not random error but mirrors substance. That is, only analyses at the momentary level revealed consistent patterns of interpersonal dynamics between participants and role-players that further predicted performance ratings in high-fidelity simulations as well as job-related performance ratings beyond self-ratings of personality and single-point estimates of interpersonal behavior.

Limitations

Some limitations of this study need to be noted. First, the interrater reliabilities for our CAID codings were low. Average cross-correlations between time series of different coders for the same target were $r = .22$ for affiliation and $r = .36$ for dominance. At the level of individual coders, the reliability of momentary interpersonal behavior thus appears

disappointing. However, we aggregated time series for each target across four coders (average harmonic mean of coders per target: 3.92). Thereby, the principle of aggregation reduces the relative amount of coder-specific idiosyncrasies and generates a more reliable time series of interpersonal behavior (see, for example, Eisenkraft, 2013).

Second, the construct-related validity evidence for overall interpersonal behavior derived from the CAID is not consistent across all four role-plays and for both dimensions of interpersonal behavior. That is, we did not find evidence for significant positive correlations between corresponding SBI and CAID scales for affiliation in two role-plays and for dominance in one role-play. As a potential resolution, we had to select a limited number of items for each of the SBI scales that were then rated by the role-players. Although we selected SBI items per dimension that were most relevant in each of the role-plays, it is possible that several interpersonal behaviors within the role-plays were not captured by the selected SBI items. In contrast, CAID coders based their codings upon all verbal and nonverbal expressions of interpersonal behavior of participants. Hence, CAID codings might have more fully captured the rich variety of interpersonal behavior within the role-plays. This difference in breadth of construct coverage might explain the inconsistent results of construct-related validity for our CAID codings. However, for interpersonal behavior averaged across all four role-plays, we found significant positive intercorrelations between corresponding CAID and SBI scales for both dimensions. Further, momentary interpersonal dynamics between participants and role-players followed the principle of complementarity and predicted job-related performance ratings in line with theoretically derived hypotheses. Thus, overall, we have no doubt about the construct-related validity of the CAID codings in our study.

Third, CAID codings did not fully support our manipulation to confront participants with four role-plays that capture all four quadrants of the Interpersonal Circumplex at the overall level. Although this manipulation was successful for affiliation with two role-plays

confronting participants with a friendly and two role-plays confronting participants with an unfriendly role-player, role-players overall showed dominant behavior across all four role-plays. However, confidence intervals indicated that participants were confronted with three distinct degrees of dominance across the four role-plays. Further, theoretical arguments as well as our results show that interpersonal behavior and dynamics are most important at the momentary level. Given that we find consistent evidence for intraindividual variability in interpersonal behavior of role-players, we have support that participants were confronted with different interpersonal demands in our study.

Fourth, several variables have not been assessed in our studies that might work as mediators between complementarity and outcomes. We thus encourage future research to assess possible mediators such as ratings of liking of participants and the satisfaction with interactions on behalf of role-players as well as instructors and peers who provide criterion ratings, or status- and role-assignments in job-related situations.

Implications for Practice

Our results attest to the relevance of complementarity in high-fidelity simulations because they contribute to explain performance ratings in high-fidelity simulations and predict job-related performance ratings. Although we call for further replications of our results, practitioners might thus assess participants' individual tendency to show complementarity in high-fidelity simulations. Given that complementarity is a dyadic phenomenon, though, we urge caution and recommend to assess complementarity only across various high-fidelity simulations in which participants interact with different role-players. An economic way to do so might be to assess complementarity in Multiple Speed Assessments (Herde & Lievens, 2018). In Multiple Speed Assessments, participants involve in multiple, short, interpersonal simulations that are introduced by different role-players. Multiple Speed Assessments might thus serve to assess the degree of complementarity that participants show with many different

human actors in a short amount of time (e.g., 20 simulations in 1 hour). In this way, complementarity across multiple short high-fidelity simulations could be utilized as a new angle to the assessment of interpersonal adaptability. Complementarity might thus complement previous self- and other reports of interpersonal adaptability in practice.

Our results further outline that practitioners might best apply continuous assessments of interpersonal behavior because there is meaningful intraindividual variability in interpersonal behavior that drives and shapes interpersonal dynamics between participants and role-players. We thus concur with others who recommend the use of continuous assessments in organizational science and practice (Gabriel et al., 2017; Jebb & Tay, 2017). Although we acknowledge that continuous assessments consume enormous efforts, our results showcase that these efforts pay off in terms of gained insights. That is, investigations of interpersonal behavior appear to draw a more accurate picture of the nature of interpersonal dynamics and interpersonal dynamics at the momentary level add incremental variance to the prediction of performance ratings in high-fidelity simulations and in job-related situations.

Future Research Avenues

Future research might investigate the utility of our results for developmental interventions. That is, our results outline that interpersonal behavior and dynamics are best captured at the momentary behavioral level. This provides good opportunity for developmental interventions because interventions that focus on actual behavior have been shown to produce favorable training effects (e.g., Burke & Day, 1986; Taylor, Russ-Eft, & Chan, 2005). Participants might involve in interactions with different role-players in high-fidelity simulations and videos could be recorded. Afterwards, participants might take part in a training about Interpersonal Theory and the principles of complementarity as well as their contribution to successful interactions. Participants might then review the video tapes together with a coach. In line to decompose the different processes that influence the degree of

complementarity shown (Sadler et al., 2011; see also Oliver & Lievens, 2014), it could then be analyzed (a) whether participants had sent adequate interpersonal signals to their interaction partners, (b) how participants had perceived their interaction partners' interpersonal behavior at any particular moment, (c) how participants reacted to their interaction partners' interpersonal behavior. In this way, one might also analyze the benefit of CAID codings for coders' interpersonal perception skills. The coach could then provide feedback about all these aspects and work with the participant to improve the degree of complementarity that participants establish.

Future research might further investigate interpersonal dynamics in other high-fidelity simulations than role-plays. For example, it might be interesting to investigate interpersonal dynamics in leaderless group discussions because they involve more human actors than role-plays. Interpersonal dynamics between all group members might be more complicated, but might also provide further insights about participants' job-related behavior.

Conclusion

This study investigated interpersonal behavior and interpersonal dynamics between participants and role-players in high-fidelity simulations. We found that participants and role-players show intraindividual variability in their interpersonal behavior and that the principles of complementarity consistently describe interpersonal dynamics in high-fidelity simulation, but only if the level of measurement accounts for the intraindividual variability in interpersonal behavior. Further, participants who show stronger degrees of complementarity in high-fidelity simulations turn out to receive higher ratings of job-related performance criteria such as interpersonal adaptability and task-performance. In other words, when it comes to interpersonal behavior and interpersonal dynamics, there is no best time to investigate it: The magic happens continuously.

References

Ansell, E. B., Kurtz, J. E., & Markey, P. M. (2008). Gender differences in interpersonal complementarity within roommate dyads. *Personality and Social Psychology Bulletin*, *34*, 502–512. https://doi.org/10.1177/0146167207312312

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

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, *83*, 377–391. http://dx.doi.org/10.1037/0021-9010.83.3.377

Barry, B., & Stewart, G. L. (1997). Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology*, *82*, 62–78. https://doi.org/10.1037/0021-9010.82.1.62

Bendersky, C., & Hays, N. A. (2012). Status conflict in groups. *Organization Science*, *23*, 323–340. https://doi.org/10.1287/orsc.1110.0734

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., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology*, *74*, 957–963. https://doi.org/10.1037/0021-9010.74.6.957

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology*, *71*, 232–245. http://dx.doi.org/10.1037/0021-9010.71.2.232

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

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

Charbonnier-Voirin, A., & Roussel, P. (2012). Adaptive performance: A new scale to measure individual performance in organizations. *Canadian Journal of Administrative Science, 29*, 280–293. https://doi.org/10.1002/cjas.232

Dryer, D. C., & Horowitz, L. M. (1997). When do opposites attract? Interpersonal complementarity versus similarity. *Journal of Personality and Social Psychology, 72*, 592–603. https://doi.org/10.1037/0022-3514.72.3.592

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

Estroff, S., & Nowicki, S. (1992). Interpersonal complementarity, gender of interactants, and performance on puzzle and word tasks. *Personality and Social Psychology Bulletin*, *18*, 351–356. https://doi.org/10.1177/0146167292183012

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology*, *24*, 387–407. https://doi.org/10.1007/s10869-009-9123-3

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

Gabriel, A. S., Diefendorff, J. M., Bennett, A. A., & Sloan, M. D. (2017). It's about time: The promise of continuous rating assessments for the organizational sciences. *Organizational Research Methods*, *20*, 32–60. https://doi.org/10.1177/1094428116673721

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

Glomb, T. M., & Welsh, E. T. (2005). Can opposites attract? Personality heterogeneity in supervisor-subordinate dyads as a predictor of subordinate outcomes. *Journal of Applied Psychology, 90,* 749–757. https://doi.org/10.1037/0021-9010.90.4.749

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

Grant, A. M., Gino, F., & Hofmann, D. A. (2011). Reversing the extraverted leadership advantage: The role of employee proactivity. *Academy of Management Journal, 54,* 528–550. https://doi.org/10.5465/amj.2011.61968043

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50,* 327–347. https://doi.org/10.5465/amj.2007.24634438

Gurtman, M. B., & Lee, D. L. (2009). Sex differences in interpersonal problems: A circumplex analysis. *Psychological Assessment, 21,* 515–527. https://doi.org/10.1037/a0017085

Herde, C. N., & Lievens, F. (2018). Multiple Speed Assessments: Theory, practice, & research evidence. *European Journal of Psychological Assessment,* Advance online article. https://doi.org/10.1027/1015-5759/a000512

Hopwood, C. J., Harrison, A. L., Amole, M., Girard, J. M., Wright, A. G. C., Thomas, K. M., … Kashy, D. A. (2018). Properties of the Continuous Assessment of Interpersonal Dynamics across sex, level of familiarity, and interpersonal conflict. *Assessment*, 107319111879891. https://doi.org/10.1177/1073191118798916

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

Jansen, A., Lievens, F., & Kleinmann, M. (2011). Do individual differences in perceiving situational demands moderate the relationship between personality and Assessment Center dimension ratings? *Human Performance*, *24*, 231–250. https://doi.org/10.1080/08959285.2011.580805

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

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*, 61–94. https://doi.org/10.1177/1094428116668035

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

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, *78*, 988–993. http://dx.doi.org/10.1037/0021-9010.78.6.988

Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of Assessment Centers: A conceptual review. *Annual Review of Organizational Psychology and Organizational Behavior, 6,* 349–372. https://doi.org/10.1146/annurev-orgpsych-012218-014955

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1,* 128–146. https://doi.org/10.1177/2041386610387000

Krause, D. E., & Gebert, D. (2003). A comparison of Assessment Center practices in organizations in German-speaking regions and the United States. *International Journal of Selection and Assessment, 11,* 297–312. https://doi.org/10.1111/j.0965-075X.2003.00253.x

Krause, D. E., Rossberger, R. J., Dowdeswell, K., Venter, N., & Joubert, T. (2011). Assessment Center practices in South Africa. *International Journal of Selection and Assessment, 19,* 262–275. https://doi.org/10.1111/j.1468-2389.2011.00555.x

Krause, D. E., & Thornton III, G. C. (2009). A cross-cultural look at Assessment Center practices: Survey results from Western Europe and North America. *Applied Psychology, 58,* 557–585. https://doi.org/10.1111/j.1464-0597.2008.00371.x

Kristof-Brown, A., Barrick, M. R., & Stevens, C. K. (2005). When opposites attract: A multi-sample demonstration of complementary person-team fit on extraversion. *Journal of Personality, 73,* 935-957. https://doi.org/10.1111/j.1467-6494.2005.00334.x

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

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. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*, 102–121. https://doi.org/10.1080/13594320802058997

Lievens, F., & Klimoski, R. J. (2001). Understanding the Assessment Center process: Where are we now? In C. L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245–286). Chicester: John Wiley & Sons, Ltd.

