Measuring Distinct Social Skills via Multiple Speed Assessments – A Behavior-Focused Personnel Selection Approach

Simon M. Breil[1], Boris Forthmann[1], & Mitja D. Back[1]

[1]University of Münster, Germany

This is an unedited manuscript accepted for publication in the European Journal of Psychological Assessment. The manuscript will undergo copyediting, typesetting, and review of resulting proof before it is published in its final form.

Author Note

We embrace the values of openness and transparency in science (www.researchtransparency.org). Therefore, we follow the 21-word solution (Simmons et al., 2012). All raw data and scripts that are necessary to reproduce the results reported in this manuscript can be found at osf.io/jy5wd.

This work was supported by the Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research, Germany; project number: 01GK1801B)

Acknowledgments: We thank Bernhard Marschall, Helmut Ahrens, Anike Hertel-Waszak, Eva Schönefeld, and Britta Brouwer for their help collecting the data for this study. Furthermore, we thank Kirsten Gehlhar, Leonie Hater, Christoph Herde, Mirjana Knorr, and Theresa Seifert for their insightful comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Simon Breil (simon.breil@wwu.de, University of Münster, Institute for Psychology, Psychological Assessment and Personality Psychology, Fliednerstr. 21, 48149 Münster, Germany).

Abstract

Social skills (e.g., persuading others, showing compassion, staying calm) are of key importance in work and education settings. Accordingly, the goal of many selection processes is to identify candidates who excel in desired skills. For this, high-fidelity simulations such as assessment centers (ACs) are regarded as ideal procedures because they can be used to evoke, observe, and evaluate candidates' actual behavior. However, research has repeatedly shown that observed performance differences in ACs are not sufficiently driven by the specific skill dimensions that are defined for assessment. Building on multiple speed assessments and incorporating insights from behavioral personality science, we offer an alternative approach for the reliable and valid assessment of distinct social skills. We hereby (a) selected skills on the basis of a bottom-up analysis of observable and distinguishable interpersonal behaviors and (b) specifically designed exercises around these skills (i.e., one skill per exercise, multiple exercises per skill). Here, we present initial results of this newly developed procedure across three samples in a high-stakes selection context (N = 589). Generalizability theory analyses showed that a substantial amount of variance in assessor ratings could be attributed to the selected skills. This underlines the importance of more behaviorally focused selection procedures.

Keywords: multiple speed assessments, assessment center, behavioral personality science, social skills, interpersonal behavior

Measuring Distinct Social Skills via Multiple Speed Assessments – A Behavior-Focused Personnel Selection Approach

Social skills (i.e., the entire range of skills that promote effective functioning in interpersonal situations; interpersonal skills) are of key importance in work and education settings as they predict relevant outcome criteria such as job performance and academic success (e.g., Arthur et al., 2003; Lievens & Sackett, 2012). Thus, many organizations aim to identify and select individuals who excel in desired social skills (e.g., persuading others, showing compassion, staying calm). The prime methods for this are high-fidelity simulations such as assessment centers (ACs) as they can be used to evoke, observe, and evaluate individuals' actual social behavior. This is based on the general idea that by observing behavioral expressions in several relevant situations (i.e., situations that call for specific behavioral responses), one can draw conclusions about individuals' social skills (Kanning, 2009). However, years of research have shown that ACs are currently not particularly suited to reliably measure *distinct* social skills. Performance differences in ACs mainly depend on the simulation exercises that are used but not on the social skill dimensions that are defined for assessment (e.g., Jackson et al., 2016; Putka & Hoffman, 2013). That is, if decision makers aim to assess the skills to persuade others, to show compassion, and to stay calm via a typical AC procedure, they would generally not be able to reliably identify individuals' distinct performances on these skills. In this article, we aim to take a closer look at this phenomenon by focusing on how skill selection and exercise creation might influence the dimensionality and reliability of social skill assessment. We then present initial results concerning the development of a new procedure (based on the multiple speed assessment approach; Herde & Lievens, 2020) that incorporates insights from behavioral personality science and directly focuses on assessing distinct social skills.

Assessment of Social Skills via ACs

In ACs, assessees (i.e., individuals who are evaluated) face a variety of different exercises (e.g., role-plays, group discussions) and are observed and evaluated by different assessors (i.e., individuals who evaluate) on a variety of (often social) skill dimensions (for a recent overview, see Kleinmann & Ingold, 2019). Given the lack of a parsimonious and consensual overarching framework of social skills, the selection of skills is mostly based on idiosyncratic choices that depend on the work or educational context. This results in a myriad of differently labeled social skills in research and practice (Arthur et al., 2003; Klein et al., 2006). Whereas ACs have repeatedly been shown to reliably capture between-person differences that predict important outcomes (e.g., Sackett et al., 2017), the selected skill dimensions seem to play a negligible role. This has most recently been investigated by using generalizability theory (Brennan, 2001), in which the impact of different AC components (e.g., skill dimensions, assessee main effects, assessor effects) and their interactions are estimated by identifying how much they contribute to the variance in assessor ratings. Here, the amount of variance that can be attributed to differences in skill dimension-specific performance is typical low (Jackson et al., 2016; Putka & Hoffman, 2013). This is due to (a) little differentiation between different skills within an exercise (low discriminant validity) and (b) no convergence between the same skills across exercises (low convergent validity; Bowler & Woehr, 2006). That is, high ratings on persuading others within one exercise would typically covary with high ratings on showing compassion within the same exercise. Across exercises, there would be little consistency in persuasiveness or compassion ratings. Reasons for this phenomenon have been discussed extensively (see Lance, 2008 and associated commentaries) and can be broken down into two main arguments: potential *assessor biases* and potential *inconsistencies in assessee behavior*.

Assessor biases refer to the idea that assessors are not sufficiently skilled or have too much cognitive load. Along with the fact that assessor and exercise variance are often

confounded (i.e., an assessor typically does not evaluate each assessee on each exercise), this might lead to inaccurate and undifferentiated ratings. And of course, training assessors (e.g., frame of reference) and carefully designing ACs (e.g., including behavioral checklists, reducing the number of skill dimensions) will undoubtedly improve the reliability and validity of assessors' ratings (Lievens, 2009). However, years of research have shown that assessors are actually quite accurate when it comes to observing and evaluating interpersonal behavior (Back & Nestler, 2016; Lievens, 2009). For example, Lievens (2002) showed that when assessees were specifically instructed to act in a cross-situationally consistent manner, assessors picked up on this pattern and rated the assessees accordingly. Thus, a sole focus on assessors cannot explain the inability to reliably assess distinct skills in ACs.

