Project working title: A Deeper Look into Construct Validity of Assessment Centers - Which Behaviors Do Actually Matter?

Authors:

Affiliation:

Date: ***Originally uploaded on March 08, 2019 (osf.io/rj4z6; so far not public). This is an identical copy with masked authors and masked affiliations. The original preregistration will be made public after the review process.***

A. Background, Aim, and Hypotheses

What is and what can be measured within Assessment Centers (ACs)? This question has engaged selection researchers for the last decades but left them with confusion and disagreement (e.g., Jackson, Michaelides, Dewberry, & Kim, 2016). To move beyond the general exercise vs. dimension discussion we took a deeper look into (interpersonal) behavior shown in ACs. Specifically, we aimed at analyzing 1) the structure of interpersonal behaviors shown in ACs and 2) their impact on subsequent AC-judgments.

Doing so, we will use real-life AC data and videos from the selection of medical students. Over the course of this AC (which already happened), applicants faced several short stations (i.e., 5 minutes) which included interactions with professional role players. Applicants were subsequently evaluated by trained professionals (physicians) on three different dimensions (overall competence/intuitive judgement, relationship building, information handling / emotional management). For our analyses we will use data of three different AC exercises which vary in setting (take care of a stranger after an accident, persuade a patient, deliver bad news). As all the interactions were videotaped, interpersonal behaviors were coded by trained independent raters after the selection procedure took place.

We already used the (behavioral) data from one station (i.e., take care of a stranger after an accident) for multiple bachelor and master theses. This research builds on a master thesis (see *blinded* for the respective preregistration) that investigated whether behaviors belonging to the dimensions of friendliness, dominance, expressiveness, and arrogance can be reduced to the constructs of agency and communion. This prediction was confirmed (using a confirmatory factor analysis) for this one station. We have since expanded the model by (coded) behaviors that should generally be associated with nervousness as well as intellectual competence resulting in a (good fitting) model with 31 behaviors being assigned to the 4 factors of agency, communion, nervousness, and intellectual competence. These factors were all related to performance judgments within the station. When controlling for the respective other factors we found independent effects of communion and intellectual competence.

We now aim to test whether this model can be transferred to the two other stations used as part of this AC. We hereby investigate the following hypotheses:

1) We assume that 31 specific behaviors can be reduced to the four factors of agency, communion, nervousness, and intellectual Competence.
2) We expect (uncontrolled/zero-order) effects for all these four factors on AC judgments within the respective stations. We will also investigate the multiple influence of all four factors on AC judgments. Furthermore, we will test how generic (all AC judgments within one station as one latent factor) or dimension specific (specific factors for each judgment dimension) these relationships are. In additional analyses, we will control and examine the influence of other variables such as (but not limited to) attractiveness, gender, type of study (human medicine / vs. dentistry), or GPA.

In case that the factor structure does not replicate across stations, we will exploratorily investigate which structure and / or model configuration lead to satisfactory model fits and then explore how the resulting factors are related to AC judgments.

B. Method

Participants

Targets. Targets were 220 women and men that applied for medical school beginning in April 2017. Out of those 220 candidates, 160 applied for human medicine and 60 for dentistry. 203 applicants (143 female) allowed the use of their data to be analyzed for scientific purpose. Age ranged from 18 to 29 years (M = 19.56; SD = 1.67). The pre-selection was based solely on GPA. Specifically, in this sample, a GPA of 1.3 for human medicine and 1.7 for dentistry was necessary to be invited to the selection procedures (whereas 1.0 is the best possible high-school grade and 4.0 the worst achievable grade). 110 candidates of the 220 were eventually accepted to attend medical school.

Observers. Observers for the selected exercises were 36 professional physicians (eight female; age: 27 – 67, M = 48.79, SD = 10.20; 20 years average professional experience) who attended observation training prior to the assessment center. Different observers were assigned to different stations and different applicants. Teams of two observers evaluated 40 candidates (30 candidates for dentistry) each.

