Do face-to-face panel interviews in medical school admission help us select empathetic students? Results of a cross-sectional study

Josefin Wagner, Alex Mommert, Jürgen Westermann

Empathy is a key attribute of competent physicians. Yet, research on empathy as medical school admission criterion in Germany is still scarce. Lübeck Medical School admits 60 percent of students via standardised, structured face-to-face panel interviews designed to measure non-academic aspects of aptitude for medical school. We hypothesise that we can detect a difference in empathy levels between interview-admitted and non-admitted students. In our sample (N = 420), admitted compared to non-admitted applicants show significantly higher empathy scores on an established standardised self-assessment tool. The effect persists when relevant covariates (age, sex and pre-university grade point average) are included in the analysis. Of the covariates, only sex had a significant influence. Despite limited generalisability, our study indicates that interviews do help select empathetic medical students and contributes to the field of medical school admission research with implications for practice.

1 Introduction

Medical school admission in Germany is highly competitive as the number of applications is some four times larger than study place capacities (SFH, 2019). Given this scarcity on the one hand, and the societal demand for medical doctors on the other, selection becomes inevitable in medical school admission and should focus on an applicant's aptitude for a successful course of study and the medical profession.

The necessity to emphasise interpersonal skills development throughout medical school and residency makes an assessment of candidates' non-cognitive attributes a feasible approach in order to select suitable candidates with the highest potential to become competent physicians (Hojat, 2014). Among desirable non-cognitive traits and qualities, empathy has recently become a focus of interest in teaching, training and admission. Empathy facilitates informative patient-physician communication (Coulehan et al., 2001; Neumann et al., 2012) and can ultimately lead to better physical and psychosocial health outcomes for patients (Elliott et al., 2018; Hojat et al., 2011; Neumann et al., 2012) as well as higher patient satisfaction (Menendez et al., 2015; Wang et al., 2018; Zachariae et al., 2003).

With regard to student selection for medical school admission, we expect candidates who show a high level of empathy at admission to have an advantage in their further empathy development during medical teaching and training. Hojat (2014) argues in favour of selecting applicants with an existing empathic orientation over those who need additional intensive training from a cost-effectiveness perspective. Some studies show a decline in empathy scores during medical training (for details see section 2). Yet, the decline was less steep in those with initially higher empathy scores (Chen et al., 2012). However, empathy has been studied only sparingly in the context of the German medical school admission system and results are insufficient to confidently introduce measures of empathy (Knorr et al., 2018; Obst et al., 2016).

We undertook our investigation at Lübeck Medical School (LMS), where since 2012 approximately 60% of medical students have been admitted via a school-specific selection procedure (*Auswahlverfahren der Hochschulen*) that, at its core, relies on face-to-face panel interviews. The previous interest had been whether personality and psychometric inventories could be utilised for our selection procedure (Kötter et al., 2017; Obst et al., 2016). Naturally, we are also interested in the effectiveness of the panel interview itself to detect empathetic students. Assuming that panel interviews are a capable instrument to ascertain a candidate's aptitude, we hypothesise that we can detect a difference in empathy levels between interview-admitted and non-admitted students. As the admission decision is a result of a candidate's overall interview score, it is conceivable that suitable candidates with high overall interview scores possess other desirable traits and qualities – such as empathy – detectable in the panel interview. Alternatively, certain domains that are explicitly addressed in the panel interview may be interrelated with empathy.

In this paper, we first outline the importance of empathy in the clinical context, followed by an overview of the German admission system and the specificities of the LMS selection procedure in sections 2 and 3. Section 4 describes the study design, methods and analyses. Results are presented and discussed in sections 5 and 6. The final section summarises implications for research and practice.