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 Management* (Vol. 28, pp. 99–152). Bingley: Emerald Group Publishing.

Locke, K. D., & Sadler, P. (2007). Self-efficacy, values, and complementarity in dyadic interactions: Integrating interpersonal and social-cognitive theory. *Personality and Social Psychology Bulletin*, *33*, 94–109. https://doi.org/10.1177/0146167206293375

Markey, P. M., Funder, D. C., & Ozer, D. J. (2003). Complementarity of interpersonal behaviors in dyadic interactions. *Personality and Social Psychology Bulletin*, *29*, 1082–1090. https://doi.org/10.1177/0146167203253474

Markey, P. M., & Kurtz, J. E. (2006). Increasing acquaintanceship and complementarity of behavioral styles and personality traits among college roommates. *Personality and Social Psychology Bulletin*, *32*, 907–916. https://doi.org/10.1177/0146167206287129

Markey, P., Lowmaster, S., & Eichler, W. (2010). A real-time assessment of interpersonal complementarity. *Personal Relationships*, *17*, 13–25. https://doi.org/10.1111/j.1475-6811.2010.01249.x

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

Markey, P. M., & Markey, C. N. (2007). Romantic ideals, romantic obtainment, and relationship experiences: The complementarity of interpersonal traits among romantic partners. *Journal of Social and Personal Relationships*, *24*, 517–533. https://doi.org/10.1177/0265407507079241

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

McEvoy, G. M., & Cascio, W. E. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology*, *74*, 11–17. https://doi.org/10.1037/0021-9010.74.1.11

McFarland, L., Ryan, A. M., & Kriska, S. D. (2003). Impression management use and effectiveness across assessment methods. *Journal of Management, 29*, 641–661. https://doi.org/10.1016/S0149-2063(03)00030-8

McFarland, L., Yun, G. J., Harold, C. M., Viera, L., & Moore, L. G. (2005). An examination of impression management use and effectiveness across assessment center exercises:

The role of competency demands. *Personnel Psychology*, *58*, 949–980. https://doi.org/10.1111/j.1744-6570.2005.00374.x

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

Moskowitz, D. S. (1994). Cross-situational generality and the interpersonal circumplex. *Journal of Personality and Social Psychology*, *66*, 921–933. http://dx.doi.org/10.1037/0022-3514.66.5.921

Moskowitz, D. S., Ho, M.-h. R., & Turcotte-Tremblay, A.-M. (2007). Contextual influences on interpersonal complementarity. *Personality and Social Psychology Bulletin*, *33*, 1051–1063. https://doi.org/10.1177/0146167207303024

Moskowitz, D. S., & Sadikaj, G. (2011). Event-contingent recording. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 160–175). New York: Guilford Press.

Moskowitz, D. S., & Zuroff, D. C. (2005). Assessing interpersonal perceptions using the Interpersonal Grid. *Psychological Assessment*, *17*, 218–230. https://doi.org/10.1037/1040-3590.17.2.218

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

Nowicki, S., & Manheim, S. (1991). Interpersonal complementarity and time of interaction in female relationships. *Journal of Research in Personality*, *25*, 322–333. https://doi.org/10.1016/0092-6566(91)90023-J

O'Connor, B. P., & Dyce, J. (1997). Interpersonal rigidity, hostility, and complementarity in musical bands. *Journal of Personality and Social Psychology*, *72*, 362–372. http://dx.doi.org/10.1037/0022-3514.72.2.362

Oliver, T. (2012). *Applying a framework of interpersonal adaptability for assessment* (doctoral dissertation). Retrieved from https://atrium2.lib.uoguelph.ca/xmlui/handle/10214/4960

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. https://doi.org/10.1177/0149206314525207

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.

Pennings, H. J. M., van Tartwijk, J., Wubbels, T., Claessens, L. C. A., van der Want, A. C., & Brekelmans, M. (2014). Real-time teacher–student interactions: A dynamic systems approach. *Teaching and Teacher Education, 37*, 183–193. https://doi.org/10.1016/j.tate.2013.07.016

Pincus, A. L., Sadler, P., Woody, E., Roche, M. J., Thomas, K. M., & Wright, A. G. C. (2014). Multimethod assessment of interpersonal dynamics. In C. J. Hopwood & R. F. Bornstein (Eds.), *Multimethod clinical assessment* (pp. 51–91). New York: Guilford.

Ployhart, R. E., & Bliese, P. D. (2006). Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In S. Burke, L. Pierce, & Salas (Eds.), *Understanding adaptability: A prerequisite for effective performance within complex environments* (pp. 3–39). Bingley: Emerald Publishing Group

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624. https://doi.org/10.1037//0021-9010.85.4.612

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

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

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology, 97,* 1005–1020. https://doi.org/10.1037/a0016232

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.

Sadler, P., & Woody, E. (2003). Is who you are who you're talking to? Interpersonal style and complementarily in mixed-sex interactions. *Journal of Personality and Social Psychology*, *84,* 80–96. https://doi.org/10.1037/0022-3514.84.1.80

Sadler, P., & Woody, E. (2016). *Manual for the Continuous Assessment of Interpersonal Dynamics (CAID) Joystick Monitor Program*.

Sadler, P., Woody, E., McDonald, K., Lizdek, I., & Little, J. (2015). A lot can happen in a few minutes: Examining dynamic patterns within an interaction to illuminate the

interpersonal nature of personality disorders. *Journal of personality disorders, 29,* 526–546. https://doi.org/10.1521/pedi.2015.29.4.526

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

Shechtman, N., & Horowitz, L. M. (2006). Interpersonal and noninterpersonal interactions, interpersonal motives, and the effect of frustrated motives. *Personality and Social Psychology Bulletin, 32,* 1126–1139. https://doi.org/10.1177/0146167206288669

Smelser, W. T. (1961). Dominance as a factor in achievement and perception in cooperative problem solving interactions. *Journal of Abnormal and Social Psychology, 62,* 535–542. https://doi.org/10.1037/h0049303

Strong, S. R., Hills, H. I., Kilmartin, C. T., DeVries, H., Lanier, K., Nelson, B. N., … Meyer III, C. W. (1988). The dynamic relations among interpersonal behaviors: A test of complementarity and anticomplementarity. *Journal of Personality and Social Psychology, 54,* 798-810. http://dx.doi.org/10.1037/0022-3514.54.5.798

Sullivan, H. S. (1953). *The interpersonal theory of psychiatry.* New York: Norton.

Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology, 90,* 692–709. https://doi.org/10.1037/0021-9010.90.4.692

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

Thomas, K. M., Hopwood, C. J., Woody, E., Ethier, N., & Sadler, P. (2014). Momentary assessment of interpersonal process in psychotherapy. *Journal of Counseling Psychology*, *61,* 1–14. https://doi.org/10.1037/a0034277

Thornton III, G. C., & Rupp, D. E. (2006). *Assessment Centers in Human Resource Management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum Publishers.

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

Tracey, T. J. (1994). An examination of the complementarity of interpersonal behavior. *Journal of Personality and Social Psychology, 67*, 864–878. https://doi.org/10.1037/0022-3514.67.5.864

Tracey, T. J. G. (2004). Levels of interpersonal complementarity: A simplex representation. *Personality and Social Psychology Bulletin, 30,* 1211–1225. https://doi.org/10.1177/0146167204264075

Tracey, T. J. G., Albright, J. M., & Sherry, P. (1999). The interpersonal process of cognitive-behavioral therapy: An examination of complementarity over the course of treatment. *Journal of Counseling Psychology*, *46,* 80–91. http://dx.doi.org/10.1037/0022-0167.46.1.80

Tracey, T. J. G., Bludworth, J., & Glidden-Tracey, C. E. (2012). Are there parallel processes in psychotherapy supervision? An empirical examination. *Psychotherapy, 49,* 330–343. https://doi.org/10.1037/a0026246

Tracey, T. J. G., Ryan, J. M., & Jaschik-Herman, B. (2001). Complementarity of interpersonal circumplex traits. *Personality and Social Psychology Bulletin*, *27*, 786–797. https://doi.org/10.1177/0146167201277002

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

Wiggins, J. S. (1980). Circumplex models of interpersonal behavior. In L. Wheeler (Ed.), *Review of personality and social psychology* (pp. 265–294). Beverly Hills, CA: SAGE.

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

Yao, Q., & Moskowitz, D. S. (2015). Trait agreeableness and social status moderate behavioral responsiveness to communal behavior. *Journal of Personality*, *83*, 191–201. https://doi.org/10.1111/jopy.12094

Yaughn, E., & Nowicki, S. (1999). Close relationships and complementary interpersonal styles among men and women. *Journal of Social Psychology*, *139*, 473–478. https://doi.org/10.1080/00224549909598406

Yukl, G. (1999). An evaluative essay on current conceptions of effective leadership. *European Journal of Work and Organizational Psychology*, *8*, 33–48. https://doi.org/10.1080/135943299398429

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

CHAPTER 6: GENERAL DISCUSSION

Across time, the frequency and complexity of interpersonal interactions on the job has increased. This stresses the importance to perform well in interpersonal situations and to adapt to different interpersonal demands (Griffin, Neal, & Parker, 2007; Pulakos, Arad, Donovan, & Plamondon, 2000). Therefore, this dissertation addressed several objectives to add knowledge about how personnel selection and development procedures might help organizations to master this challenge. In particular, this dissertation (a) investigated whether low-fidelity simulations such as Situational Judgment Tests (SJTs) that assess procedural knowledge about working well with others and adapting to change can be developed in line with a combined emic-etic approach to gain SJTs that can be compared across geographical regions, (b) defined Multiple Speed Assessments as an umbrella term that encompasses various approaches that integrate multiple, short high-fidelity simulations and specified their common characteristics, theoretical fundament, as well as possible application areas, (c) provided an empirical investigation of the reliability and validity of a face-to-face format of Multiple Speed Assessments, and (d) explored interpersonal behavior as well as interpersonal dynamics in high-fidelity simulations at the continuous moment-to-moment level and investigated its relations to performance ratings in high-fidelity simulations, interpersonal adaptability, and task performance in interpersonal settings.

The following sections provide a short summary of the main findings of each of the chapters, reflect on implications for theory, highlight possible implications for practice, describe several limitations, provide a more general agenda for future research, and finally, state a final conclusion.

Main Findings

Objective 1: Provide an empirical test of the combined emic-etic approach to develop SJTs to measure procedural knowledge about interpersonal performance and (interpersonal) adaptability across geographical regions

To address this objective, chapter 2 investigated measurement invariance across participants from Latin America and Europe for five different SJTs that assessed procedural knowledge about one competency each that is crucial for success in entry-level jobs. One of these SJTs assessed procedural knowledge about "working well with others" and one of these SJTs assessed procedural knowledge about "adapting to change". The results showed evidence for configural and metric measurement invariance for all of the five SJTs across participants from Latin America and Europe. This means that the same factorial structure served to explain SJT scenario scores for both regional groups and that the SJT scenarios scores assessed the latent factor(s) equally across both regional groups (Byrne & Stewart, 2006; Byrne & van de Vijver, 2010). Thus, participants from Latin America and Europe construed the SJT scenarios and response options equally and attributed equal meaning to these scenarios and response options (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Objective 2: Provide a conceptual overview of Multiple Speed Assessments, including their shared characteristics, theoretical fundaments, possible design variations as well as application areas, and an agenda for future research

Chapter 3 defined Multiple Speed Assessments as an umbrella term for different assessment approaches that build upon multiple, short interpersonal simulations to gain insights into the behavioral repertoire of a target person in a predefined domain. In a review of different forms of Multiple Speed Assessments, it was proposed that the common characteristics of Multiple Speed Assessments are (a) applying multiple simulations (e.g., 20),

(b) applying short simulations (often less than 5 minutes), (c) structuring simulations via standardized role-player actions or statements (prompts) that are consistently displayed to participants, (d) streamlining evaluations of performance (e.g., only using one single evaluation of performance per simulation and using role-players as assessors), and (e) integrating the multiple short simulations into an overarching background.