Some authors (e.g., Lance, 2008) have built on this argument (i.e., that assessors' ratings are not generally flawed) and suggested that it must be assessees' behavior that is *inherently* cross-situationally inconsistent. This reasoning is backed by the general finding that performance differences between assessees depend strongly on the situational demands of the respective exercise (e.g., some assessees just perform better in Exercise A, whereas other assessees perform better in Exercise B; see Lance et al., 2004; Lievens et al., 2009; Lievens, 2009). However, AC research has predominantly analyzed assessors' performance ratings and has not considered actual behavioral variation. That is, just because ratings on the chosen skills are undifferentiated within exercises and inconsistent across exercises does not mean that the same holds for the actual expressed behaviors. For example, even if behavior is expressed consistently across exercises (e.g., individuals who make more supportive statements in Exercise A also make more supportive statements in Exercise B), the resulting skill ratings (e.g., on a compassion dimension) may be inconsistent. This could be due to the use of exercise-specific rating sheets (i.e., the resulting skill ratings might not actually be based on the same behaviors) or the case if resulting skill ratings are influenced by other aspects of the exercises (e.g., whether the overall exercise performance was good or bad).

In fact, a recent study (Breil, Lievens, et al., 2021) showed that basic behavioral differences (e.g., warm behavior, assertive behavior, interpersonal calm behavior) can be differentiated in interpersonal AC exercises and are actually quite stable across AC exercises. This is supported by decades of research from behavioral personality science that has shown moderate to high behavioral consistency in a variety of behavioral domains across many different tasks and situations (e.g., Borkenau et al., 2004; Furr & Funder, 2004; Leikas et al., 2012; Sherman et al., 2010). These results indicate that (a) the unreliable assessment of distinct skills is not solely rooted in inconsistent assessee behavior, and (b) the variety of behaviors that are evoked by different exercises are not necessarily the behaviors that are being evaluated. Thus, it might rather be the typical way in which ACs are designed (i.e., the interplay of behaviors, skill selection, and exercise creation) that limits the interpretability of skill dimensions.

Interplay of Behaviors, Skills, and Exercises

The skills included in ACs are foremost based on job analysis or competency modeling and reflect bundles of behaviors that are deemed important (Lievens, 2017; Rupp et al., 2015). Even though skills are sometimes organized according to theoretical frameworks (e.g., Arthur et al., 2003; Meriac et al., 2014), the typical way of selecting skills is not without its drawbacks as it does not take the empirical structure of behaviors into account (Lance et al., 2004). On the one hand, the skills that are defined for assessment (e.g., compassion, empathy, social sensitivity) might share most of their underlying behaviors (jangle fallacy). On the other hand, one skill (e.g., communication) might be represented by different behaviors depending on the exercise (jingle fallacy). Thus, there might be little differentiation between different skills because the underlying behaviors are often identical or overlap, and there might be no consistency across the same skills because the underlying behaviors differ (across exercises; see Lievens et al., 2009; Melchers et al., 2012).

These effects will likely be further exaggerated due to the typical selection and creation of exercises. Most AC exercises are designed for their face validity (e.g., tasks related to the job) rather than for their capacity to expose relevant behavioral differences (Brannick, 2008; Lievens & Klimoski, 2001). That is, it is unclear how well the variety of different AC exercises that are used can evoke reliable differences in desired behaviors. For example, if one aims to assess persuasiveness within an exercise in which no one must be persuaded (e.g., because the role-player is instructed to act submissive), there will be little reliable behavioral variance related to persuasiveness (see also research on trait relevance and situational strength: Lievens et al., 2006; Meyer et al., 2010). Furthermore, specific aspects of exercises might actively increase the co-dependence of skill ratings. That is, within ACs it is mainly the exercise outcome that is evaluated (Brannick, 2008; Lance et al., 2004) and if there is only one way to "solve" an exercise, assessees are likely to receive high/low ratings on all skills assessed with this exercise. Also, if good performance on an exercise can only be achieved via exercise-specific behaviors (e.g., the interaction partner will only be persuaded by putting forth a content specific argument), the cross-situational consistency of skill ratings will decrease.

Implications for Assessment Designs

If one aims to assess distinct social skills, the presented findings have several implications for assessment design. That is, skill selection should not only be based on job analysis but also acknowledge empirical findings concerning the structure of behavioral expression. Accordingly, one should consider how behaviors that are expressed in interpersonal interactions will cluster. Whereas this (bottom-up) approach has received little attention in selection settings, it has been a key focus of behavioral personality science. Here, results suggest that most interpersonal behavior can be represented by the two underlying factors of agency (i.e., dominant and assertive behavior) and communion (i.e., warm and friendly behavior; Dawood et al., 2018; Wiggins, 1979). These two factors are often

supplemented by a third factor pertaining to interpersonal calmness (emotionally stable and relaxed behavior; Leising & Bleidorn, 2011), which seems to be especially profound in stressful situations such as assessment settings (Hirschmüller et al., 2015). All three behavioral domains have been shown to emerge across a variety of interpersonal settings and tend to be quite stable across situations (e.g., Borkenau et al., 2004; Leikas et al., 2012). Thus, we suggest (a) not assessing multiple skills per behavioral domain (e.g., compassion and empathy, both of which pertain primarily to communal behavior) and (b) not assessing skills that overlap across the behavioral domains (e.g., broad concepts such as "leadership" will likely include communal and agentic behaviors; thus, for a more distinct assessment of skills, it would be preferable to assess these leadership facets separately).

When it comes to exercises, it is clear that assessees will vary in how they approach each of these specific situations and solve the implied specific tasks. This will undoubtfully impact ratings on all skills assessed with the respective exercise (Lance et al., 2004; Lance, 2008). Thus, it is important to build exercises in such a way that high performance on an exercise can only be achieved by individuals who express skill-related behaviors. For example, if one aims to assess candidates' compassion with an exercise that involves delivering unpopular news, it is important that the exercise is designed in a way that a positive outcome can only be achieved through compassionate behaviors (e.g., active listening, supportive statements) and not with other behaviors (e.g., using specific arguments, controlling the interaction). However, doing this for multiple skills within an exercise would likely result in the co-occurrences of related behaviors leading to little differentiation between skills. For this reason, we propose that *only one skill in each exercise* be assessed and that the exercises be built exclusively around the chosen skill (see Brannick, 2008, for similar reasoning). In line with axioms of trait activation theory (Tett & Guterman, 2000), this would involve using multiple skill-specific environmental cues that directly trigger the expression of skill-relevant behaviors (e.g., having someone ask for help will likely trigger differences in

communal behavior). For a more reliable assessment of skills, it would then be beneficial to use multiple exercises per skill (i.e., have exercises nested within skills, e.g., three exercises to assess compassion, three additional exercises to assess persuasiveness). This would not necessarily mean having very similar exercises (which might negatively impact validity; Speer et al., 2014). Instead, it allows for a relatively comprehensive assessment of each skill expressed across different contexts that represent the variety of specific environmental cues that trigger individual differences in the skill of interest.