2 Empathy in the clinical context

Empathy is one of the key attributes of competent physicians. In a clinical context, empathy involves understanding a patient's situation, perspective and feelings, communicating this understanding and acting accordingly in a therapeutically helpful way (Mercer & Reynolds, 2002). Being able to understand and reflect another person's thoughts and emotions facilitates fruitful physician-patient communication and improved patient outcomes (Banja, 2006; Kerasidou & Horn, 2016; Zachariae et al., 2003). Consequently, empathy has seen an increased relevance in medical training

and a variety of measures strengthening students' and residents' empathy has been introduced in recent years. International examples of medical education frameworks, such as the CanMeds Framework (Frank, 2005) and the Accreditation Council for Graduate Medical Education Outcome Project (Swing, 2007), emphasise on the relevance of interpersonal and communication skills. The *Master Plan for Medical Studies 2020 (Masterplan Medizinstudium 2020)* is a similar transformation currently underway in Germany (BMBF, 2017).

Nonetheless, conceptualising empathy is not easy as definitions and aspects of dimensionality vary. In general, concepts largely concur that empathy involves a cognitive component (Davis, 1980; Dohrenwend, 2018; Hojat et al., 2001; Mercer & Reynolds, 2002), while some authors additionally emphasise the emotional facet (Davis, 1980) or moral and behavioural aspects (Mercer & Reynolds, 2002). Hemmerdinger et al. define empathy as "*a personality trait that enables one to identify with another's situation, thoughts, or condition by placing oneself in their situation*" (2007, p. 2). Universally and regardless of the age group studied or measure utilised, empathy in females has been shown to exceed levels in males (Chen et al., 2012; Davis, 1980; Davis & Franzoi, 1991; Hojat et al., 2011).

Particularly in studies concerning empathy in medical students and medical education, self-reported measures are commonplace, among which the Interpersonal Reactivity Index (IRI) (Davis, 1983) and the Jefferson Scale of Empathy (JSE) (Hojat et al., 2001) are the most widely used. The JSE was developed to assess empathy specifically in healthcare providers and views empathy as an interpersonal skill relevant to the concept of clinical competence (Hojat et al., 2001). In contrast, the IRI was developed in a general population setting (Davis, 1983). Supporting the underlying assumption that empathy is a stable construct, Davis and Franzoi (1991) have shown maturity in trait empathy development that renders IRI scores stable in high-school students. Hojat et al. (2005) found a moderate statistically significant correlation in overall IRI and JSE scores ($r = .45$, $p < 0.01$). This empirical overlap supports that the scales measure different but related constructs. Ultimately, the research question and the target population should inform the choice of instrument (Hemmerdinger et al., 2007; Pedersen, 2009).

Results on the trajectories of medical students' empathy throughout the course of study are mixed. The most recent meta-analysis shows inconclusive results on empathy increase/decline with little effect on actual bedside behaviour (Ferreira-Valente et al., 2017). Particularly longitudinal studies show a decline of empathy scores on the JSE throughout the course of study with a significant drop upon entering the clinical phase (Ferreira-Valente et al., 2017; Hojat et al., 2009; Papageorgiou et al., 2019). A deterioration was observed in a systematic review that also included cross-sectional

studies (Neumann et al., 2011). Commonly, reasons for empathy decline are related to distress stemming from the "hidden/informal curriculum" (e. g. high work-load, time pressure, mistreatment by superiors, a shift away from humanistic aspects of care towards technology and objectivity, lacking role models) or negative learning experiences (West & Shanafelt, 2007). Paradoxically, more frequent patient contacts in the beginning of the clinical phase and the related confrontation with clinical reality have been found to diminish empathy (Hojat et al., 2009; Neumann et al., 2011). A qualitative study elicited additional reasons which negatively affect medical students' empathy such as individual patient characteristics, professional development and general doubts about the health care system (Pohontsch et al., 2018).