Chapter 3 further outlined that the zero acquaintance and thin slices paradigm serve as theoretical fundament for the possibility to obtain reliable and valid ratings in short simulations (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992; Back & Nestler, 2016; Connelly & Ones, 2010; Funder, 2012), the trait-activation theory as theoretical fundament how different short simulations and role-player prompts activate individual differences in the domain that is to be sampled (Lievens, Tett, & Schleicher, 2009; Tett & Burnett, 2003), and the principle of aggregation as theoretical fundament for the reliability of performance ratings that are aggregated across multiple different simulations and assessors/role-plays (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014).

In addition, chapter 3 provided an overview of possible design variations of Multiple Speed Assessments. For example, stimuli can be presented to participants via dynamic audiovisual or face-to-face interactive stimuli. Participants' responses can then be recorded in an audiovisual constructed or face-to-face interaction format. Amongst others, possible types of simulations include (asynchronous) role-plays, clinical scenarios, interviews, fact-finding exercises, and short presentations. Further, many different domains, such as the interpersonal (leadership) domain, performance in (healthcare) study programs, or job-related behavior (e.g., of healthcare practitioners) can be sampled. Across different variations of Multiple Speed Assessments, the number of simulations varies between 3 and 40 with durations per simulation that often do not last longer than seven minutes on average (Cucina, Su, Busciglio,

Harris Thomas, & Thompson Peyton, 2015; Knorr & Hissbach, 2014; Lievens, De Corte, & Westerveld, 2015; Patrício, Julião, Fareleira, & Carneiro, 2013; Rees et al., 2016).

As the main purpose of Multiple Speed Assessments, chapter 3 identified the assessment of overall behavior across situations as well as the assessment of participants' intraindividual variability across situations in current and possible application areas such as personnel selection and development, educational settings, such as the certification and selection of medical students, research on short-term personality change and related interventions, as well as clinical applications.

Finally, chapter 3 formulated a research agenda on Multiple Speed Assessments. To gain further knowledge about Multiple Speed Assessments, further research is necessary about their (a) reliability, including investigations of interrater reliability and the contribution of different sources of reliable and unreliable variance to overall score variance, (b) validity, including relations to other forms of simulation-based assessment approaches and criterion-related validity to measures of job-performance, (c) applicant reactions, including perceptions of face validity and the opportunity to perform in Multiple Speed Assessments, and (d) subgroup differences, including investigations of differences due to age, gender, and ethnicity as well as investigations whether short simulations increase the relative influence of stereotypes in assessors' judgments.

Objective 3: Provide knowledge about the reliability and validity of a face-to-face format of Multiple Speed Assessments to sample the leadership domain which includes components of interpersonal performance and (interpersonal) adaptability

Chapter 4 investigated the reliability and validity of a face-to-face format of Multiple Speed Assessments that sampled the leadership domain and required to appropriately adapt (interpersonal) behavior to different role-players and situational demands. In terms of reliability, results showed that single-rater reliabilities were low to at best moderate. However,

the picture changed when we investigated interrater reliabilities averaged across three to four independent assessors for each role-play. Then, reliabilities for the averaged ratings were moderate to high. This pattern thus replicates a key result from the zero acquaintance/thin slices paradigm (e.g., Connelly & Ones, 2010): Although role-players and assessors in our study participated in an intensive training, used behavioral observation aids, and behavior elicitation and evaluation of participants' behavior was facilitated via standardized role-player prompts, ratings from independent assessors need to be aggregated to obtain reliable ratings (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014).

A generalizability study that decomposed different sources of variance further attested to this phenomenon, because the amount of reliable variance in our study only reached a value of 37 %. The generalizability study further indicated that the largest amount of reliable variance (61 %) was explained by a participant x simulation interaction effect. This source of variance thus descriptively contributed more to reliable variance than a participant main effect. The results of the generalizability study were further used to run a decision study that investigated how many simulations and ratings from independent assessors per simulation need to be aggregated to obtain an overall reliable estimate of performance in the leadership domain. Results showed that an overall $G \geq .70$ was obtained when ratings from at least fourteen simulations were aggregated across two independent assessors or if ratings from at least nine simulations were aggregated across three independent assessors.

In terms of validity, the results mirrored the pattern that was obtained for the single-rater reliabilities: Relations between performance ratings in single simulations on the one hand and measures of cognitive ability, personality, as well as criterion performance were highly inconsistent and showed overall low to moderate correlations. However, when performance ratings were aggregated across all simulations, performance ratings in Multiple Speed Assessments showed substantial relations to cognitive ability, extraversion, and

agreeableness. Further, such aggregated performance ratings across all simulations predicted instructor and peer ratings of performance. In fact, such an overall Multiple Speed Assessment performance score explained additional 14 % and 9 % variance in instructor and peer ratings of performance beyond other predictors that tapped into a similar domain.

Objective 4: Provide knowledge about the interpersonal behavior of participants and the interpersonal dynamics they establish with other human actors in high-fidelity simulations at the continuous moment-to-moment level as well as their relations to ratings of performance in high-fidelity simulations, interpersonal adaptability and task performance in interpersonal settings

Chapter 5 investigated intraindividual variability in interpersonal behavior as well as interpersonal dynamics between participants and role-players in four distinct high-fidelity simulations (role-plays). Results showed that the interpersonal behavior of both participants and role-players showed significant intraindividual variability within each of the four analyzed high-fidelity simulations. Further, it appeared that the intraindividual variability at the momentary level drove interpersonal dynamics between participants and role-players because participants and role-players adapted their interpersonal behavior at the momentary level in line with the two principles of complementarity. That is, following the principle of correspondence in affiliation, participants expressed higher degrees of affiliation at the momentary level when role-players expressed higher degrees of affiliation, and vice versa. On top of that, following the principle of reciprocity in dominance, participants expressed lower degrees of dominance at the momentary level when role-players expressed higher degrees of dominance and vice versa.

Moreover, interpersonal dynamics between participants and role-players predicted performance ratings in the high-fidelity simulations as well as instructor and peer ratings of interpersonal adaptability and task-performance in interpersonal settings. In particular, there

appeared a trend that complementarity at the momentary level was related to higher performance ratings in the high-fidelity simulations. However, for momentary complementarity in affiliation, this trend was only marginally significant. For momentary complementarity in dominance, this trend was only visible for performance ratings made by role-players and not for performance ratings made by independent assessors who did not interact with the participants themselves. Further, momentary complementarity in affiliation predicted interpersonal adaptability rated by instructors. Momentary complementarity also added incremental validity in the prediction of interpersonal adaptability beyond controls and other interpersonal predictors. Finally, momentary complementarity also predicted instructor and peer ratings of task performance. Momentary complementarity in dominance positively predicted instructor rated task performance and added incremental variance beyond control variables and other interpersonal predictors. Momentary complementarity in affiliation, however, was negatively related to task performance rated by peers.

Implications for Theory

The main results of the four chapters of this dissertation provide various theoretical contributions to the use of low-fidelity simulations and high-fidelity simulations, such as Multiple Speed Assessments, to assess performance in interpersonal situations and (interpersonal) adaptability. First, chapter 2 exemplified that the combined emic-etic approach to the development of SJTs can serve to obtain SJTs which scores can be used to compare participants across different geographical regions. These SJTs included SJTs that assessed the procedural knowledge about working well with others and adapting to change. Chapter 2 therefore extends similar results from studies on the cross-cultural transportability of measures of personality (Cheung, Cheung et al., 2008; Cheung, Fan et al., 2008; Cheung et al., 1996; Schmit et al., 2000) to SJTs. This is noteworthy because of the highly contextualized nature of SJTs (e.g., Lievens, 2006; Ployhart & Weekley, 2006) and the fact

that previous investigations of the cross-cultural transportability of SJTs that did not include cross-regional/cultural input across all steps of SJT development showed mixed results (Lievens, Corstjens, et al., 2015; Such & Schmidt, 2004).

Second, chapter 3 defined Multiple Speed Assessments as an umbrella term that encompasses different approaches that have emerged across various contexts (e.g., Brannick, Erol-Korkmaz, & Prewett, 2011; Knorr & Hissbach, 2014; Lievens, De Corte, & Westerveld, 2015). Thereby, different fields are connected with each other that ultimately serves to concentrate research efforts and advances knowledge about Multiple Speed Assessments.

Third, chapter 3 explicated the various theoretical fundaments of Multiple Speed Assessments. In particular, this chapter describes how Trait Activation Theory (Lievens et al., 2009; Tett & Burnett, 2003) explains how multiple short simulations serve to elicit a sufficient number of behaviors on behalf of participants, as well as how the principle of aggregation (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014) and research on the zero acquaintance/thin slices paradigm (e.g., Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992; Connelly & Ones, 2010) explain the reliability and ultimately validity of Multiple Speed Assessments that serve to assess overall behavior across situations or intraindividual variability across situations. Fourth, chapter 3 summarized previous research evidence on different variations of Multiple Speed Assessments and formulated a research agenda to further enhance our knowledge about Multiple Speed Assessments.

Fifth, chapter 4 illustrates the crucial role of the principle of aggregation in multiple speed assessments. That is, ratings from single assessors in single, short simulations show only low to moderate reliability. However, if ratings are aggregated across many different simulations or assessors, reliability increases because such an aggregation maximizes the amount of systematic variance in ratings that is shared across situations (Epstein, 1979; Kuncel & Sackett, 2014) and assessors (Eisenkraft, 2013). In this way, aggregating ratings

across simulations and assessors in multiple speed assessments generates a reliable and valid indicator of performance.

Sixth, chapter 4 exemplifies how and why Multiple Speed Assessments provide "good information" for assessors to evaluate performance in a predefined domain. That is, the multiple short simulations capture different situational demands that elicit behavior that is relevant for a vast set of qualitatively different aspects of the criterion domain. In this way, the various simulations do not serve as alternate measures of one another, which is also evident in the large contribution of the participant x simulation interaction effect to reliable variance, but optimize the behavioral sampling and point-to-point correspondence between predictor and criterion (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Seventh, chapter 5 showed that participants and role-players in high-fidelity simulations show substantial intraindividual variability in interpersonal variability at the continuous momentary level. Thereby, these results extend the evidence from different studies in social and clinical psychological (lab) settings (Markey, Lowmaster, & Eichler, 2010; Sadler, Ethier, Gunn, Duong, & Woody, 2009; Tracey, 2004) to the context of high-fidelity simulations. Further it extends the notion of intraindividual variability *across* high-fidelity simulations (e.g., Gibbons & Rupp, 2009; Jackson, Michaelides, Dewberry, & Kim, 2016; Lance, 2008; Lievens, 2009; Putka & Hoffman, 2013) to the perspective of intraindividual variability within high-fidelity simulations. The results further show that role-players' interpersonal behavior within high-fidelity simulations is not stationary although role-players often receive instructions to display specific interpersonal behavior and prompts (Lievens et al., 2015; Schollaert & Lievens, 2011, 2012).

Eighth, chapter 5 contributed to our knowledge about interpersonal dynamics between participants and role-players in high-fidelity simulations. The results showed that the principles of complementarity (i.e., correspondence in affiliation and reciprocity in

dominance) appear to be powerful heuristics to describe interpersonal dynamics between role-players and participants at the continuous momentary level although high-fidelity simulations that sample work-related situations are characterized by strong task and organizational situational demands that might interfere with participants' tendency to establish such dynamics (see, for example, Meyer, Dalal, & Hermida, 2010; Moskowitz, Ho, & Turcotte-Tremblay, 2007; Oliver, Hausdorf, Lievens, & Conlon, 2016).