A procedure that is ideal for such an approach involves multiple speed assessments (Brannick, 2008; Herde & Lievens, 2020). In contrast to classic ACs, multiple speed assessments revolve around a larger number of very short interpersonal simulations that elicit overt behavior in a standardized way. The basic idea behind multiple speed assessments is that (a) even brief behavioral observations provide assessors with enough information to make relatively accurate judgments (Breil, Osterholz, et al., 2021; Funder, 2012; Ingold et al., 2018) and (b) by increasing the number of situations/exercises (i.e., sampling behavior across a variety of contexts), the reliability and validity of the overall ratings increase (i.e., principle of aggregation; Epstein, 1983).

Present Research

With the current research, we aimed to conduct a first empirical investigation of a more behavior-focused approach that directly addresses the presented implications. Specifically, we tested whether the selection of skills according to the structure of cross-situationally consistent behavioral differences and the creation of exercises around these specific skills would allow for a more reliable assessment of distinct social skills. To do so, we used a multiple speed assessment procedure to identify the best applicants for a spot in medical school across three samples. We hereby chose to assess the three social skills of *warmth*, *assertiveness*, and *resilience* that were found to be desired skills for future physicians (Hertel-Waszak et al., 2017) and that exclusively revolve around behaviors from the three

domains of agency (*assertiveness*), communion (*warmth*), and interpersonal calmness (*resilience*). For each of these skills, we designed two exercises that involved multiple cues to specifically evoke skill-related behaviors.

Here, the basic premise was to create exercises in which specific behavioral responses were needed to perform effectively. For example, within the exercises revolving around assertiveness, agentic behavior was required to achieve a positive outcome. Assessees with a high level of assertiveness skill were expected to (a) *know* that agentic behavior is needed and (b) *act* accordingly. By observing interindividual differences in expressed agentic behavior (within the relevant exercises), it was then possible to draw conclusions about assessees' assertiveness skill levels. Such assessed individual differences in social skills refer to maximal performance (i.e., what someone is capable of doing when it matters) and should not be equated with individual differences in personality traits (i.e., what someone tends to do in general, typical performance; Breil et al., 2017; Soto et al., in press).

To investigate the distinct assessment of the selected social skills, we analyzed the reliability of the skill ratings, the correlations between the skill ratings, and variance components (i.e., identifying the amount of variance that could be attributed to different AC components; e.g., individual differences depending on skill dimensions). For a more reliable assessment of distinct social skills, one would expect that (a) there would be at least acceptable reliability in the individual skill ratings, (b) the correlations between ratings on exercises belonging to the same skill would be higher than the correlations between ratings on exercises belonging to different skills, and (c) there would be a substantial amount of variance in assessor ratings that could be attributed to individual differences in the different social skills.

Method

Participants

Assessees

The data were based on three independent selection procedures that took place 6 months apart. Out of an initial 652 applicants, 589 (age: M = 18.86, SD = 1.43; 406 female) gave informed consent for further data analysis (Sample 1: 202; Sample 2: 191; Sample 3: 196). All assessees applied for a place in medical school (human medicine N = 436; dentistry N = 153) and were preselected by their GPA. The university's institutional review board approved this study. Further information (data, R code, supplemental material) can be found on the projects' Open Science Framework page (osf.io/jy5wd).

Assessors

In each selection sample, up to 60 professional physicians evaluated the assessees. The assessors were allocated to teams of two, assigned to one exercise and one room (six teams/rooms per exercise per sample), and evaluated up to 40 assessees. That is, each assessee was evaluated by two assessors per exercise, and there were different assessors for each exercise and each sample (see Figure 1 for a visual representation).

Procedures

The multiple speed assessment was developed in a manner similar to other approaches (e.g., Herde & Lievens, 2020; Knorr et al., 2018) and consists of multiple interpersonal role-play exercises. All assessors received assessor training (2 hr; see Rupp et al., 2015), which included a lecture (e.g., rater biases, separating observation and evaluation, establishing a frame-of-reference) and practice (e.g., viewing example videos followed by moderated discussion and feedback). Before each exercise, assessees had 90 s to read the instructions and then took part in the exercises, each of which lasted for 5 min. The general sequence of exercises was identical for all assessees, but the starting exercises were randomized. In the exercises, assessees interacted with role-players who all had multiple years of professional

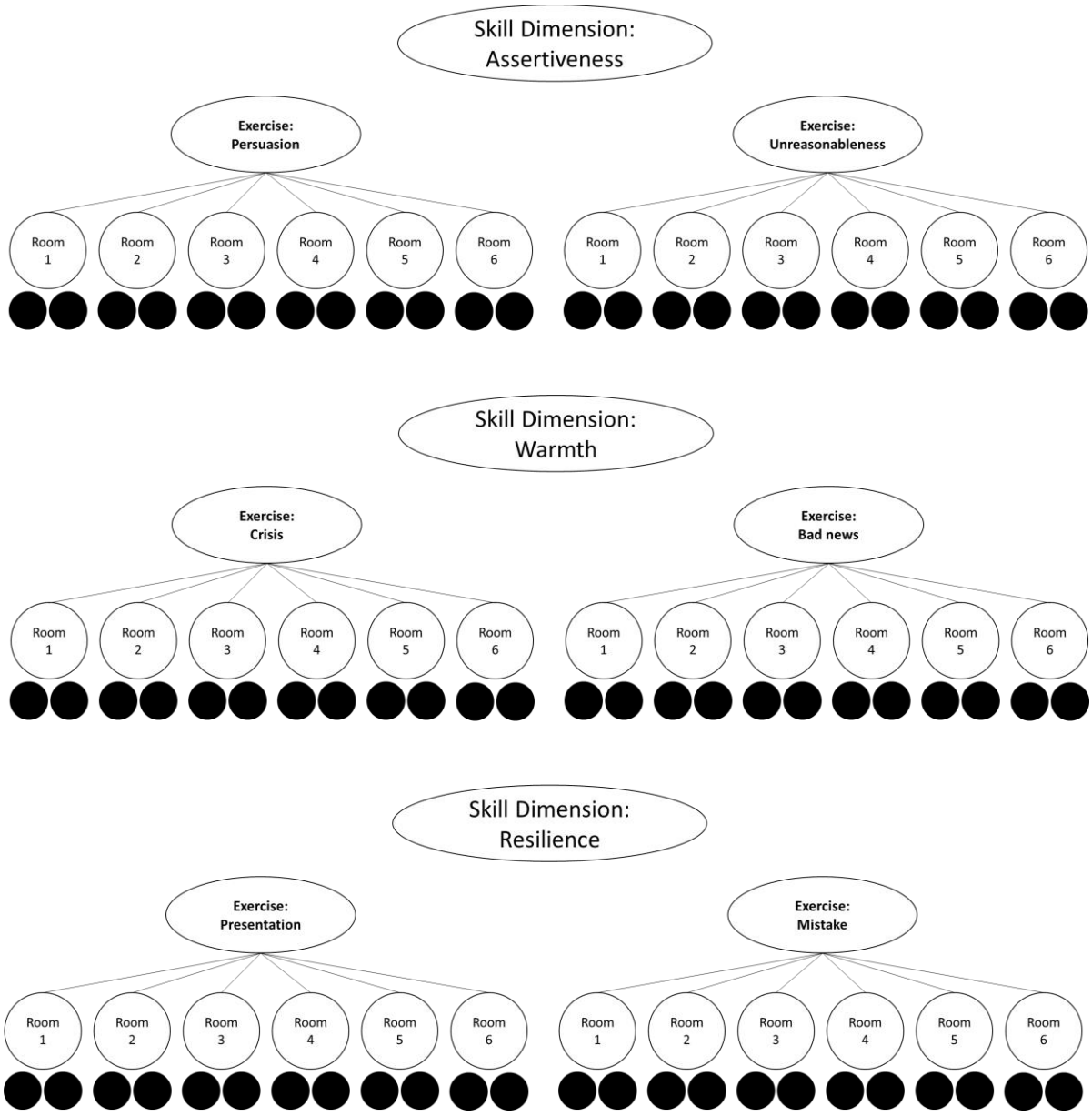
acting experience and were specifically instructed concerning their roles and behavior. At the same time, assessees were observed by assessors through a one-way mirror (assessees knew that they were being observed and evaluated). After each exercise, participants switched to the next exercise, and assessors were asked to rate assessees on the corresponding social skill dimension.