3 Admission to medical school in Germany: A national framework and local adaptations

3.1 Transition from predominantly cognitively-oriented to aptitude-related admission criteria

Until recently, negligible attention had been paid to non-cognitive attributes such as empathy in the selection of medical students in the German admissions system. On the contrary, an emphasis was put on cognitively-oriented measures, particularly on applicants' secondary school leaving grade (*Abitur*), hereafter referred to as pre-university grade point average (pu-GPA). As the average of two years of high school performance across a broad canon of subjects, pu-GPA is universally considered indicative of general cognitive functions and has demonstrated acceptable predictive validity for academic success in the pre-clinical phase of medical school (Kadmon et al., 2014; Patterson et al., 2016; Trapmann et al., 2007). The predominance of the pu-GPA becomes evident in that 20% of study places were allocated solely based on pu-GPA. It was also a secondary criterion in the 20% quota for applicants with accrued waiting time since high school graduation. For another 60% of study places, medical schools were at liberty to employ individual selection schemes, including criteria such as aptitude or knowledge tests, interviews, previous health-related vocational training and employment, or grades of individual subjects (e.g. biology, physics). Yet, pu-GPA had to be a deciding factor in this quota and, for some individual medical schools, was lawfully used as the sole criterion (for an overview see Schwibbe et al., 2018). Besides pu-GPA, aptitude or knowledge tests, namely the German Test for Medical Studies (TMS) and the Hamburg Natural Science Test (Ham-Nat) have been broadly applied in the admission process across medical schools in Germany (for an overview see Schwibbe et al., 2018). Both tests have demonstrated predictive validity regarding academic success during the pre-clinical stage (Kadmon & Kadmon, 2016; Werwick et al., 2018). One major disadvantage, however, is their limited value in predicting students' academic performance in the clinical phase and clinically relevant practical skills (Saguil et al., 2015; Sladek et al., 2016).

Thanks to a ruling by the German Constitutional Court in 2017, which declared parts of the German admissions system for medical studies unconstitutional, the system has undergone major changes, including the reshaping of the admission quotas and the reconfiguration of admission criteria (Bundesverfassungsgericht, 2017). Overall, a shift can be observed towards criteria that are better predictors of an applicant's aptitude during preclinical and clinical phases as well as the actual medical profession. While the pu-GPA quota is augmented to 30%, the selection quota now requires at least two additional criteria alongside pu-GPA. The former waiting time is fully replaced by a 10% quota for candidates selected without any grade-based criteria (Staatsvertrag, 2019). This overhaul opens up a window of opportunity for the introduction of more diverse selection methods aimed at non-cognitive traits, such as interview formats of varying degrees of standardisation and structure.

Interviews in general are a nationally and internationally widely accepted selection instrument that aims to measure non-academic competencies crucial in competent physicians (Patterson et al., 2016; Schwibbe et al., 2018). Typical formats in student selection are panel interviews and multiple mini-interviews (MMI). While in panel interviews multiple raters simultaneously assess aspects of a candidate's strengths, weaknesses and overall aptitude, MMIs follow a multiple sample approach in which a number of raters evaluate candidates in various interview stations (Eva et al., 2004). Patterson et al. (2016) point out consistent findings of predictive validity evidence for MMIs. Likewise, in panel interviews, higher degrees of structure and competence-based assessment are associated with improved reliability and prognostic validity (Albanese et al., 2003; Huffcutt et al., 2013; Levashina et al., 2014; Patterson et al., 2016). Particularly for panel interviews, countermeasures addressing potential weaknesses such as social desirability effects or rater bias are well established (Cook, 2016).

3.2 University-specific selection at LMS via panel interviews

The selection procedure focuses on capturing non-academic competencies and communication skills and balancing those with indicators of academic performance in school (Brüheim et al., 2012). The initial goal was to identify applicants whose aptitudes, inclinations and interests align best with our school's profile and values while succeeding academically with regard to study progress and exam success. Selected students performed well (Mommert et al., 2020) and many of them anecdotally attributed being highly motivated to the positive experience of having been selected.