Ninth, chapter 5 outlines that interpersonal dynamics in high-fidelity simulations are relevant because the degree to which participants establish interpersonal dynamics with role-players at the continuous momentary level that follow the principles of complementarity predict performance ratings in high-fidelity simulations as well as job-related performance outside of the context of high-fidelity simulations. Especially, momentary complementarity in high-fidelity simulations predicted ratings of interpersonal adaptability. The extent to that participants establish interpersonal dynamics with role-players across different high-fidelity simulations that follow the principles of complementarity at the continuous momentary level thus serves as a new angle to the assessment of interpersonal adaptability. Such a theoretically driven, contextualized approach to the assessment of interpersonal adaptability thus complements previous measures of interpersonal adaptability (Charbonnier-Voirin & Roussel, 2012; Griffin et al., 2007; Ployhart & Bliese, 2006; Pulakos et al., 2000) that usually rely on self- and other-report questionnaires that tap into a general tendency to adapt one's behavior to interaction partners without relating specific interpersonal demands (e.g., specific interpersonal behavior of interaction partners) to specific interpersonal behavior (see Oliver & Lievens, 2014 for similar arguments).

Tenth, chapter 5 stresses the relevance of continuous assessments (see Gabriel et al., 2017; Jebb & Tay, 2017). In particular, chapter 5 highlights the significant intraindividual variability of participants and role-players' interpersonal behavior in high-fidelity simulations

at the continuous level that did not appear to be random error but instead appeared to indicate at least in part how participants and role-players adapt their interpersonal behavior to each other. That is because the interpersonal behavior of participants and role-players was entrained at the continuous momentary level in line with the principles of complementarity. Further, only analyses at the momentary level revealed consistent patterns of interpersonal dynamics between participants and role-players that further predicted performance ratings in high-fidelity simulations as well as job-related performance ratings beyond self-ratings of personality and single-point estimates of interpersonal behavior.

In sum, this dissertation shows that low-fidelity simulations can be used to assess procedural knowledge about performance in interpersonal settings and adapting to change across geographical regions and cultures. Further, Multiple Speed Assessments as new approaches of high-fidelity simulations can serve to gain reliable and valid insights into participants' performance in interpersonal settings that include adapting (interpersonal) behavior to different interpersonal actors or demands. Especially, the interpersonal dynamics that participants establish with role-players at the continuous momentary level provide a new angle to the assessment of interpersonal adaptability.

Limitations

This dissertation is not without limitations. First, the SJTs that were investigated for their transportability across geographical regions did not assess performance in interpersonal settings or adapting to different (interpersonal) demands per se, but procedural knowledge about these performance domains. Future research should therefore investigate whether other assessment approaches that assess actual performance in these domains, such as Constructed Response Multimedia Tests (e.g., Cucina et al., 2015; Lievens et al., 2015; Oostrom, Born, Serlie, & van der Molen, 2010, 2011) can be compared across regions and cultures if they are developed according to a combined emic-etic approach.

Second, the available data only allowed investigations of the transportability of these SJTs across regions of Europe and Latin America. Clearly, it is of importance to extend this evidence to further geographical regions and cultures.

Third, this dissertation only provides empirical evidence about the reliability of Multiple Speed Assessments that consist of role-play simulations and sample the leadership domain. As summarized in chapter 3, Multiple Speed Assessments can also integrate other high-fidelity simulations, such as short fact-finding exercises, presentations, or Constructed Response Multimedia Tests (Cucina et al., 2015; Lievens et al., 2015; Oostrom et al., 2010, 2011). Further, Multiple Speed Assessments can sample other domains, such as domains of integrity or of performance in specific job fields such as business consultants or call center agents. Related to the validity of Multiple Speed Assessments to predict performance in these domains, we would expect positive validity evidence as long as the different (types of) multiple short simulations cover different aspects of the targeted domain and ultimately ensure a high point-to-point correspondence between predictor and criterion domain (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Fourth, the mechanisms that explain the relations between complementarity in high-fidelity simulations on the one hand and role-player or assessor ratings of performance in these simulations as well as ratings of interpersonal adaptability and task performance made by instructors and peers on the other hand could not be fully investigated in this dissertation. That is because several potential mediators have not been assessed. Future research might therefore shed a light on these mechanisms. For example, future studies might implement assessments of potential mediators such as satisfaction with interactions or liking of participants on behalf of role-players and assessors in high-fidelity simulations, or on behalf of instructors and peers who provide criterion ratings. This might provide answers to the question whether complementarity might tap into true performance in high-fidelity

simulations or introduce a bias on behalf of role-players. Peers who collaborate with participants in job-related situations and provide ratings of job-related performance, such as task performance, might also indicate the clarity of role-assignments in job-related situations. This might provide further clarity whether reciprocity in dominance in high-fidelity simulations predicts task performance in interpersonal settings because reciprocity in dominance implies agreeing upon each other's status and assigned role in a given interaction (Bendersky & Hays, 2012; Carson, 1969; Kiesler, 1983; Locke & Sadler, 2007).

Implications for Practice

Although we stress the necessity of further replications, this dissertation provides several implications for practice in personnel selection and development.

First, practitioners in personnel selection and development who want to apply SJTs across geographical regions and cultures might best follow a combined emic-etic approach to the development of SJTs. Although this approach to the development of SJTs demands high efforts because it requires cross-cultural input and feedback loops across all stages of test development, this dissertation exemplifies that it pays off in terms of SJT scores that show metric measurement invariance across geographical regions and cultures. Thus, applying a combined emic-etic approach to the development of SJTs contributes to the phenomenon that participants from different geographical regions and cultures interpret SJT scenarios and response options in the same way and attribute equal meaning to them. This is an important precondition to limit the confounding of measurement effects and true mean differences in the procedural knowledge that is to be assessed (see, for example, Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Therefore, the interpretation of possible mean differences across (regional/cultural) groups is facilitated.

Second, this dissertation suggests that organizations might adopt Multiple Speed Assessments for personnel selection and development purposes to gain insights about

participants' performance related to (interpersonal) performance and adapting to different (interpersonal) demands. However, this dissertation also showcases that reliable and valid performance ratings in Multiple Speed Assessments strongly require aggregations across multiple simulations and independent assessors. Practitioners should therefore not use single short simulations in which a single assessor provides ratings of participants' performance. Instead, the results showcase that ratings from at least fourteen simulations with two independent assessors each or from at least nine simulations with at least three independent assessors each need to be aggregated to gain an overall Multiple Speed Assessment score with a reliability of $G \geq .70$.

Third, this dissertation suggests that practitioners in personnel selection and development should not only focus on performance ratings in high-fidelity simulations, but might also consider to assess the interpersonal dynamics that participants establish with role-players. That is because this dissertation showed that interpersonal dynamics between participants and role-players at the continuous momentary level predict job-related performance. Especially, organizations might consider the assessment of the degree of complementarity that participants establish with role-players on the continuous momentary level as a new angle to the assessment of interpersonal adaptability that could complement or replace traditional self- or other reports of interpersonal adaptability (Charbonnier-Voirin & Roussel, 2012; Griffin et al., 2007; Ployhart & Bliese, 2006; Pulakos et al., 2000). An assessment of interpersonal adaptability via the degree of following the principles of complementarity at the continuous momentary level implies at least two advantages compared to decontextualized self- or other reports of interpersonal adaptability. First, it resembles a theoretically driven approach to the assessment of interpersonal adaptability (see also Oliver & Lievens, 2014; Pincus et al., 2014; Sadler, Ethier, & Woody, 2011). Second, the clear prescription of how specific interpersonal demands (e.g., different degrees of affiliation and

dominance expressed by a role-player) might best be mastered by showing specific interpersonal behavior (e.g., different degrees of affiliation and dominance expressed by the participant) via the principles of complementarity provides a more contextualized assessment approach that better captures the nature of the construct of interpersonal adaptability (Oliver & Lievens, 2014). However, practitioners might best investigate and aggregate the degree of complementarity that participants show across different simulations and role-players because complementarity is a dyadic phenomenon that is not only influenced by the participant, but also by the role-player. Multiple Speed Assessments in which participants interact with many different role-players in multiple short simulations might thus be an efficient way to assess interpersonal dynamics and the degree of complementarity that participants establish with different role-players.

Fourth, this dissertation suggests that practitioners might implement continuous assessments of behavior in high-fidelity simulations (Gabriel, Diefendorff, Bennett, & Sloan, 2017; Jebb & Tay, 2017). As an example, this dissertation showed that interpersonal behavior significantly varies at the continuous momentary level. Further, interpersonal dynamics at the continuous momentary level added incremental validity to the prediction of performance in high-fidelity simulations and job-related performance beyond traditional assessments of interpersonal behavior that involved self-ratings of personality and interpersonal dynamics assessed at the overall level that do not account for continuous variability at the momentary level. However, we acknowledge that assessments of behavior at the continuous momentary level consume high efforts. For example, this dissertation exemplified that continuous assessments need to be made and aggregated across multiple independent coders to gain reliable indicators (see also Sadler, Ethier, Gunn, Duong, & Woody, 2009). Future technological developments might provide opportunities for practitioners to facilitate

continuous assessment of behavior in ways that complement or replace human assessors (e.g., Schmid Mast, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015).

Directions for Future Research

This dissertation suggests various directions for future research. First, it encourages investigations of intraindividual variability in behavior and performance across simulations in Multiple Speed Assessments. Various theoretical frameworks assume that intraindividual variability across situations does not only indicate random measurement error, but does also reveal substantive, meaningful variance. For example, the Cognitive-Affective Personality System Theory (Mischel & Shoda, 1995) explains intraindividual variability across situations as differences in how people construe varying situations in different ways, which ultimately causes individuals to respond differently to these varying situations. Building upon this notion, Whole Trait Theory (Fleeson & Jayawickreme, 2015) posits that personality should not only be conceived in terms of a general cross-situational and cross-temporal consistent tendency to show specific behaviors, which might be represented by a mean score across situations and time, but also in terms of individuals variability across different situations and time. This variability might be represented by a standard deviation or a density distribution across different situations and time, or more sophisticated modeling approaches, such as IRT Tree Models (Lang, Lievens, De Fruyt, Zettler, & Tackett, 2019; Lievens et al., 2018).

In line with these theoretical assumptions, several studies already provided evidence for substantial intraindividual variability in behavior and performance across situations (Dalal, Bhave, & Fiset, 2014; Fleeson, 2001; Fournier, Moskowitz, & Zuroff, 2008; Gibbons & Rupp, 2009; Judge, Simon, Hurst, & Kelley, 2014; Minbashian & Luppino, 2014; Moskowitz & Zuroff, 2004; Smith, Shoda, Cumming, & Smoll, 2009). Results from chapter 4 showcase that substantial variability across situations can also be found in Multiple Speed Assessments. That is, a participant x simulation interaction effect explained the largest part of reliable

variance in assessor ratings. This indicates that participants show substantial variability in performance across different simulations. Future research might apply different strategies to validate various forms of intraindividual variability that might be captured in Multiple Speed Assessments by relating variability indicators derived from Multiple Speed Assessments to (a) validated measures of variability, (b) variability across situations on and off the job, or (c) relevant outcomes, such as job-performance or training performance (Lievens, 2017). Only recently, Lievens et al. (2018) assessed individuals' intraindividual variability in terms of variability of responses across different SJT items. Individuals variability across different SJT items (a) related to self-reports of functional flexibility, (b) predicted intraindividual variability in an experience sampling study across ten days, and (c) predicted performance ratings beyond mean scores across SJT items. This study provides encouraging support for the notion that intraindividual variability across different (low- or high-fidelity) simulations indicates variability that is not only random error but reveals meaningful variance that relates to important outcomes.