Overall, six different interpersonal exercises were used across the three samples. That is, for each of the three social skills (i.e., assertiveness, warmth, resilience), two exercises were developed. Building on the assumptions of trait activation theory (Tett & Guterman, 2000), exercises were specifically designed to be highly relevant for the desired social skills. For this, we relied on research and findings from behavioral personality science (e.g., interpersonal theory; Dawood et al., 2018). For example, for the exercises related to resilience, we included multiple triggers in the exercises (e.g., social pressure, uncertainty) that have been shown to provoke differences in expressed behaviors related to resilience (e.g., Hirschmüller et al., 2015). Furthermore, emphasis was placed on ensuring that the exercises revolved as exclusively as possible around the respective skills and that skill-related behavior would be associated with good performance within each exercise. For an overview of all exercises and behavioral triggers see Figure 2.[1]

[1] Please note that in Sample 1, only the exercises for assertiveness and warmth were used (resulting in four exercises for the first sample). In Samples 2 and 3, only one of the assertiveness exercises (i.e., persuasion) was used (resulting in up to five exercises for the second and third samples). Furthermore, applicants who applied for dentistry participated in only one of the warmth exercises (i.e., crisis) in Sample 2 and only one of the resilience exercises (i.e., presentation) in Sample 3. This procedure may have led to the confounding of exercise and skill dimensions in some cases (e.g., the assertiveness skill dimension was only assessed via one exercise in Samples 2 and 3). To investigate the potential impact of this, we employed two alternative ways of analyzing the data. First, we excluded participants who did not have multiple skills assessed via two exercises per skill and excluded skill dimensions (within samples) that were assessed via only one exercise. Second, we used multiple imputation to consider the missing at random pattern of missing values. Neither the imputation nor the employment of the exclusion criteria substantially changed the estimation of the variance components (see Online Supplement S1, osf.io/jy5wd), so we only report results for the full sample using all observations (without imputation).

Figure 1

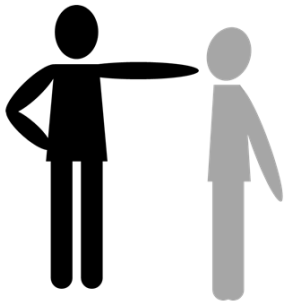
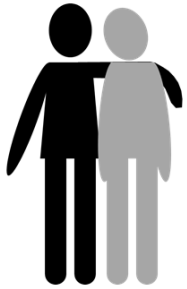
Overview of Skill Dimensions, Exercises, and Assessor Allocation



Note. Black dots represent different assessors. Rooms 1 to 4 were used for human medicine, Rooms 5 and 6 for dentistry.

Figure 2

Overview of Social Skills, Corresponding Behaviors, Triggers, and Exercises

Definition	Behavioral anchors	Interpersonal triggers	Exercises
Assertiveness The extent to which one is confident, determined, and energetic in one's actions. 	• Self-confident and upright posture; confident expressions and gestures; confident flow of speech • Leads/controls the interaction; clear statements	• Reactance • Conflicts • Decisions needed	**Persuasion:** Assessees had to pass on important information and convince someone to do something. The role-player acted distracted and unconvinced. **Unreasonableness:** Assessees had to stop someone from doing something and reason with the person. The role-player acted unreasonable and carefree.
Warmth The extent to which one treats others with love, kindness, and compassion. 	• Attentive, positive attention; friendly gestures and expressions • Active listening; positive feedback; statements of support	• Need for help • Sadness • Observed stress	**Crisis:** Assessees had to take care of someone after a crisis situation and provide support. The role-player acted overwhelmed and shocked. **Bad news:** Assessees had to deliver bad news to someone and calm and soothe the person. The role-player acted sad and insecure.
Resilience The extent to which one controls one's emotions, handles stress well, and reacts in a calm manner. 	• Relaxed position; calm expressions and gestures; does not break up sentences • No uncertain queries; no justifications; no oversensitive reactions	• Time pressure • Social stressors • Performance-related stressors • Uncertainty	**Presentation:** Assessees had to give a presentation in front of someone but had little time to prepare. The role-player acted unimpressed and cynical. **Mistake:** Assessees had to react to a mistake they made and deal with criticism and stress. The role-player acted angry and disappointed.

Measures

Assessors rated assessees on one social skill per exercise (i.e., the skill that the exercise was designed to assess) via one global rating. For the ratings of the social skill dimensions, we relied on global behaviorally anchored rating scales (e.g., Please rate the candidate's assertiveness) ranging from 0 to 5. Behavioral anchors were based on previous research concerning the selected social skills (e.g., Hirschmüller et al., 2015; Oliver et al., 2016) and were identical across the respective exercises (e.g., assertiveness: shows self-confident posture, confident flow of speech, leads the interaction).[2] For all social skills, this included a mixture of nonverbal (e.g., warmth: friendly expression) and verbal (e.g., warmth: statements of support) behaviors. The exact behavioral anchors (in a condensed form) are displayed in Figure 2.