In a pre-interview stage, a bonus system awards deductions off the pu-GPA for criteria such as TMS scores, completed health-related vocational training and extracurricular activities. Invitations to face-to-face panel interviews are extended to the 240 highest-ranking applicants. Interviewees are randomly assigned to one of twelve

three-person mixed-gender panels consisting of two faculty members and one student. All panel members receive instructions to utilise biographical and situational question techniques, as those are less prone to influences from social desirability and impression management (Barrick et al., 2009; Levashina et al., 2014). Particularly, new panel members participate in an interview training that provides theoretical background on sources of bias and rating errors and familiarises them with interview techniques and the instruments used. A fully standardised, competence-based score sheet has been specifically developed at LMS. The score sheet encompasses five primary domains, namely motivation, knowledge about the course of study, social engagement, (self-) reflection and communication. In addition, an interview guide gives a brief description of desirable manifestations of each primary domain and example questions for each score sheet item. While the 30-minute-interviews are semi-structured to give panel members the opportunity to probe interviewees' statements, their reasoning and rationales, the assessment is fully structured. Interviewers individually rate each of the five items per domain on a five-point verbally and numerically anchored rating scale ranging from 0 (not at all) to 4 (entirely) resulting in a maximum of 100. Inter-rater reliability is routinely assessed as part of the quality assurance system using intraclass correlation coefficients (ICC). For the two cohorts, overall inter-rater reliability was excellent with values of ICC $(1,3)_{2018}$ = .92 and ICC $(1,3)_{2019}$ = .90. The post-interview score sums interview and pu-GPA scores. Interview scores are calculated as the mean of the independent panel members' ratings, then adjusted for the overall strictness or leniency of the panel and finally fit to a 0 to 30 range via linear transformation. Pu-GPA is transformed to a score on a 31-point-scale with a deduction of one point for every tenth off the maximum pu-GPA (range: 1.0 (best) to 4.0 (worst)). About half of the interviewees with the highest-ranking post-interview scores are admitted.

4   Study design

Our study investigates whether there are differences in empathy levels in admitted and non-admitted interviewees of the 2018 and 2019 LMS selection procedure (n_{2018} = 226; n_{2019} = 228). All were asked to participate in a post-interview evaluation and quality assurance survey. Additionally, we asked them to anonymously and voluntarily fill out the questionnaire for this study. Participating interviewees were given written information detailing the purpose of data collection and data handling. The information also contained an explicit statement that participation in the evaluation or the study would not influence selection. A numerical alias allowed linkage of the study questionnaire (dependent measure) and procedural data such as sex, age, pu-GPA, interview score and admission status. To ensure privacy, an in-house data custodian conducted this linkage. The local ethics committee approved this procedure in both years (references 18-213 and 19-257).

4.1 Dependent measure: empathy

As we were interested in empathy as a personality trait rather than a goal of professional development, for the purpose of this study, we adopt the conceptualisation of empathy as a personality trait and therefore stable construct in accordance with Hemmerdinger et al. (2007). Hence, to test our hypothesis that differences in empathy scores between admitted and non-admitted interviewees exist, we used the German 16-item short form version of the IRI, called *Saarbrücker Persönlichkeitsfragebogen (SPF-IRI)* as a measurement of empathy (Paulus, 2009). The SPF-IRI had already been employed in the context of the LMS selection procedure (Obst et al., 2016). The reported reliability (α = .78) of the SPF-IRI is comparable to the original IRI (Davis, 1980; Paulus, 2009). It contains the original four subscales:

1. *perspective taking (PT)*: assessment of one's tendency to spontaneously adopt the perspective or point of view of others

2. *fantasy (FS)*: assessment of one's tendency to identify and empathise with fictitious characters in books or movies

3. *empathic concern (EC)*: assessment of one's tendency to experience "other-oriented" emotional reactions (e.g. compassion, concern, pity) witnessing others in distress

4. *personal distress (PD)*: assessment of one's tendency to experience "self-oriented" emotional reactions (e.g. anxiety, discomfort, uneasiness) when witnessing others in distress or in tense and close interpersonal settings

Items are answered on a five-point Likert scale ranging from 1 (never applicable) to 5 (always applicable) in reference to the item statement. To compute an overall empathy score, only the EC, FS and PT subscale scores were added as suggested by Paulus (2012).