Procedure

The selection procedure took part on different days for human medicine and dentistry applicants, but the procedures were mostly identical. First, perceivers were given observation training. Then, the assessment center took place. The AC was developed similarly to approaches at other universities (cf. Eva, Rosenfeld, Reiter, & Norman, 2004; Knorr & Hissbach, 2014; Rees et al., 2016). Over the course of the AC, participants had to face ten different exercises (i.e., stations) that lasted 5 minutes each and included different tasks. They were observed by physicians behind a one-way mirror. For this research, we will focus on three stations that included interaction with professional role-players and were videotaped. These stations are: Taking care of a stranger after an accident, persuade a patient, and deliver bad news. After each 5-minute situation, participants switched to the next station and observers were asked to judge participants on three dimensions.

Measures

As the selection process already took place, we already obtained the videos as well as the observer judgment ratings. Furthermore, the behavioral codings have been finished for two stations (take care of a stranger after an accident, persuade a patient) and are on the way for the third station (deliver bad news). However, we have so far only investigated the data of one station (i.e., accident).

Observer judgement dimensions. Depending on the stations, judges had to rate applicants on different dimensions. For the selected stations this included *relationship building* (i.e., build and preserve a good relationship with the interaction partner), *handling of information* (i.e., gather and pass on necessary information), as well as an overall intuitive judgment. For the station *persuade a patient*, the dimension *handling of information* was changed to *emotional management* which still included passing on relevant information but with a focus on understanding the interaction partner's perspective. The anchor specifications were based on aspects of the individual situation (e.g., accident station, handling of information: applicant inquires all necessary information regarding the accident; the ambulance call involves all necessary information, …) and simultaneously constructed to complement the anchor descriptions in the other stations. The overall intuitive judgment dimension was identical in all situations and enquired whether one could imagine this applicant in her / his practical year ranging from 0 (not very imaginable) to 5 (very imaginable).

Behavioral codings. For the behavioral coding, independent coders counted and rated 37 items that were allocated to six interpersonal behavioral domains. These domains were derived from the interpersonal circumplex (i.e., expressive, dominant, arrogant, warm; Wiggins, 1979) and supplemented by nervous (e.g., Leising & Bleidorn, 2011) and intelligent/competent (e.g., Borkenau & Liebler, 1995) behaviors. Potentially suitable behaviors were taken from available micro behaviors in the M-BeCoSy (Grünberg, Mattern, Geukes, Küfner, & Back, 2018), the CONNECT study (Geukes et al., 2019; osf.io/2pmcr/) and other literature (e.g., Borkenau & Liebler, 1995; Gifford, 1994). We first selected (or altered) potentially suitable behaviors regarding the performance context. We then did a bottom up analysis of example videos and selected five to seven behaviors for each behavioral domain that 1) were observable in the videos, and 2) varied between applicants. For a detailed overview of all assessed behaviors see Table 1. Behaviors were either counted, e.g. "clear statements that indicate a certain direction regarding content" or rated. Ratings such as "shows self-confident/dominant gestures" were made using a scale from 1 (very little) to 6 (very much). 18 teams of two coders (one team for every domain in every station) received extensive trainings and coded the behavior of all applicants within the respective domain and station. A few items were excluded from further analyses based on low ICCs and / or low intercorrelations.

Attractiveness was rated by 40 independent raters (each rater judged 101 or 102 targets within the accident station). Ratings were based on the first 15 seconds of interaction within the accident station. Attractiveness was operationalized through the three items: attractiveness of body, attractiveness of face, and neatness/trimness of hair and face.

Table 1.