Analytic Strategy

To quantify the extents to which the different components contributed to the amount of reliable variance (i.e., individual differences in overall performance, individual differences in skill dimension-specific performance, individual differences in exercise-specific performance), we estimated the variance components on the basis of the assessors' ratings. Similar to recent studies (e.g., Breil et al., 2020; Jackson et al., 2016; Putka & Hoffman, 2013), we decomposed the variance by specifying the assessees, assessors, skill dimensions, exercises, samples, and all their possible interactions as crossed random factors in a Bayesian linear random effects model with the R package brms (Bürkner, 2017). Bayesian approaches have the advantage of being applicable to very complex data structures with a large number of estimated variance components (see Jackson et al., 2016). A description and interpretation of the different variance components is summarized in Table 1. Please note that (per design)

[2] The reliability and validity of the selected behaviors were tested in another study (see AUTHORS, 2021). Here, it was shown that these behaviors were observable within interpersonal role-plays, could be clearly assigned to distinguishable behavioral factors, were expressed relatively consistently across different exercises, and predicted future interpersonal performance.

exercises were nested within dimensions, *and* assessors were nested within exercises. This resulted in fewer variance components in comparison with studies with more classical AC designs (e.g., designs in which exercises and dimensions were crossed; e.g., Jackson et al., 2016). Data and R code can be found at osf.io/jy5wd.

Table 1

Decomposition of Observed Variance in Assessor Ratings

Variance component	Meaning
Reliable variance	
σ^2 assessee	Individual differences in overall performance: *Some assessees perform better than others irrespective of exercises or skill dimensions.*
σ^2 assessee × skill dimension	Individual differences in skill dimension-specific performance: *Some assessees perform better on some skill dimensions than others irrespective of exercises.*
σ^2 assessee × exercise	Individual differences in exercise-specific performance: *Some assessees perform better on some exercises (nested in skill dimensions) than others.*
Unreliable variance	
σ^2 assessee × assessor	Assessor disagreement: *Some assessees are rated higher by some assessors (nested in exercises and skill dimensions) than others.*
σ^2 assessor	Assessor main effects: *Some assessors (nested in exercises and skill dimensions) tend to give more lenient/severe ratings than others.*
σ^2 other components	Exercise, skill dimension, sample main effects: *Some exercises, skill dimensions, or samples (and their interactions) have higher average ratings.* *Here, this does not influence reliability.*

Note. This table was adapted from Putka and Hoffman (2013).

Results

Table 2 presents descriptive statistics and reliabilities (based on the $G(q,k)$ estimator; Putka et al., 2008).[3] All average interrater reliability estimates (calculated between the two assessors per exercise) indicated acceptable to good reliability according to common cut-offs ($G(q,2)_{mean}$ = .69; e.g., Cicchetti, 2001). The correlations between the skill ratings on the different exercises showed that ratings on exercises revolving around the same social skills were more strongly correlated (r_{mean} = .23) than ratings on exercises referring to different social skills (r_{mean} = .10). This difference was significant ($r_{\text{difference}}$ = .14; p = .016, 95% CI [.03, .24])[4]

Table 3 (column 2) shows the decomposition of variance components into reliable variance and unreliable variance (see Table 1 for more information). Column 3 shows the percentages of variance rescaled within the reliable components. Furthermore, column 4 shows the percentage of reliable variance when taking the aggregation level into account (i.e., given that final skill dimension scores are typically based on aggregated ratings from multiple exercises). The largest amount of reliable variance was accounted for by assessee-exercise effects (i.e., some assessees performed better than others, depending on the exercise, 62%). The second largest amount was explained by assessee-dimension effects (i.e., some assessees performed better than others, depending on the skill dimension, 20%). Assessee main effects (i.e., some assessees performed better than others irrespective of exercise or skill dimension)

[3] As additional analyses, we also investigated potential relations between skill ratings and assessees' gender and age. Results showed that females (compared to males) received higher skill ratings on both warmth exercises (crisis r = .17, bad news: r = .14). Furthermore, older assessees received slightly higher skill ratings on the mistake exercise (r = .11). Apart from these results, there were no significant gender or age effects. See Online Supplement S2 on osf.io/jy5wd for all results)

[4] For this, we relied on Fisher-z-transformed correlations of skill ratings across exercises based on multiple imputation. This was done to account for the missing at random structure (see Footnote 1). These correlations were then averaged and defined as new parameters in a structural equation model. This included correlations of ratings concerning the same social skills (Parameter 1: three correlations) and correlations of ratings concerning different social skills (Parameter 2: the remaining twelve correlations). The difference between these average correlations was then tested with standard errors based on the Delta-Method (Dorfman, 1938).

accounted for 18% of the variance. Taking aggregation level into account (i.e., dimension-level scores), the reliable variance attributable to social skill dimensions increased to 28%. On this aggregated level, 26% of the variance was accounted for by assessee main effects and 45% by assessee-exercise effects. All these results were robust across the three samples (see the Online Supplement S1 with separate results for the three samples on osf.io/jy5wd).

Table 2

Overview of Skill Dimension Ratings: Descriptive Statistics

Dimension Rating	*N*	*M*	*SD*	*G* (*q*,*k*)	2	3	4	5	6
1 Assertiveness: Persuasion	589	3.05	1.17	.76	**.36 (.39)**	.15 (.18)	.08 (.08)	.16 (.14)	.14 (.17)
2 Assertiveness: Unreasonable	202	3.37	1.21	.76		.16 (.14)	.12 (.12)	- -	- -
3 Warmth: Crisis	589	3.54	1.10	.74			**.18 (.18)**	.05 (.08)	.12 (.08)
4 Warmth: Bad news	535	3.51	0.98	.55				.00 (.04)	.09 (.02)
5 Resilience: Presentation	387	3.43	1.06	.70					**.19 (.27)**
6 Resilience: Mistake	341	3.48	0.94	.59					

Note. These results refer to skill dimension ratings across all samples (pairwise deletion). For the correlations in parentheses, the respective ratings were first standardized within each subsample (three samples, two majors) per exercise. Correlations concerning the same social skills are in bold. Correlational results obtained from multiple imputation (see Footnote 4 and R code) and based on pairwise deletion (reported here) did not differ in any meaningful way from each other.