As one possible perspective on intraindividual variability across simulations, Multiple Speed Assessments might shed a light on participants' adaptability (Baard, Rench, & Kozlowski, 2014; Jundt, Shoss, & Huang, 2015; Ployhart & Bliese, 2006; Pulakos et al., 2000). To this end, one might develop Multiple Speed Assessments that build upon more general situational taxonomies (Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014) or situational taxonomies for a specific domain, such as the leadership domain (Yukl, 1989, 2010). Then, one might observe whether participants follow different behavioral approaches to solve the various problems represented in these different situations. Researchers might then explore any systematic patterns of specific situation-behavior linkages (i.e., behavioral signatures; Mischel & Shoda, 1995; Shoda, Mischel, & Wright, 1994; Smith et al., 2009) and how they relate to performance within and across the different simulations as well as self- and

other ratings of adaptability (Lievens et al., 2018). However, such research endeavors that intent to investigate or approve successful patterns how behavior is adapted to different situational demands call for further developments of situational taxonomies and might benefit from theoretically prescriptions how these different situations might best be handled to ensure high performance. In fact, one might even view Multiple Speed Assessments as a viable tool to further explore and test hypotheses about specific situation-behavior linkages. In this way, Multiple Speed Assessments could serve to facilitate further theory development.

As another perspective on intraindividual variability in Multiple Speed Assessments, one might generate various indicators that capture different forms of variability in performance across time and/or different situations. An investigation of variability in task performance across 36 experience sampling studies showed that the majority of variance appears to be attributable to within-person variance (Dalal et al., 2014). Further, there is evidence that mean performance and different indicators of performance variability need to be distinguished from each other. For example, typical performance (i.e., mean performance) and maximum performance appear to be only moderately correlated with each other (Beus & Whitman, 2012). On top of that, individual differences and situational variables appear to predict different indicators of performance, such as typical and maximum performance to various extents (Marcus, Goffin, Johnston, & Rothstein, 2007; Witt & Spitzmüller, 2007). Moreover, different indicators of performance, such as typical performance, maximum performance, and performance variability have been found to relate differently to meaningful outcomes, such as compensation (Barnes & Morgeson, 2007).

In a recent review of the literature on job performance variability, Dalal et al. (2014) concluded that we still lack a sound understanding of performance variability and the within-person structure of performance. Given that Multiple Speed Assessments allow to observe participants' performance across multiple simulations that can capture different situations,

they might provide a viable tool to advance our knowledge about job-related performance variability. For example, one might derive indicators of participants mean level of performance, maximum performance, minimum performance, and standard deviation or range in performance across the different simulations. One might then examine relations to individual differences or situational predictors that might explain individual differences in intraindividual variability in performance or relate these different performance indicators to relevant outcomes (Dalal et al., 2014). Further, one might investigate whether indicators of intraindividual variability derived from Multiple Speed Assessments predict corresponding indicators of typical performance, maximum performance, minimum performance and standard deviation or range in performance on the job. Therefore, examining different indices of performance and performance variability in Multiple Speed Assessments might provide further insights into the nature of variability in job-related performance.[21]

Another perspective on intraindividual variability in Multiple Speed Assessments relates to different performance trajectories across the course of multiple simulations. For example, some participants might need to involve in many short simulations in which they show moderate levels of performance until they figure out how problems might best be solved in such a short amount of time and ultimately show higher levels of performance. Other participants, however, might more quickly develop appropriate approaches to solve problems in the short simulations and might show higher levels of performance much quicker (i.e., after less simulations). One might then investigate whether different performance trajectories relate to self- and other reports of learning agility (DeRue, Ashford, & Myers, 2012) that describes

[21] Some might argue that selection situations create strong situational demands on behalf of participants because they are aware of being evaluated and might thus be highly motivated to "put their best foot forward" in order to receive favorable performance ratings and increase chances for employment (Sackett, Zedeck, & Fogli, 1988; Smith-Jentsch, 2007). In line with this argument, one might expect rather low variability in performance across different stimulations in Multiple Speed Assessments. However, motivation alone might not be the only variable that causes variability in performance across different situations. Instead, variability in performance might also depend on how participants construe a given situation (Mischel & Shoda, 1995), and adapt their behavior to this situation, which also relates to differences in skill repertoires of participants (Gibbons & Rupp, 2009).

how well and quickly participants learn from prior experience. In a similar manner, one might investigate whether some participants might show a slowly or suddenly occurring drop in performance after several role-plays, whereas other participants might be able to show consistent levels of performance across the multiple simulations across time. Such different performance trajectories might also provide insights into participants' stress/psychological resilience (Fletcher & Sarkar, 2013) because Multiple Speed Assessments that assess performance across multiple simulations can be expected to create stressful situations. One might therefore investigate how different performance trajectories across time in Multiple Speed Assessments relate to self- and other reports or even physiological indicators of stress resilience.

In general, Multiple Speed Assessments provide interesting opportunities to the investigation of intraindividual variability in behavior and performance because Multiple Speed Assessments present the same situations (i.e. simulations) to all participants. In this way, the situation is kept constant across individuals. It can thus be excluded that individual differences in intraindividual variability occur solely because of different varieties of situations that different individuals face or select (Lievens, 2017). This provides a key advantage of Multiple Speed Assessments compared to experience sampling methods that are often used to study intraindividual variability (Wrzus & Mehl, 2015).

Second, future research might investigate the applicability of Multiple Speed Assessments for development purposes. Test developers might build upon situational taxonomies to capture various parts of the domain to be sampled with several simulations that target similar parts of the domain or similar situational demands. For example, if Multiple Speed Assessments should sample participants' behavior in the leadership domain, one might draw upon the Multiple Linkage Theory (Yukl, 2010) that proposes six prominent leadership challenges (i.e., role ambiguity, immediate crisis, inadequate skills, inadequate cooperation,

scarce resources, weak task commitment). A Multiple Speed Assessment could then be designed to sample behavior in each of these six prominent leadership situations with three simulations each. Then, participants might involve in the first six simulations that would confront them with each of these six different situational demands. A coach might observe participants during the simulations and provide feedback afterwards how the problems might be solved more effectively. Then, participants could try to improve their behavior in the next set of six simulations. This procedure could repeat several times to observe participants' behavior and performance, provide them with feedback and to investigate whether they might learn from feedback and prior experience. In this way, Multiple Speed Assessments might provide the opportunity to emerge as an efficient development intervention tool.

In a similar manner, future research might investigate whether Multiple Speed Assessments could be used as development interventions to improve individuals' tendency to establish complementarity at the continuous momentary level. Given that this dissertation showed that establishing complementarity at the continuous momentary level relates to important outcomes, such as performance ratings in high-fidelity simulations and ratings of interpersonal adaptability or task performance in interpersonal settings, improving the tendency to establish complementarity across interactions with different interaction partners appears crucial in today's world of work that confronts individuals with the challenge to perform well in interpersonal situations and to adapt to different (interpersonal) demands because the frequency and complexity of interpersonal interactions on the job has increased (Griffin et al., 2007; Pulakos et al., 2000). Further, the focus on complementarity at the momentary behavioral level might be a promising opportunity for developmental interventions because interventions focusing on actual behavior produced favorable training effects (Burke & Day, 1986; Taylor, Russ-Eft, & Chan, 2005).

Conclusion

This dissertation addressed several objectives to contribute to our knowledge how personnel selection and development procedures might help organizations to assess individuals' performance in interpersonal settings and performance related to adapting to different (interpersonal) demands. Results showed that low-fidelity simulations such as SJTs that follow the combined emic-etic approach to the development of SJTs that includes cross-regional/cultural input across all stages of test development can contribute to SJT scores that can be compared across geographical regions and cultures, including SJTs that assess procedural knowledge about working well with others and adapting to change. Next, this dissertation introduced the umbrella term of Multiple Speed Assessments that encompass varies approaches that build upon multiple, short, and often integrated simulations to get insights into the behavioral repertoire of participants in situations sampled from a given domain. Further results showed that assessor ratings in a face-to-face version of Multiple Speed Assessment were reliable and valid indicators of performance, but only if ratings were aggregated across many situations and independent assessors. Finally, this dissertation showed that interpersonal behavior of participants and role-players in high-fidelity simulations varies at the continuous moment-to-moment level and that acknowledging this intraindividual variability within high-fidelity simulations is crucial because interpersonal dynamics at the momentary level predict performance ratings in high-fidelity simulations as well as ratings of interpersonal adaptability and task performance in interpersonal settings. In conclusion, this dissertation shows that low-fidelity simulations such as SJTs and high-fidelity simulations such as Multiple Speed Assessments provide organizations with various opportunities to assess individuals' performance in interpersonal settings and performance related to adapting to different (interpersonal) demands to assess, select or develop a workforce that is capable of mastering interpersonal challenges of today's world of work.

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, UK: Cambridge University Press.

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance, 20,* 259–274. https://doi.org/10.1080/08959280701333289

Bendersky, C., & Hays, N. A. (2012). Status conflict in groups. *Organization Science, 23,* 323–340. https://doi.org/10.1287/orsc.1110.0734

Beus, J. M., & Whitman, D. S. (2012). The relationship between typical and maximum performance: A meta-analytic examination. *Human Performance, 25,* 355–376. https://doi.org/10.1080/08959285.2012.721831

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores: Reliability of objective

structured clinical examination scores. *Medical Education*, *45*, 1181–1189. https://doi.org/10.1111/j.1365-2923.2011.04075.x

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology*, *71*, 232–245. http://dx.doi.org/10.1037/0021-9010.71.2.232

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 287–321. https://doi.org/10.1207/s15328007sem1302_7

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107–132. https://doi.org/10.1080/15305051003637306

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

Charbonnier-Voirin, A., & Roussel, P. (2012). Adaptive performance: A new scale to measure individual performance in organizations. *Canadian Journal of Administrative Science*, *29*, 280–293. https://doi.org/10.1002/cjas.232

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

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

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

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

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

Fletcher, D., & Sarkar, M. (2013). Psychological resilience: A review and critique of definitions, concepts, and theory. *European Psychologist*, *18*, 12–23. https://doi.org/10.1027/1016-9040/a000124

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

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

Gabriel, A. S., Diefendorff, J. M., Bennett, A. A., & Sloan, M. D. (2017). It's about time: The promise of continuous rating assessments for the organizational sciences. *Organizational Research Methods*, *20*, 32–60. https://doi.org/10.1177/1094428116673721

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

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal*, *50*, 327–347. https://doi.org/10.5465/amj.2007.24634438

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

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*, 61–94. https://doi.org/10.1177/1094428116668035

Judge, T. A., Simon, L. S., Hurst, C., & Kelley, K. (2014). What I experienced yesterday is who I am today: Relationship of work motivations and behaviors to within-individual variation in the five-factor model of personality. *Journal of Applied Psychology*, *99*, 199–221. https://doi.org/10.1037/a0034485

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

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., Lievens, F., De Fruyt, F., Zettler, I., & Tackett, J. L. (2019). Assessing meaningful within-person variability in likert-scale rated personality descriptions: An IRT tree approach. *Psychological Assessment, 31,* 474–487. http://dx.doi.org/10.1037/pas0000600

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 279–300). Mahwah, NJ: Erlbaum.