Behavioral Domain	**Behavior**	**Counted or Rated**	**Transformation**	**Par-cel**
Dominance	interrupts others to steer conversation in another direction/to finish others sentences	Counted	Excluded	
Dominance	clear statements that indicate a certain direction regarding content	Counted	Box-Cox trans. then standardized	2
Dominance	upright, dominant posture of body	Rated	Standardized	1
Dominance	dominantly leans forward or turns towards other person	Rated	Standardized	1
Dominance	shows self-confident/dominant gestures	Rated	Standardized	2
Dominance	addresses the other person immediately and leads the interaction	Rated	Standardized	1
Dominance	stable, confident flow of words	Rated	Standardized	2
Friendliness	agrees, makes responsive sounds while the patient talk	Counted	Box-Cox trans. then standardized	1
Friendliness	expresses politeness	Counted	Excluded	
Friendliness	confirmative/friendly nodding and smiling	Rated	Standardized	2
Friendliness	active listening (positive paraphrasing, behaves attentive, listens to the other person)	Rated	Standardized	2
Friendliness	offering to help, statements of support	Rated	Standardized	1
Friendliness	turns to others in an attentive manner, shows positive-trusting attention	Rated	Standardized	1

Expressiveness	makes - appropriate to the situation - easing or humorous statements	Counted	Excluded	
Expressiveness	talks a lot	Rated	Standardized	1
Expressiveness	expressive lively facial expressions	Rated	Standardized	1
Expressiveness	dynamic, (not nervous!) movements of hands, arms and the body	Rated	Standardized	2
Expressiveness	expresses a positive basic attitude and optimism (i.e., covered not to the other person but covered to itself)	Rated	Standardized	2
Arrogance	interrupts the person, cuts others off	Counted	Box-Cox trans. then standardized	1
Arrogance	arrogant-patronizing comments	Counted	Box-Cox transformation then standardized	2
Arrogance	not nervous, bored gestures	Counted	Excluded	
Arrogance	behaves in an arrogantly detached/bored manner is aloof/ not to be impressed	Rated	Standardized	1
Arrogance	Paternalism / Ignore wishes of the other person	Rated	Standardized	2
Arrogance	takes an rejecting posture (crossing arms, turning away)	Rated	Standardized	1
Arrogance	aggressive-challenging, arrogant- gestures and facial expressions	Rated	Standardized	2
Nervousness	breaking up sentences, getting muddled, stammering, repeating sentences or words	Counted	Behaviors were aggregated then Box-Cox trans then standardized	1
Nervousness	uses expletives (ehm, mhm,...)			
Nervousness	reinsurances	Counted	Excluded	
Nervousness	nervous and/or purposeless change of position	Rated	Standardized	2
Nervousness	frequent change arm position, hand position; self-touch	Rated	Standardized	1
Nervousness	nervous facial expression	Rated	Standardized	1
Nervousness	seems rigid, does not act a lot/freezes	Rated	Standardized	2
Intelligence/ Competence	explains own arguments and positions, uses words of reasoning (e.g. "because", "since"...)	Counted	Box-Cox transformation then standardized	1
Intelligence/ Competence	fluent, clear way of speaking; is eloquent and articulate	Rated	Standardized	1
Intelligence/ Competence	fast, well-fitting answers to questions, reactions to comments etc.	Rated	Standardized	2
Intelligence/ Competence	behaves task- and goal-orientated; asks reasonable questions	Rated	Standardized	1
Intelligence/ Competence	puts perspectives, arguments, solutions next to each other and compares	Rated	Standardized	2

C. Analysis plan

Here, we describe how we preprocessed and analyzed the data in the accident station. For the two other stations that we now aim to investigate, we will follow the exact same steps.

Data pre-processing

Observer Judgement dimensions. The observer judgments were aggregated across the two observers.

Behavioral codings. We aggregated all behavioral items between the two coders and computed ICCs (3,k) as well as intercorrelations. Based on these statistics, we decided to exclude five behaviors that showed low ICCs and low intercorrelations with other behaviors from the respective domains (see Table 1). Furthermore, the two counting items in the nervousness domain were aggregated to one score. This resulted in a final sample of 31 unique behaviors. As the counting items were heavily right skewed (and included extreme values), we used Box-Cox transformation on these items. In addition, all behaviors were standardized.