Table 3

Variance Components and Reliability Estimates of Ratings

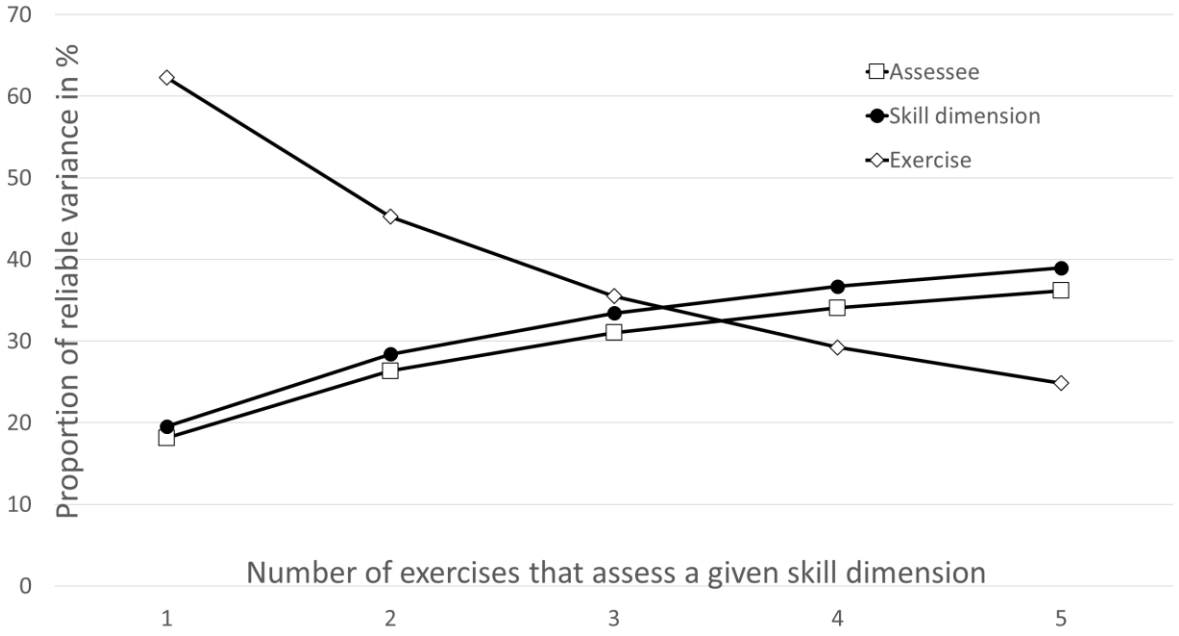
Variance component	Total between-assessee variance (%)	Reliable variance (%)	Reliable dimension variance (%)
Reliable variance			
σ^2 assessee	9.56	18.15	26.36
σ^2 assessee × skill dimension	10.30	19.55	28.40
σ^2 assessee × exercise	32.80	62.29	45.24
Unreliable variance			
σ^2 assessor *	8.28		
σ^2 assessee × assessor + residual	39.06		
Estimated $G(q,1)$		.53	.61
Estimated $G(q,2)$		.69	.75

Note. N = 589. Estimated $G(q,k)$ = reliable variance subtotal/(reliable variance subtotal + unreliable variance subtotal/k). Assessors were nested in exercises. Exercises were nested in dimensions. All assessees and assessors were nested within six subsamples (i.e., three samples x two majors). For the reliable dimension variance (i.e., dimension-level scores), all variance components concerning exercises were divided by the number of exercises that assessed a given skill dimension (i.e., 2). The variance estimations attributable to other components (e.g., exercise main effects) are presented in the Online Supplement S3 (osf.io/jy5wd).
* This variance component contributed primarily to unreliable variance because assessees were not fully crossed with assessors. This contribution was rescaled (see Putka et al., 2008).

Using the results of this study, we further visually represented how assessee-dimension variance would increase as a function of using more than two exercises per skill dimension (see Figure 3). Here, reliable variance related to assessee-exercise effects was divided by the number of exercises that assessed a given skill (for more information about the impact of aggregation, see Kuncel & Sackett, 2014; Putka & Hoffman, 2013). Results suggested that relying on four or more exercises per skill dimension would lead to assessee-dimension effects surpassing assessee-exercise effects as the main source of reliable variance in assessors' ratings. As a final visualization, we compared the amount of reliable assessee-dimension variance across different AC studies that relied on generalizability theory (see Figure 4).

Figure 3

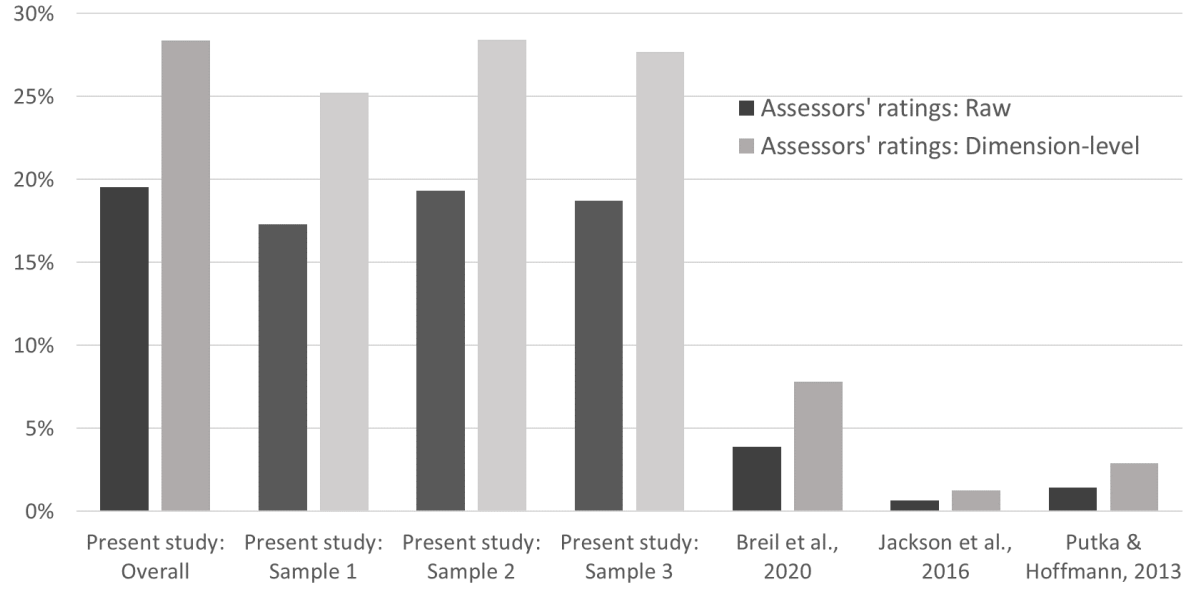
Amount of Reliable Variance as a Function of Aggregating Across Exercises



Note. Total reliable variance = σ^2 assessee + σ^2 assessee × skill dimensions + σ^2 assessee × exercises / number of exercises that assess a given skill dimension.

Figure 4

Amount of Reliable Assessee-Dimension Variance: Comparison of Studies



Note. Dimension-level variance components were rescaled on the basis of the number of exercises that assessed a given skill dimension (two for the present studies, three for the other studies).

Discussion

Aiming at a differentiated behavioral assessment of social skills, we derived, implemented, and evaluated several changes in current assessment practices. This included (a) selecting skills on the basis of a bottom-up analysis of observable and distinguishable interpersonal behaviors (in addition to a job analysis) and (b) assessing these via exercises that were designed to exclusively measure the different skills (i.e., multiple exercises nested within skills). Here, we aimed to conduct a first empirical investigation of this adapted procedure and tested it in a high-stakes selection context via multiple speed assessments (i.e., short interpersonal simulations).

Assessment of Distinct Social Skills

Our results showed that implementing the presented changes improved the assessment of distinct social skills. This was evident from a substantial amount of variance reflecting individual differences in performance that depended on the specific skill dimensions. In fact, this crucial variance was considerably higher than in previous studies that followed a more classical AC design (e.g., Breil et al., 2020; Jackson et al., 2016; Putka & Hoffman, 2013; see Figure 4). That is, assessees showed variability not only in their general performance (some assessees were better than others across all exercises) or in their exercise-related performance (some assessees performed better on specific exercises) but also in their skill-related performance (e.g., some assessees performed better on exercises related to assertiveness, whereas other assessees performed better on exercises related to warmth).