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18,* 102–121. https://doi.org/10.1080/13594320802058997

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

Locke, K. D., & Sadler, P. (2007). Self-efficacy, values, and complementarity in dyadic interactions: Integrating interpersonal and social-cognitive theory. *Personality and Social Psychology Bulletin, 33*, 94–109. https://doi.org/10.1177/0146167206293375

Marcus, B., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. *Human Performance, 20*, 275–285. https://doi.org/10.1080/08959280701333362

Markey, P., Lowmaster, S., & Eichler, W. (2010). A real-time assessment of interpersonal complementarity. *Personal Relationships, 17*, 13–25. https://doi.org/10.1111/j.1475-6811.2010.01249.x

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

Minbashian, A., & Luppino, D. (2014). Short-term and long-term within-person variability in performance: An integrative model. *Journal of Applied Psychology, 99,* 898-914. http://dx.doi.org/10.1037/a0037402

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

Moskowitz, D. S., Ho, M. -h. R., & Turcotte-Tremblay, A.-M. (2007). Contextual influences on interpersonal complementarity. *Personality and Social Psychology Bulletin, 33,* 1051–1063. https://doi.org/10.1177/0146167207303024

Moskowitz, D. S., & Zuroff, D. C. (2004). Flux, pulse, and spin: Dynamic additions to the personality lexicon. *Journal of Personality and Social Psychology, 86,* 880–893. https://doi.org/10.1037/0022-3514.86.6.880

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. https://doi.org/10.1177/0149206314525207

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. 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

Pincus, A. L., Sadler, P., Woody, E., Roche, M. J., Thomas, K. M., & Wright, A. G. C. (2014). Multimethod assessment of interpersonal dynamics. In C. J. Hopwood & R. F. Bornstein (Eds.), *Multimethod clinical assessment* (pp. 51–91). New York: Guilford.

Ployhart, R. E., & Bliese, P. D. (2006). Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In S. Burke, L. Pierce, & Salas (Eds.), *Understanding adaptability: A prerequisite for effective performance within complex environments* (pp. 3–39). Bingley: Emerald Publishing Group.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 345-350). Mahwah, NJ: Erlbaum.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, *85*, 612–624. https://doi.org/10.1037//0021-9010.85.4.612

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

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482–486. http://dx.doi.org/10.1037/0021-9010.73.3.482

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology, 97,* 1005–1020. https://doi.org/10.1037/a0016232

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.

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science, 24,* 154–160. https://doi.org/10.1177/0963721414560811

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153–193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x

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

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

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

Such, M. J., & Schmidt, D. B. (2004). *Examining the effectiveness of empirical keying: A cross-cultural perspective.* Presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, IL.

Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology*, *90*, 692–709. http://dx.doi.org/10.1037/0021-9010.90.4.692

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

Tracey, T. J. G. (2004). Levels of interpersonal complementarity: A simplex representation. *Personality and Social Psychology Bulletin*, *30*, 1211–1225. https://doi.org/10.1177/0146167204264075

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://doi.org/10.1177/109442810031002

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52,* 372-376. http://dx.doi.org/10.1037/h0026244

Witt, L. A., & Spitzmüller, C. (2007). Person-situation predictors of maximum and typical performance. *Human Performance*, *20*, 305–315. https://doi.org/10.1080/08959280701333529

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? Measuring personality processes and their social consequences: Lab and/or field? *European Journal of Personality*, *29*, 250–271. https://doi.org/10.1002/per.1986

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

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

ENGLISH SUMMARY

Prevailing trends in our world of work, such as the ever increasing globalization that facilitates collaborations between organizations from different cultures (Cascio, 2003; Javidan, Dorfman, de Luque, & House, 2006), shifts to service-oriented business (Zeithaml & Bitner, 1996), or project based work involving formations of new teams (Hesketh & Neal, 1999; Kozlowski, Gully, Salas, & Cannon-Bowers, 1996) have stressed the importance of performing well in interpersonal situations and to adapt to varying (interpersonal) demands (Griffin, Neal, & Parker, 2007; Pulakos, Arad, Donovan, & Plamondon, 2000). As one possible way to master this challenge, organizations might apply personnel selection and development procedures that assess individuals' performance in these situations. One intriguing approach might be to apply simulation-based procedures that confront participants with various interpersonal situations. For example, low-fidelity simulations such as Situational Judgment Tests (SJTs) might be used that traditionally confront participants with written situation descriptions and various response options that need to be rated, rank-ordered or chosen (Motowidlo, Dunnette, & Carter, 1990). As another simulation-based procedure that confronts participants with various interpersonal situations, Multiple Speed Assessments have been developed across different fields (Brannick, Erol-Korkmaz, & Prewett, 2011; Knorr & Hissbach, 2014; Lievens, De Corte, & Westerveld, 2015) and were recently added to the portfolio of selection practitioners (Byham, 2016). Multiple Speed Assessments sample participants' actual behavior in a predefined domain across multiple, short interpersonal simulations.

To add knowledge how personnel selection and development procedures such as SJTs and Multiple Speed Assessments might serve to assess performance in interpersonal settings and adapting to different (interpersonal) demands, this dissertation addressed four objectives. A first objective was to investigate whether a combined emic-etic approach to the

development of SJTs that incorporates cross-regional and cross-cultural input across all stages of test development serves to develop SJTs that can be compared across regions and cultures. A second objective was to provide a conceptual overview of Multiple Speed Assessments, including their common characteristics, theoretical fundaments, design variations, application areas and an agenda for future research. A third objective was to provide knowledge about the reliability and validity of a face-to-face variant of Multiple Speed Assessments. A fourth objective was to provide knowledge about the interpersonal behavior of participants and the interpersonal dynamics they establish with role-players in high-fidelity simulations at the continuous momentary level as well as about relations between these interpersonal dynamics on the one hand and ratings of performance in high-fidelity simulations, interpersonal adaptability and task performance in interpersonal settings on the other hand. Each of these objectives is addressed in one separate chapter of this dissertation.

Chapter 2 addressed Objective 1. Five SJTs were developed to assess procedural knowledge about five different competencies that are crucial to flourish in entry-level jobs across regions, industries, or professions. These SJTs included assessments of procedural knowledge about “adapting to change” and “working well with others”. In line with the combined emic-etic approach (Cheung, Fan, Cheung, & Leung, 2008; Schmit, Kihm, & Robie, 2000), cross-regional/cultural input was incorporated at all stages of test development. Chapter 2 then empirically demonstrated configural and metric measurement invariance for these five SJTs across participants from Europe and Latin America. Hence, the same factor structure explained SJT scenario scores for participants from Europe and Latin America and the latent factor(s) were equally assessed across both regional groups (Byrne & Stewart, 2006; Byrne & van de Vijver, 2010). In other words, participants from Latin America and Europe interpreted the SJT scenarios and response options in the same way and attributed the same meaning to them (see also Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Therefore, this chapter showcases the value of the combined emic-etic approach to the development of SJTs to obtain SJT scores that can be used to compare participants across regions and cultures.

Chapter 3 addressed Objective 2. It defined Multiple Speed Assessments as an umbrella term for various assessment approaches that build upon multiple, short interpersonal simulations to gain insights into the behavioral repertoire of participants in a predefined domain. It further summarized the common characteristics of Multiple Speed Assessments as (a) applying multiple simulations (e.g., 20), (b) applying short simulations (often less than 5 minutes), (c) applying structured simulations, (d) streamlining evaluations of performance, and (e) integrating the simulations into an overarching background. As the theoretical fundaments of Multiple Speed Assessments, this chapter identified the zero acquaintance/thin slices paradigm (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992; Back & Nestler, 2016; Connelly & Ones, 2010), trait-activation theory (Lievens, Tett, & Schleicher, 2009; Tett & Burnett, 2003) as well as the principle of aggregation (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014). Across various application areas and different design variations, Multiple Speed Assessments aim to assess overall behavior across situations or participants' intraindividual variability across situations. By providing this overview of Multiple Speed Assessments, chapter 3 connects different fields with each other. To further promote research on and advance knowledge about Multiple Speed Assessments, chapter 3 further proposed an agenda for future research.

Chapter 4 addressed Objective 3. A face-to-face format of Multiple Speed Assessments that samples the leadership domain was developed and investigated in terms of its reliability and validity. The different simulations requested to perform well in interpersonal situations and adapt to different (interpersonal) demands because participants faced different role-players across the different simulations. For performance ratings made by single

assessors (role-players) in single short simulations, results showed low to moderate interrater reliabilities. However, when performance ratings were aggregated across three to four assessors per simulation, interrater reliabilities were moderate to high. Further, the decomposition of different sources of variance in performance ratings showed that the majority of reliable variance in Multiple Speed Assessments reflects a participant x simulation interaction effect which indicates that differences in participants' performance vary across different simulations. Regarding validity, a similar pattern emerged. That is, only when ratings were aggregated across all simulations, relations to cognitive ability and personality as well as to performance ratings from instructors and peers followed theoretically derived hypotheses and were moderate to high. Such an overall Multiple Speed Assessment score further added incremental validity in predicting instructor and peer rated performance beyond traditional predictors that tap into a similar domain. This chapter therefore exemplifies the key role of the principle of aggregation in Multiple Speed Assessments (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014). Further, it illustrates that Multiple Speed Assessments provide good information for assessors to provide performance ratings in a given domain because multiple speed assessments capture different situational demands via multiple simulations that tap into qualitatively different aspects of the criterion domain. Thereby, Multiple Speed Assessments optimize the point-to-point correspondence between predictor and criterion via multiple short simulations which mirrors an efficient form of the behavioral sampling approach (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Finally, chapter 5 addressed Objective 4. It investigated participants' and role-players' intraindividual variability in interpersonal behavior as well as interpersonal dynamics in four high-fidelity simulations. Participants' and role-players' interpersonal behavior significantly varied at the momentary level in all of the high-fidelity simulations. Further, this intraindividual variability did not appear to mirror random error only, because momentary

interpersonal behavior of participants and role-players were entrained in line with the principles of complementarity. That is, the interpersonal dynamics between participants and role-players consistently followed the principles of correspondence in affiliation and reciprocity in dominance at the continuous momentary level in all four high-fidelity simulations. Finally, the degree to which participants followed the principles of complementarity across the four high-fidelity simulations at the continuous momentary level predicted performance ratings in the high-fidelity simulations as well as ratings of interpersonal adaptability and task performance in interpersonal settings provided by instructors and peers. This chapter therefore expands the notion of intraindividual variability *across* high-fidelity simulations (Gibbons & Rupp, 2009; Jackson, Michaelides, Dewberry, & Kim, 2016; Lance, 2008; Lievens, 2009; Putka & Hoffman, 2013) to the level of intraindividual variability *within* high-fidelity simulations. It further contributes to our knowledge about the nature and relevance of continuous, momentary interpersonal dynamics in high-fidelity simulations. Especially, momentary interpersonal dynamics in high-fidelity simulations that follow the principles of complementarity might provide a new angle to assess interpersonal adaptability (Oliver & Lievens, 2014). In all these different ways, chapter 5 highlights the importance of assessing (interpersonal) behavior at the continuous momentary level (Gabriel, Diefendorff, Bennett, & Sloan, 2017; Jebb & Tay, 2017).

In sum, this dissertation contributes to our knowledge how individuals' performance in interpersonal settings and adapting to different (interpersonal) demands can be assessed via low-fidelity simulations such as SJTs and high-fidelity simulations such as Multiple Speed Assessments. In this way, this dissertation equips organizations with valuable knowledge to assess, select or develop individuals to more successfully master the interpersonal challenges of today's world of work.

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

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.

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

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

Byrne, B. M., & Stewart, S. M. (2006). TEACHER'S CORNER: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 287–321. https://doi.org/10.1207/s15328007sem1302_7

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107–132. https://doi.org/10.1080/15305051003637306

Cascio, W. F. (2003). Changes in workers, work, and organizations. In R. J. Klimoski, W. C. Borman, & D. R. Ilgen (Eds.), *Handbook of Psychology* (Vol. 12, pp. 401–422). Hoboken, NJ: Wiley & Sons.