In a next step, we created two parcels for each behavioral domain. We generally used the balance approach (i.e., allocated items based on their factor loading on the respective behavioral domain; in the order 1 2 2 1 1 2 2 1., e.g., Little, Cunningham, Shahar, & Widaman, 2002). We however differed from this approach for expressiveness and arrogance (the two domains that should load on two factors, see below). Here, we created parcels that showed equal loadings on both factors. Furthermore, for the nervousness domain we created a parcel that included both nervous change of position as well as rigid/freezing behavior. This was done because theoretically one would assume that nervousness is either expressed in a frequent change of position or in rigid/freezing behavior and not in both behaviors at the same time. Thus, we created a parcel that included both behaviors so that a low score would then represent individuals who show neither nervous change of positions nor freezing, while a high score would represent individuals with either nervous change of position or rigid/freezing behaviors. Please see Table 1 for the exact allocations.

Analysis of research questions

Hypothesis 1. Using the R program lavaan (Rossell, 2012), we build a model with the 12 parcels loading on the four latent factors agency (dominance, expressiveness, arrogance parcels), communion (friendliness, expressiveness, arrogance parcels), nervousness (nervousness parcels), and intelligence (intelligence parcels). We expected cross loadings for expressiveness (i.e., positive loading on agency and communion) and arrogance (i.e., positive loading on agency and negative loading on communion). This is because theoretically expressiveness (i.e., gregarious, extraverted) and arrogance (i.e., calculating) lie between the poles of agency (i.e., status) and love (i.e., communion) and should thus be related to both factors. Theoretically, the behaviors that lie between the poles should load less strongly on the respective factors compared to the behaviors that lie on one of the poles (i.e., expressiveness and arrogance should load less strongly on agency compared to dominance; expressiveness and arrogance should load less strongly on communion compared to friendliness).

As the parcels for expressiveness and arrogance loaded on two factors, we allowed correlations among residuals for the parcels expressiveness 1 and 2 as well as for the parcels arrogance 1 and 2. We hereby accounted for the fact that the respective parcels were rated by the same team of judges who attended the same judgment training and discussed the same example videos. This might have led to different levels in rating characteristics such as leniency between the rating teams (i.e., shared method variance) which are not captured with the cross loadings. Variances of the latent factors were fixed

to 1. No further restrictions were made. Parameter estimated were based on maximum likelihood estimation with robust standard errors (MLM).

We will test this model (with the exact same specifications) in the other two stations and evaluate its performance based on common fit indices. A good fit would be indicated by CFI > .95, RMSEA < .06; SRMR < .08 (cf. Hu & Bentler, 1999).

Hypothesis 2. We created a latent judgment factor with the three observer judgment variables. In a first step, we included this factor to the previous described model and investigated (uncontrolled) relationships between all factors. Here, we tested the relationships between the latent judgment factor and each of the four behavioral factors. In a next step, we used the four behavioral factors to predict the judgment factor and investigated their respective influence. We repeated this procedure with three observer judgments as separate factors (single indicator approach). In additional models, we controlled for gender and / or attractiveness.

We will test all these relationships for significance ($p < .05$) in the other two stations.

D. Literature

Borkenau, P., & Liebler, A. (1995). Observable attributes as manifestations and cues of personality and intelligence. *Journal of Personality, 63*(1), 1-25.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. *Medical Education, 38*(3), 314-326.

Geukes, K., Breil, S. M., Hutteman, R., Nestler, S., Küfner, A.C.P., Back, M.D. (2019). Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The PILS and the CONNECT study. *PlosOne, 14(1)*, e0210424

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology, 66*(2), 398-412.

Grünberg, M., Mattern, J., Geukes, K., Küfner, A., & Back, M. (2018). Assessing group interactions in personality psychology: The Münster Behavior Coding-System (M-BeCoSy). In E. Brauner, M. Boos, & M. Kolbe (Eds.), *Cambridge Handbook of Group Interaction Analysis* (pp. 602–611). Cambridge: Cambridge University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Jackson, D. J., 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.

Leising, D., & Bleidorn, W. (2011). Which are the basic meaning dimensions of observable interpersonal behavior?. *Personality and Individual Differences, 51*(8), 986-990.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151-173.

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. *Medical Education, 48*(12), 1157-1175.

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*(5), 443-455.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*(3), 395-412.