The finding that ratings across different social skills were generally positively related (i.e., assessee main effects) was not unexpected, as some assessee characteristics such as cognitive ability (Collins et al., 2003), or being able to assess situational demands (Jansen et al., 2013) are likely related to effective performance irrespective of the specific social skill dimensions. Of course, assessees' performance ratings were not perfectly consistent even

across exercises that belonged to the same skill dimension. This is also not surprising as the exercises were not designed to be completely parallel (and, thus, also more redundant) measures (i.e., they included different tasks, role-players, and affordances). Still, compared with studies on behavioral consistency in personality science (e.g., Leikas et al., 2012; Sherman et al., 2010), the assessor rating consistencies were a bit lower. Thus, the skill ratings might also have been influenced by exercise-specific aspects not directly related to the desired behaviors. For example, the specific arguments put forth in the "mistake" exercises may have influenced the skill ratings in addition to the actually expressed differences in resilient behaviors. However, different performances within exercises belonging to the same broad skills should not generally be treated as meaningless or unreliable as they likely also reflect context-dependent variability in behavioral expressions. That is, some individuals might be more resilient when it comes to time pressure, whereas other individuals might be more resilient concerning criticism. Aggregating across these different aspects of resilience arguably allows for a more comprehensive—and, thus, more valid—skill assessment compared with an approach that maximizes consistency at the expense of restricting the breadth of the skill construct (Asendorpf, 1988; Clifton, 2020; Speer et al., 2014).

Implications

Generally, the findings and suggestions do not imply that the current best practices should be abandoned when it comes to the development of assessment procedures. ACs show predictive validity, and this is likely driven by the variety of job-related exercises that evoke job-related behaviors. However, if the aim is not only to predict job performance but to assess distinguishable (social) skills (e.g., for development purposes or to give feedback on assessees' strengths and weaknesses), we put forth the "multiple exercises nested in skills speed assessment" as an alternative procedure. This procedure essentially follows mixed-

model ACs that acknowledge the importance of exercises *and* skill dimensions (see Hoffman et al., 2011; Melchers et al., 2012).

In this study, skill selection was based on a bottom-up analysis of observable and distinguishable interpersonal behavior, which serves as an important addition to top-down approaches of reducing assessment dimensions (e.g., Arthur et al., 2003; Meriac et al., 2014). This does not mean that the same skills (i.e., assertiveness, warmth, resilience) should be chosen in all assessment contexts but rather suggests that more emphasis should be placed on how the proposed behavioral structure of desired skills aligns with the empirical and theoretical structure of actually expressed behaviors. This includes identifying the specific situational triggers that evoke desired behavioral differences. Here, one can build on previous research: For example, interpersonal theory (Dawood et al., 2018) has provided many findings on situations in which differences in agentic or communal behaviors should be especially profound.

Exercise effects (i.e., individual differences in performance that depend on the specific context) will always play an important role in assessment contexts. In the presented assessment approach, exercise effects are channeled by assigning multiple exercises to specific skills. This gives meaning to the still prevalent exercise-related performance differences as they can be interpreted as context-dependent facets of the broader social skills. Accordingly, feedback given to assessees will benefit from acknowledging this nested structure (e.g., “Overall, you showed above-average performance. This was especially true for exercises related to assertiveness and warmth. On the exercises related to resilience, you varied in your performance, showing low resilience when it comes to time pressure”).

Concerning the specific type of assessment procedure, multiple speed assessments have been discussed as an ideal way to assess the broad behavioral repertoires of assessees (Herde & Lievens, 2020). Even though relatively few exercises were used in the current

implementation, the results indicate that multiple speed assessments can also be used to assess distinguishable behavioral domains. Generally, as the number of exercises assessing each skill increases, exercise-related performance (i.e., context-dependent skill facets) will have less influence on the overall skill scores, resulting in more reliable skill estimations. Thus, there is always a benefit to including more exercises in the assessment procedure, limited only by cost and time constraints. Here, we identified four exercises per skill dimensions as the sweet spot, leaving individual differences in assessees' skills as the main source of variance in ratings.

Limitations and Future Research

In this study, we focused on the assessment of social skills as expressed in a specific high-stakes selection context. Naturally, to increase generalizability, the present results should be replicated across different selection contexts. Besides this, the present study highlights several important directions for future research.

First, the situational triggers (e.g., role-players acting sad and insecure to evoke differences in communal behavior) were based on theoretical considerations, and we did not directly test whether the triggers we included actually led to an increase in the desired behavioral variance. Thus, future research might specifically change the situational triggers and investigate how this affects behavioral expressions and subsequent ratings. For example, one could directly rate relevant behavioral expressions in exercises with or without the specific triggers (e.g., rate communal behavior in the exercise revolving around warmth but also in the exercises revolving around assertiveness). Furthermore, the perception and behavioral reaction to situational triggers could be further disentangled. That is, some assessees might not even perceive the situation as demanding a specific behavioral response (see Jansen et al., 2013), whereas other assessees might theoretically know how to behave, but do not manage to react as planned. Such an investigation of how different triggers work

and affect behavioral expression will (a) advance the theoretical understanding of behavioral functioning and (b) benefit an effective process of exercise-creation.

Second, based on previous research concerning multiple speed assessments (e.g., Herde & Lievens, 2020; Oliver et al., 2016), the skills we assessed were relatively broad and rated via one global item. Here, it would certainly be possible to further zoom in on the specific skills and more directly assess different aspects (e.g., by differentiating between assertive verbal content, assertive nonverbal behavior, and assertive paraverbal behavior), which might help to explain different behavioral expressions between exercises. Furthermore, one can zoom in on the behaviors from a time perspective. That is, a continuous assessment of behavior via joystick approaches over the course of an exercise (i.e., CAID method; Sadler et al., 2009) would enable a more fine-grained perspective on the underlying behavioral processes. This would not only allow for an analysis of (within situation) behavioral variance and its relationship with performance measures but also help in identifying relevant situational triggers.

Third, we provided an in-depth analysis of the reliability and dimensionality of the different skill ratings but did not investigate predictive validity. Whereas a lot of research has suggested that AC performance ratings predict future performance, this will likely vary depending on the exercises and skills that are included (Speer et al., 2014). Thus, it would be beneficial to investigate the predictive validity of the presented procedure and focus on differences between skills. That is, one might expect different relations depending on the type of skill dimension (e.g., warmth might be more strongly related to short-term customer satisfaction, whereas assertiveness might be related to long-term success; see Wirz et al., 2020).