4.2 Statistical analysis, independent measures and covariates

For descriptive analyses, means (*M*) and standard deviations (*SD*) were calculated for all continuous variables; two-tailed t-tests were used for comparisons. For dichotomous items, data were analysed using χ^2-tests (Fisher's exact test); relative frequencies (%) are presented. Scale reliability was assessed computing Cronbach's α. We used a 5%-significance level for all analyses. We first conducted analysis of variance (ANOVA) with admission status as the independent and the SPF-IRI score as the dependent measure. Second, in analysis of covariance (ANCOVA) sex, age and pu-GPA were introduced as covariates. Adjusted means (M_{adj}) and corresponding standard errors (*SE*) were obtained. Effect sizes were quantified as (partial) η^2.

5 Results

Of the 454 interviewees, 420 (92.51%; n_{2018} = 212; n_{2019} = 208) participated in our study and filled out the SPF-IRI. Participants from 2018 and 2019 did not differ regarding participation rate, sex distribution, mean age, pu-GPAs, interview and empathy scores (all ps ≥ .147).

Overall, participants were predominantly female (74.52%). The mean age was 20.49 years (SD = 2.13) and mean pu-GPA was 1.47 (SD = .22). Characteristics of admitted and non-admitted participants are displayed in Table 1. Both sex ratio and mean age were comparable in admitted and non-admitted groups (ps ≥ .50). The mean pu-GPA was slightly more favourable among admitted interviewees. Expectedly, mean interview scores differed significantly. Reliability of the SPF-IRI empathy scale was good ($α$ = .76).

Table 1: Interviewee characteristics by admission status

	N	admitted	non-admitted	X^2 (1, 420)	p
sex, % female	420	72.99	76.08	.53	.502
				t (418)	
age M (SD)	420	20.44 (2.00)	20.53 (2.25)	.37	.698
pu-GPA M (SD)	420	1.44 (.22)	1.50 (.22)	2.68	.008
interview score M (SD)	420	21.44 (4.14)	11.06 (4.77)	23.81	.000

pu-GPA: pre-university grade point average; range: 1.0 (best) to 4.0 (worst).

Regarding our main hypothesis, we found a statistically significant difference in empathy scores between admitted and non-admitted participants in the one-way ANOVA (see Table 2). Empathy scores were significantly higher in admitted compared to non-admitted participants. The difference in mean empathy scores of 1.43 translates to a small effect size ($η^2$ = .02). When age, sex and pu-GPA were included as covariates, both the significant effect and the effect size persisted in the ANCOVA. Of the three covariates, only sex turned out to be significant: female participants (M = 46.16, SD = 5.01) displayed higher empathy scores than male participants (M = 44.62, SD = 4.81).

Table 2: ANOVA and ANCOVA results on empathy by admission status (N = 420)

| | admission status | | | | |
	admitted M (SD)	non-admitted M (SD)	F (1, 418)	p	η^2
empathy	46.48 (4.90)	45.05 (5.02)	8.69	.003	.02
	admitted M_{adj} (SE)	non-admitted M_{adj} (SE)	F (1, 415)	p	η^2
empathy	46.48 (.34)	45.05 (.34)	8.67	.003	.02
sex			8.04	.005	.02
age			.63	.426	.00
pu-GPA			.21	.647	.00

empathy: computed overall empathy score (Paulus, 2012).
pu-GPA: pre-university grade point average.