Cheung, F. M., Fan, W., Cheung, S. F., & Leung, K. (2008). Standardization of the cross-cultural Chinese Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment. *Acta Psychologica Sinica, 40,* 839–852. http://dx.doi.org/10.3724/SP.J.1041.2008.01639

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

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

Gabriel, A. S., Diefendorff, J. M., Bennett, A. A., & Sloan, M. D. (2017). It's about time: The promise of continuous rating assessments for the organizational sciences. *Organizational Research Methods*, *20,* 32–60. https://doi.org/10.1177/1094428116673721

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

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal*, *50*, 327–347. https://doi.org/10.5465/amj.2007.24634438

Hesketh, B., & Neal, A. (1999). Technology and performance. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 21–55). San Francisco, CA: Jossey-Bass.

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

Javidan, M., Dorfman, P. W., de Luque, M. S., & House, R. J. (2006). In the eye of the beholder: *Academy of Management Perspectives*, *20*, 67–90. https://doi.org/10.5465/AMP.2006.19873410

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*, 61–94. https://doi.org/10.1177/1094428116668035

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

Kozlowski, S. W. J., Gully, S. M., Salas, E., & Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In M. Beyerlein, S. Beyerlein, & D. Johnson (Eds.), *Advances in interdisciplinary studies of work teams: Team leadership* (Vol. 3, pp. 251–289). Greenwich, CT: JAI Press.

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

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

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*, 102–121. https://doi.org/10.1080/13594320802058997

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

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

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.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, *85*, 612–624. https://doi.org/10.1037//0021-9010.85.4.612

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

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153–193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x

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

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://doi.org/10.1177/109442810031002

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372-376. http://dx.doi.org/10.1037/h0026244

Zeithaml, V. A., & Bitner, M. J. (1996). *Services marketing*. New York: McGraw-Hill.

NEDERLANDSTALIGE SAMENVATTING

Multiple Speed Assessment:

Een nieuwe benadering om interpersoonlijk gedrag en aanpassingsvermogen te meten?

Er zijn enkele trends in onze werkomgeving, die het belang hebben benadrukt van goed presteren in interpersoonlijke situaties en van zich aanpassen aan variërende (interpersoonlijke) eisen (Griffin, Neal, & Parker, 2007; Pulakos, Arad, Donovan, & Plamondon, 2000). Voorbeelden voor deze trends zijn de steeds toenemende globalisering die samenwerking tussen organisaties uit verschillende culturen mogelijk maakt (Cascio, 2003; Javidan, Dorfman, de Luque, & House, 2006), de verschuiving naar servicegericht werk (Zeithaml & Bitner, 1996) en het projectmatig werken met formaties van nieuwe teams (Hesketh & Neal, 1999; Kozlowski, Gully, Salas, & Cannon-Bowers, 1996). Als een mogelijke manier om deze uitdaging aan te gaan, kunnen organisaties personeelsselectie- en ontwikkelingsprocedures toepassen die de prestaties van individuen in deze situaties beoordelen. Een intrigerende benadering zou kunnen zijn om op simulatie gebaseerde procedures toe te passen die deelnemers confronteren met verschillende interpersoonlijke situaties. Er kunnen bijvoorbeeld “low-fidelity” simulaties zoals Situational Judgement Tests worden gebruikt die deelnemers traditioneel confronteren met schriftelijke situatiebeschrijvingen en verschillende reactie-opties die moeten worden beoordeeld, gerangschikt of gekozen (Motowidlo, Dunnette, & Carter, 1990). Een andere op simulatie gebaseerde procedure die deelnemers confronteert met verschillende interpersoonlijke situaties, zijn Multiple Speed Assessments die ontwikkeld zijn in verschillende gebieden (Brannick, Erol-Korkmaz, & Prewett, 2011; Knorr & Hissbach, 2014; Lievens, De Corte, & Westerveld, 2015) en die onlangs werden toegevoegd aan de portefeuille van selectie organisaties (Byham, 2016). Multiple Speed Assessments geven een voorbeeld van het

werkelijke gedrag van deelnemers in een vooraf gedefinieerd domein in meerdere, korte interpersoonlijke simulaties.

Om kennis toe te voegen aan hoe personeelsselectie- en ontwikkelingsprocedures zoals SJT's en Multiple Speed Assessments kunnen dienen om de prestaties in interpersoonlijke contexten te beoordelen en aan te passen aan verschillende (interpersoonlijke) eisen, ging dit proefschrift in op vier doelstellingen. Een eerste doelstelling was om te onderzoeken of een gecombineerde “emic-etic” benadering van de ontwikkeling van SJT's die interregionale en interculturele input in alle fasen van testontwikkeling omvat, dient om SJT's te ontwikkelen die kunnen worden vergeleken tussen regio's en culturen. Een tweede doelstelling was om een conceptueel overzicht te geven van Multiple Speed Assessments, inclusief hun gemeenschappelijke kenmerken, theoretische fundamenten, ontwerpvariaties, toepassingsgebieden en een agenda voor toekomstig onderzoek. Een derde doelstelling was om kennis te verschaffen over de betrouwbaarheid en validiteit van een face-to-face variant van Multiple Speed Assessments. Een vierde doelstelling was om kennis te verschaffen over het interpersoonlijke gedrag van deelnemers en de interpersoonlijke dynamiek die zij met rollenspelers in “high-fidelity” simulaties op het momentane niveau opbouwen, evenals over relaties tussen deze interpersoonlijke dynamiek enerzijds en beoordelingen van prestaties in high-fidelity simulaties, interpersoonlijke aanpasbaarheid en taakuitvoering in interpersoonlijke contexten anderzijds. Elk van deze doelstellingen wordt behandeld in een afzonderlijk hoofdstuk van dit proefschrift.

Hoofdstuk 2 ging over doelstelling 1. Er werden vijf SJT's ontwikkeld om procedurele kennis te beoordelen over vijf verschillende competenties die cruciaal zijn om te bloeien in jobs op instapniveau in verschillende regio's, bedrijfstakken of beroepen. Deze SJT's omvatten beoordelingen van procedurele kennis over "aanpassing aan verandering" en "goed samenwerken met anderen". In overeenstemming met de gecombineerde emic-etic benadering

(Cheung, Fan, Cheung, & Leung, 2008; Schmit, Kihm, & Robie, 2000), werd interregionale/interculturele input opgenomen in alle fasen van testontwikkeling. Hoofdstuk 2 demonstreerde vervolgens empirische configuratie- en metrische meetinvariantie voor deze vijf SJT's bij deelnemers uit Europa en Latijns-Amerika. Daarom verklaarde dezelfde factorstructuur SJT-scenario scores voor deelnemers uit Europa en Latijns-Amerika en werden de latente factor(en) gelijk beoordeeld over beide regionale groepen (Byrne & Stewart, 2006; Byrne & van de Vijver, 2010). Met andere woorden, deelnemers uit Latijns-Amerika en Europa interpreteerden de SJT-scenario's en reactie-opties op dezelfde manier en gaven er dezelfde betekenis aan (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Daarom laat dit hoofdstuk de waarde zien van de gecombineerde emic-etic benadering van de ontwikkeling van SJT's om SJT-scores te verkrijgen die kunnen worden gebruikt om deelnemers in verschillende regio's en culturen te vergelijken.

Hoofdstuk 3 ging over doelstelling 2. Het definieerde Multiple Speed Assessments als een overkoepelende term voor verschillende beoordelingsbenaderingen die voortbouwen op meerdere, korte interpersoonlijke simulaties om inzicht te krijgen in het gedragsrepertoire van deelnemers in een vooraf bepaald domein. Het vat de gemeenschappelijke kenmerken van Multiple Speed Assessments verder samen als (a) het toepassen van meerdere simulaties (bv. 20), (b) het toepassen van korte simulaties (vaak minder dan 5 minuten), (c) het toepassen van gestructureerde simulaties, (d) het stroomlijnen van evaluaties van prestaties, en (e) het integreren van de simulaties in een overkoepelende achtergrond. Als de theoretische grondslagen van Multiple Speed Assessments, identificeerde dit hoofdstuk het zero acquaintance/thin slices paradigma (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992; Back & Nestler, 2016; Connelly & Ones, 2010), trait activation theory (Lievens, Tett, & Schleicher, 2009; Tett & Burnett, 2003) en het principe van aggregatie (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014). Over verschillende

toepassingsgebieden en verschillende ontwerpvariaties zijn Multiple Speed Assessments bedoeld om het algehele gedrag in verschillende situaties of de intra-individuele variabiliteit van de deelnemers in verschillende situaties te beoordelen. Door dit overzicht van Multiple Speed Assessments te geven, verbindt hoofdstuk 3 verschillende velden met elkaar. Om onderzoek naar en meer kennis over Multiple Speed Assessments verder te bevorderen, stelde hoofdstuk 3 verder een agenda voor toekomstig onderzoek voor.

Hoofdstuk 4 ging over doelstelling 3. Een face-to-face variatie van Multiple Speed Assessments waarin voorbeelden uit het leiderschapsdomein werden verwerkt, werden ontwikkeld en onderzocht op betrouwbaarheid en validiteit. De verschillende simulaties vereisten van deelnemers om goed te presteren in interpersoonlijke situaties en zich aan te passen aan diverse (interpersoonlijke) eisen, omdat deelnemers in de verschillende simulaties geconfronteerd werden met differente rollenspelers. Voor prestatiebeoordelingen van individuele beoordelaars (rollenspelers) in individuele korte simulaties, toonden de resultaten een lage tot matige interbeoordelaarsbetrouwbaarheid. Wanneer de prestatiebeoordelingen echter werden geaggregeerd over drie tot vier beoordelaars per simulatie, waren de interbeoordelaarsbetrouwbaarheid matig tot hoog. Verder toonde de ontleding van verschillende variantie componenten in prestatiebeoordelingen aan dat het merendeel van de betrouwbare variantie in Multiple Speed Assessments een interactie-effect van deelnemer x simulatie weerspiegelt, wat aangeeft dat verschillen in prestaties van deelnemers variëren tussen verschillende simulaties. Met betrekking tot de validiteit is een vergelijkbaar patroon naar voor gekomen. Dat wil zeggen, alleen wanneer beoordelingen over alle simulaties werden geaggregeerd, volgden de resultaten de theoretisch afgeleide hypothesen met betrekking tot cognitieve vaardigheden, persoonlijkheid, en prestatiebeoordelingen van instructeurs en collega's. De relaties waren eveneens matig tot hoog. Een dergelijke Multiple Speed Assessment-score voegde verder incrementele validiteit toe bij het voorspellen van de

prestatiebeoordelingen van instructeurs en collega's naast traditionele voorspellers die gebruikmaken van een soortgelijk domein. Dit hoofdstuk illustreert daarom de sleutelrol van het aggregatieprincipe in Multiple Speed Assessments (Eisenkraft, 2013; Epstein, 1979; Kuncel & Sackett, 2014). Verder illustreert dit hoofdstuk dat Multiple Speed Assessments goede informatie bieden voor beoordelaars om prestatiebeoordelingen in een bepaald domein te geven, omdat Multiple Speed Assessments verschillende situationele eisen representeren via meerdere simulaties die aansluiten bij kwalitatief verschillende aspecten van het criteriumdomein. Daardoor optimaliseren Multiple Speed Assessments de correspondentie tussen voorspeller en criterium via meerdere korte simulaties die een efficiënte vorm van de benadering van gedragssteekproeven weerspiegelen (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968).