Conclusion

Despite the theoretical and practical relevance of a range of different social skills in work and educational contexts (e.g., warmth, assertiveness, resilience), previous research did not find evidence for their reliable distinction across assessment situations. So, can distinct social skills be measured? And how? Building on the multiple speed assessment framework, we presented an alternative behavior-focused assessment approach. By incorporating insights from behavioral personality science and aligning behavioral expression, skill selection, and exercise creation, we showed that an assessment of distinguishable social skills is indeed possible. These results demonstrate the benefits of a more behaviorally focused selection process. We encourage researchers and practitioners to further build on these findings.

Open Science

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.

Please note that all research questions and assumptions were decided upon before data collection. Furthermore, no available data (e.g., participants, samples, exercises) were excluded for this study. That is, the sample size refers to the complete number of applicants who went through the revised selection process and gave informed consent for their data to be analyzed. This study was not preregistered.

Open Data: The study data and corresponding R code (as well as supplemental material) is available on the projects' Open Science Framework page (osf.io/jy5wd). We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results.

Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. Since the specific selection procedure is still in use, we cannot provide the original materials of this procedure (e.g., assessee instructions, role-player instruction, assessor instructions). However, this material is available from the corresponding author upon reasonable request.

References

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Asendorpf, J. B. (1988). Individual response profiles in the behavioral assessment of personality. *European Journal of Personality*, *2*(2), 155–167. https://doi.org/10.1002/per.2410020209

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

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*(4), 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*(5), 1114–1124. https://doi.org/10.1037/0021-9010.91.5.1114

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

Breil, S. M., Forthmann, B., Hertel-Waszak, A., Ahrens, H., Brouwer, B., Schönefeld, E., Marschall, B., & Back, M. D. (2020). Construct validity of multiple mini interviews – Investigating the role of stations, skills, and raters using Bayesian G-theory. *Medical Teacher*, *42*(2), 164–171. https://doi.org/10.1080/0142159X.2019.1670337

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*(5), 442–443. https://doi.org/10.1002/per.2119

Breil, S. M., Lievens, F., Forthmann, B., & Back, M. D. (2021). A closer look at interpersonal behavior in assessment center role-play exercises – Investigating the behavioral structure, consistency, and effectiveness [Manuscript submitted for publication].

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of nonverbal cues to the accurate judgment of personality traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment.* Oxford University Press.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 695–700. https://doi.org/10.1076/jcen.23.5.695.1249

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259–270. https://doi.org/10.1037/met0000236

Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment*, *11*(1), 17–29. https://doi.org/10.1111/1468-2389.00223

Dawood, S., Dowgwillo, E. A., Wu, L. Z., & Pincus, A. L. (2018). Contemporary integrative interpersonal theory of personality. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: The science of personality and individual differences* (pp. 171–202). Sage.

Dorfman, R. (1938). A note on the delta-method for finding variance formulae. *The Biometric Bulletin*, *1*, 129-127.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 260–293. https://doi.org/10.1111/j.1467-6494.1983.tb00338.x

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, *38*(5), 421–447. https://doi.org/10.1016/j.jrp.2003.10.001

Herde, C. N., & Lievens, F. (2020). Multiple speed assessments: Theory, practice, and research evidence. *European Journal of Psychological Assessment*, *36*(2), 237–249. https://doi.org/10.1027/1015-5759/a000512

Hertel-Waszak, A., Brouwer, B., Schönefeld, E., Ahrens, H., Hertel, G., & Marschall, B. (2017). Medical doctors' job specification analysis: A qualitative inquiry. *GMS Journal of Medical Education*, *34*(4), Article 43. https://doi.org/10.3205/zma001120

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*(2), 221–228. https://doi.org/10.1111/jopy.12097

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64*(2), 351–395. https://doi.org/10.1111/j.1744-6570.2011.01213.x

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*(12), 1367–1378. https://doi.org/10.1037/apl0000333

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y.-J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*(7), 976–994. https://doi.org/10.1037/apl0000102

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*(2), 326–341. https://doi.org/10.1037/a0031257

Kanning, U. P. (2009). Diagnostik sozialer Kompetenzen [Diagnostics of social skills] (2nd ed.). Hogrefe.

Klein, C., DeRouin, R. E., & Salas, E. (2006). Uncovering workplace interpersonal skills: A review, framework, and research agenda. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (Vol. 21, pp. 79–126). Wiley.

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*(1), 349–372. https://doi.org/10.1146/annurev-orgpsych-012218-014955

Knorr, M., Schwibbe, A., Ehrhardt, M., Lackamp, J., Zimmermann, S., & Hampe, W. (2018). Validity evidence for the Hamburg multiple mini-interview. *BMC Medical Education*, *18*(1), Article 106. https://doi.org/10.1186/s12909-018-1208-0.

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*(1), 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*(1), 84–97. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*(1), 22–35. https://doi.org/10.1037/0021-9010.89.1.22

Leikas, S., Lönnqvist, J.-E., & Verkasalo, M. (2012). Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal situations. *Journal of Personality and Social Psychology*, *103*(6), 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*(8), 986–990. https://doi.org/10.1016/j.paid.2011.08.003

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, *87*(4), 675–686. https://doi.org/10.1037/0021-9010.87.4.675

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*(1), 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*(5), 424–440. https://doi.org/10.1002/per.2111

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*(2), 247–258. https://doi.org/10.1037/0021-9010.91.2.247

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* (pp. 245–286). Wiley.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, *97*(2), 460–468. https://doi.org/10.1037/a0025741

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). Emerald Group Publishing Limited.

Melchers, K., Wirz, A., & Kleinmann, M. (2012). Dimensions and exercises: Theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 237–254). Routledge.

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, *40*(5), 1269–1296. https://doi.org/10.1177/0149206314522299

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Oliver, T., Hausdorf, P. A., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises. *Journal of Management*, *42*(7), 1992–2017. https://doi.org/10.1177/0149206314525207

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*(1), 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*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., Hirose, S., Kleinmann, M., Kudisch, J. D., Lanik, M., Jackson, D. J. R., Kim, M., Lievens, F., Meiring, D., Melchers, K. G., Pendit, V. G., Putka, D. J., Povah, N., Reynolds, D., . . . Thornton, G. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435–1447. https://doi.org/10.1037/apl0000236

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*(6), 1005–1020. https://doi.org/10.1037/a0016232

Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology*, *99*(2), 330–343. https://doi.org/10.1037/a0019796

Soto, C. J., Napolitano, C. M., & Roberts, B. W. (in press). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. *Current Directions in Psychological Science*.

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology*, *99*(2), 282–295. https://doi.org/10.1037/a0035213

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*(3), 395–412. https://doi.org/10.1037/0022-3514.37.3.395

Wirz, A., Melchers, K. G., Kleinmann, M., Lievens, F., Annen, H., Blum, U., & Ingold, P. V. (2020). Do overall dimension ratings from assessment centers show external construct-related validity? *European Journal of Work and Organizational Psychology*, *29*(3), 405–420. https://doi.org/10.1080/1359432X.2020.1714593