6   Discussion, conclusions and implications

Face-to-face panel interviews at LMS are designed to measure specific non-academic aspects of aptitude for medical school. The intention of this study was to examine whether we can also detect a difference in empathy levels between interview-admitted and non-admitted students. For this purpose, analyses of (co)variance were performed comparing empathy scores of admitted and non-admitted interviewees. Our results demonstrate that interviewees who were admitted via face-to-face panel interviews at LMS indeed showed higher empathy scores than their non-admitted counterparts. The magnitude of the effect was small. Even when we controlled for relevant covariates, such as participants' sex, age and pu-GPA, higher empathy scores among selected individuals persisted. Among the independent variables, only being female was independently associated with higher empathy scores.

6.1   Main finding: differences in empathy between admitted and non-admitted applicants

Our results are in line with our hypothesis in that selected interviewees indeed showed higher empathy scores than non-admitted interviewees. The effect size was small, comparable to the magnitude of the difference between females and males (η^2 = .02, respectively). We are, however, confident in the robustness of our findings. We used a rather large sample consisting of interviewees from two consecutive cohorts and the effect and magnitude of the effect persisted even when covariates were included in the analysis as statistical controls. At this stage of the investigation, it is unclear how interviewees' empathy measured by the SPF-IRI translates to higher scores

awarded by the panel members in the face-to-face panel interviews that ultimately lead to admission. We will discuss two possible mechanisms in detail.

First, as the admission decision is a function of an applicant's overall aptitude as a medical student, suitable candidates may possess other desirable non-cognitive traits and qualities, even when those are not explicitly focused on. Specifically, interviewees with a higher sum score across our five primary domains may be more empathetic as well. In this case, we would assume that empathy is a rather implicitly measured cross-sectional trait. A somewhat comparable result was obtained in a study by Heintze et al. (2004) examining the feasibility of alternative selection instruments for dentistry studies. They found a relationship between candidates' self-rated social competence and a more positive interview result ($r = 0.24$, $p < 0.05$). The study did not report a correlation coefficient for interview results and the applied two IRI scales (EC and PT), yet concluded that social abilities could certainly be assessed in interview situations. In a Swedish qualitative study that aimed to generate an impression of panel members' perspectives on the admission procedure, the author elicited that, in the semi-structured interviews, panel members "*look for such attributes as realism, motivation, maturity, attitudes, endurance, drive, engagement, empathy, i.e. non-cognitive attributes*" (Röding, 2005, p. 121). What is more, results from an MMI reliability study call into question raters' capability to score candidates accurately across desired non-cognitive traits. To an extent, the interviewers may have applied a broader, unidimensional construct, such as overall suitability for medical school (Sebok et al., 2014). It is possible that some of our constructs and empathy are intertwined enough for empathy scores to implicitly influence interview scores.

Potentially, empathy may be related to aspects that we explicitly address in the face-to-face panel interviews. From internal quality assessments, we know that admitted and non-admitted interviewees most strongly differ in the social engagement dimension. Our findings could be reasonably interpreted as a reflection of some of the traits and qualities that are explicitly measured in this dimension of the interview. It seems likely that the more empathetic someone is, the more she or he will engage socially. More empathetic interviewees will more easily be able to give biographical examples of their social engagement. Thus, the social engagement dimension of the interview does indirectly address empathy. Service or co-curricular activities, just like social engagement, also enrich personal development in students (Brazeau et al., 2011; Huang & Chang, 2004). The interest and responsiveness to learn from others and to reflect on those experiences is thereby fostered, which ultimately may lead to higher affective and cognitive empathy (Huang & Chang, 2004). As such, participation in service and social engagement activities prior to entering higher education could be beneficial in the development of empathy as well.

6.2 Covariate findings: sex, age and pu-GPA

Of the three covariates, only sex had an independent effect on empathy scores; female interviewees scored higher on empathy. This result is in accordance with the extensive body of literature that invariably points towards higher levels of empathy in females (Chen et al., 2012; Davis, 1980; Davis & Franzoi, 1991; Hojat et al., 2011). Reasons suggested for these differences may be evolutionary and associated with more prosocial behaviour as well as better emotion recognition and management (Christov-Moore et al., 2014) and/or may be a function of variations in communication skills (Graf et al., 2017). However, this effect does not influence our main finding since female and male interviewees were admitted in the same proportions. Additionally, controlling for sex did not disperse the admission effect; neither did the adjusted mean empathy scores in the ANCOVA differ from means in the ANOVA, nor did the effect size ($\eta^2 = .02$, in ANOVA and ANCOVA) of the admission effect.