Ten slotte ging hoofdstuk 5 in op doelstelling 4. Het onderzocht de intra-individuele variabiliteit van deelnemers en rollenspelers in interpersoonlijk gedrag en interpersoonlijke dynamiek in vier high-fidelity simulaties. Het interpersoonlijk gedrag van deelnemers en rollenspelers varieerde significant op het momentane niveau in alle high-fidelity simulaties. Verder leek deze intra-individuele variabiliteit niet alleen een weerspiegeling te zijn van meetfout, omdat tijdelijk interpersoonlijk gedrag van deelnemers en rollenspelers werd meegevoerd in overeenstemming met de principes van complementariteit. Dat wil zeggen, de interpersoonlijke dynamiek tussen deelnemers en rollenspelers volgde consequent de principes van correspondentie in affiliatie en wederkerigheid in dominantie op het momentane niveau in alle vier high-fidelity simulaties. Ten slotte voorspelde de mate waarin deelnemers de principes van complementariteit volgden in de vier high-fidelity simulaties op het momentane niveau prestatiebeoordelingen in de high-fidelity simulaties. Ook beoordelingen van interpersoonlijke aanpasbaarheid en taakprestaties in interpersoonlijke contexten geleverd door instructeurs en collega's werden voorspeld door de mate waarin deelnemers de principes

van complementariteit volgden. Dit hoofdstuk breidt daarom het begrip intra-individuele variabiliteit *over* high-fidelity simulaties (Gibbons & Rupp, 2009; Jackson, Michaelides, Dewberry, & Kim, 2016; Lance, 2008; Lievens, 2009; Putka & Hoffman, 2013) uit tot het niveau van intraindividuele variabiliteit *binnen* high-fidelity simulaties. Het hoofdstuk draagt verder bij aan onze kennis over de aard en relevantie van continue, tijdelijke interpersoonlijke dynamiek in high-fidelity simulaties. Vooral de continue interpersoonlijke dynamiek in high-fidelity simulaties die de principes van complementariteit volgt, kan een nieuwe invalshoek bieden om interpersoonlijke aanpasbaarheid te beoordelen (Oliver & Lievens, 2014). Op al deze verschillende manieren benadrukt hoofdstuk 5 het belang van het beoordelen van (interpersoonlijk) gedrag op het momentane niveau (Gabriel, Diefendorff, Bennett, & Sloan, 2017; Jebb & Tay, 2017).

Kortom, dit proefschrift draagt bij aan onze kennis over hoe de prestaties van individuen in interpersoonlijke situaties en het aanpassen aan verschillende (interpersoonlijke) eisen kunnen worden beoordeeld via low-fidelity simulaties zoals SJT's en high-fidelity simulaties zoals Multiple Speed Assessments. Op deze manier voorziet dit proefschrift organisaties van waardevolle kennis om individuen te beoordelen, selecteren of ontwikkelen om de interpersoonlijke uitdagingen van de huidige werkomgeving met succes te beheersen.

Referenties

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

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.

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

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

Byrne, B. M., & Stewart, S. M. (2006). TEACHER'S CORNER: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 287–321. https://doi.org/10.1207/s15328007sem1302_7

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107–132. https://doi.org/10.1080/15305051003637306

Cascio, W. F. (2003). Changes in workers, work, and organizations. In R. J. Klimoski, W. C. Borman, & D. R. Ilgen (Eds.), *Handbook of Psychology* (Vol. 12, pp. 401–422). Hoboken, NJ: Wiley & Sons.

Cheung, F. M., Fan, W., Cheung, S. F., & Leung, K. (2008). Standardization of the cross-cultural Chinese Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment. *Acta Psychologica Sinica, 40,* 839–852. http://dx.doi.org/10.3724/SP.J.1041.2008.01639

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

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

Gabriel, A. S., Diefendorff, J. M., Bennett, A. A., & Sloan, M. D. (2017). It's about time: The promise of continuous rating assessments for the organizational sciences. *Organizational Research Methods*, *20,* 32–60. https://doi.org/10.1177/1094428116673721

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

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal*, *50*, 327–347. https://doi.org/10.5465/amj.2007.24634438

Hesketh, B., & Neal, A. (1999). Technology and performance. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 21–55). San Francisco, CA: Jossey-Bass.

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

Javidan, M., Dorfman, P. W., de Luque, M. S., & House, R. J. (2006). In the eye of the beholder: *Academy of Management Perspectives*, *20*, 67–90. https://doi.org/10.5465/AMP.2006.19873410

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*, 61–94. https://doi.org/10.1177/1094428116668035

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

Kozlowski, S. W. J., Gully, S. M., Salas, E., & Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In M. Beyerlein, S. Beyerlein, & D. Johnson (Eds.), *Advances in interdisciplinary studies of work teams: Team leadership* (Vol. 3, pp. 251–289). Greenwich, CT: JAI Press.

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

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

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*, 102–121. https://doi.org/10.1080/13594320802058997

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

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

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

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, *85*, 612–624. https://doi.org/10.1037//0021-9010.85.4.612

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

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53,* 153–193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x

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

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://doi.org/10.1177/109442810031002

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372-376. http://dx.doi.org/10.1037/h0026244

Zeithaml, V. A., & Bitner, M. J. (1996). *Services marketing*. New York: McGraw-Hill.

DATA STORAGE FACT SHEETS

Data Storage Fact Sheet 1

Name/identifier study: Dissertation chapter 2; Herde et al. (2019)
Author: Christoph N. Herde
Date: 27.08.2019

1. Contact details
==========

1a. Main researcher

- name: Christoph N. Herde
- address: Ghent University, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium
- e-mail: christoph.herde@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Filip Lievens
- address: Singapore Management University, Lee Kong Chian School of Business
50 Stamford Road, 178899 Singapore
- e-mail: filip.lievens@smu.edu.sg

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

==========

* Reference of the publication in which the datasets are reported:

Herde, C. N., Lievens, F., Solberg, E. G., Harbaugh, J. L., Strong, M. H., & Burkholder, G. J. (2019). Situational judgment tests as measures of 21st century skills: Evidence across Europe and Latin America. *Journal of Work and Organizational Psychology, 35*, 65-74. https://doi.org/10.5093/jwop2019a8 Dissertation chapter 2

* Which datasets in that publication does this sheet apply to?:

This sheet applies to all data reported in the above mentioned publications

3. Information about the files that have been stored

==========

3a. Raw data

---

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO

If NO, please justify:

* On which platform are the raw data stored?

- [X] researcher PC that automatically creates back-ups
- [ ] research group file server
- [X] other (specify): computers/servers of members of SHL who developed the tests and Laureate International Universities from which study participants were recruited.

* Who has direct access to the raw data (i.e., without intervention of another person)?

- [X] main researcher
- [ ] responsible ZAP
- [ ] all members of the research group
- [ ] all members of UGent
- [X] other (specify): Members of SHL and Laureate International Universities. Copyright of the data are owned by these third parties. Thus, data cannot be shared publicly without prior explicit approval.

3b. Other files

---

* Which other files have been stored?
 - [X] file(s) describing the transition from raw data to reported results. Specify: excel sheets, SPSS syntax files
 - [X] file(s) containing processed data. Specify: data sets (SPSS data sets, text files) ready for analyses
 - [X] file(s) containing analyses. Specify: SPSS syntax files, Mplus input and output files
 - [ ] files(s) containing information about informed consent
 - [ ] a file specifying legal and ethical provisions
 - [X] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: word and excel documents describing these details
 - [ ] other files. Specify: ...

* On which platform are these other files stored?
 - [X] individual PC that automatically creates back-ups
 - [ ] research group file server
 - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
 - [X] main researcher
 - [ ] responsible ZAP
 - [ ] all members of the research group
 - [ ] all members of UGent
 - [ ] other (specify): ...

4. Reproduction

==========

* Have the results been reproduced independently?: [ ] YES / [X] NO

Data Storage Fact Sheet 2

Name/identifier study: Dissertation chapter 4
Author: Christoph N. Herde
Date: 27.08.2019

1. Contact details
==========

1a. Main researcher

- name: Christoph N. Herde
- address: Ghent University, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium
- e-mail: christoph.herde@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Filip Lievens
- address: Singapore Management University, Lee Kong Chian School of Business
50 Stamford Road, 178899 Singapore
- e-mail: filip.lievens@smu.edu.sg

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========

* Reference of the publication in which the datasets are reported:
Herde, C. N., & Lievens, F. (2019). Multiple Speed Assessments Under Scrutiny: Are Their Ratings Reliable and Valid? Chapter 4. (Doctoral dissertation). Ghent University, Ghent, Belgium

* Which datasets in that publication does this sheet apply to?:
This sheet applies to all data reported in the above mentioned publications

3. Information about the files that have been stored
==========
3a. Raw data
---

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
- [ ] researcher PC
- [ ] research group file server
- [X] other (specify): main researchers UGent network drive and UGent webshare

* Who has direct access to the raw data (i.e., without intervention of another person)?
- [X] main researcher
- [X] responsible ZAP
- [ ] all members of the research group
- [ ] all members of UGent
- [X] other (specify): Members of Hudson who developed the Multiple Speed Assessment and contributed to the data collection have access to all data except for ratings of independent assessors. Copyright of the data except for ratings of independent assessors is owned by Hudson. Thus, data cannot be shared publicly without prior explicit approval.

3b. Other files

---

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results. Specify: excel sheets, SPSS syntax files
  - [X] file(s) containing processed data. Specify: data sets (SPSS files, text files, SAS data sets): ready for analyses
  - [X] file(s) containing analyses. Specify: SPSS syntax files, Mplus input and output files, R scripts
  - [X] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [X] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: word and excel documents describing these details
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [ ] individual PC
  - [ ] research group file server
  - [X] other: main researchers UGent network drive and UGent webshare

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction

==========

* Have the results been reproduced independently?: [ ] YES / [X] NO

Data Storage Fact Sheet 3

Name/identifier study: Dissertation chapter 4
Author: Christoph N. Herde
Date: 27.08.2019

1. Contact details
==========

1a. Main researcher

- name: Christoph N. Herde
- address: Ghent University, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium
- e-mail: christoph.herde@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Filip Lievens
- address: Singapore Management University, Lee Kong Chian School of Business
50 Stamford Road, 178899 Singapore
- e-mail: filip.lievens@smu.edu.sg

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========

* Reference of the publication in which the datasets are reported:
Herde, C. N., & Lievens, F. (2019). A closer look at Intraindividual Variability in Interpersonal Behavior and Interpersonal Dynamics in High-Fidelity Simulations. Chapter 5. (Doctoral dissertation). Ghent University, Ghent, Belgium

* Which datasets in that publication does this sheet apply to?:
This sheet applies to all data reported in the above mentioned publications

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
- [ ] researcher PC
- [ ] research group file server
- [X] other (specify): main researchers UGent network drive and UGent webshare

* Who has direct access to the raw data (i.e., without intervention of another person)?
- [X] main researcher
- [X] responsible ZAP
- [ ] all members of the research group
- [ ] all members of UGent
- [X] other (specify): Members of Hudson who developed the Multiple Speed Assessment and contributed to the data collection have access to all data except for ratings of independent assessors and ratings of momentary interpersonal behavior.
Copyright of all data except for ratings of independent assessors and ratings of momentary behavior is owned by Hudson. Thus, these data cannot be shared publicly without prior explicit approval.

3b. Other files

---

* Which other files have been stored?
 - [X] file(s) describing the transition from raw data to reported results. Specify: excel sheets, SPSS syntax files
 - [X] file(s) containing processed data. Specify: data sets (SPSS files, text files, SAS files, R objects) ready for analyses
 - [X] file(s) containing analyses. Specify: SPSS syntax files, Mplus input and output files, R scripts
 - [X] files(s) containing information about informed consent
 - [ ] a file specifying legal and ethical provisions
 - [X] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: word and excel documents describing these details
 - [ ] other files. Specify: ...

* On which platform are these other files stored?
 - [ ] individual PC
 - [ ] research group file server
 - [X] other: main researchers UGent network drive and UGent webshare

* Who has direct access to these other files (i.e., without intervention of another person)?
 - [X] main researcher
 - [X] responsible ZAP
 - [ ] all members of the research group
 - [ ] all members of UGent
 - [ ] other (specify): ...

4. Reproduction

==========

* Have the results been reproduced independently?: [ ] YES / [X] NO