Conversely, age and pu-GPA had no significant effect in the ANCOVA. As for age, this finding is consistent with our assumption and most likely reflects maturity in trait empathy that renders IRI scores stable in high-school students, as shown by Davis and Franzoi (1991). Even though we did not expect an effect of pu-GPA on empathy scores, we included it as a covariate since the mean pu-GPA is slightly more favourable among those admitted, as presented in Table 1. This, however, is an artefact attributable to the Higher Education admissions law itself that mandates pu-GPAs have a significant influence in the admission decision (Hochschulzulassungsgesetz, 2016). As outlined, admission is based on the final summed score of the interview and pu-GPA scores. Interviewees who rely on multiple bonuses in pre-selection in order to gain access to an interview require relatively higher interview scores to compensate the offset from the bonuses. With regard to empathy, no influence was found in the ANCOVA.

6.3 Potential influences from faking goodness and social desirability effects

Faking goodness and portraying oneself as more favourable could pose a potential threat in high stakes selection situations. Both would endanger the procedure's validity and might entail negative implications for the selection decision itself whenever individuals attain preferential rank order positions regardless of their true aptitude. We consider the risk of faking or social desirability effects on empathy scores to be rather small in our study. For one, for all participants it was made clear in the instructions that participation in the study would not influence selection, and, therefore, we assume that participants saw little reason to falsify their responses. Second, findings by Paulus (2019) suggest that there is only little danger of impression management. A study on social desirability effects on three personality questionnaires (i.e. SPF-IRI, the

work-related stress and coping questionnaire AVEM, and the Big-Five-based personality questionnaire NEO-FFI) using a within-subjects design (Obst et al., 2016) also showed only weak evidence for faking effects. Even when study participants were prompted to fake, the overall difference in the SPF-IRI scores between the fake-good and control conditions was only on the verge of statistical significance ($p = .05$). Similarly, Ones and Viswesvaran (2011) concluded that social desirability effects are negligible in validity studies on Big-Five personality assessments. Finally, all of our participants found themselves in the same high stakes situation and thus would all be equally inclined to fake. The effect that we found using a between-subject design can, therefore, not be attributed to faking or social desirability effects.

6.4 Study limitations and further research

The findings of this study ought to be reflected in the light of its limitations and avenues for future research. One potential limitation concerns the choice of instrument to measure empathy in our study. We deliberately chose the IRI over the JSE because of the JSE's conspicuous wording. With items such as *"Willingness to imagine oneself in another person's place contributes to providing quality care"* (Hojat et al., 2001, p. 358), the construct of interest was, in our opinion, too easy to guess by the study participants posing a relevant risk of faking good. Even more importantly, we were interested in empathy as a personality trait rather than a clinical skill to be developed during medical studies, particularly with regard to the pre-medical education stage of our study participants. This said, we certainly agree, that empathic conduct in physician-patient encounters and in an interprofessional medical environment, of course, can be learned and developed and is a necessary component in teaching and training. At the same time, there is evidence of an association between empathy and burnout although causality and directionality data are inconsistent (Wilkinson et al., 2017). Thus, balancing empathy and professional distance is an equally important educational goal.

The specific nature of the LMS selection procedure poses a limitation as to the generalisability of the results to other medical schools in Germany or beyond. Overall, the effect size in the ANOVA was small, yet persisted when potentially confounding variables were included in the analysis. Our independent finding of female participants' higher level of empathy replicates a well-known effect and raises our confidence in our findings overall.

On our part, further research should focus on untangling the underlying mechanism of how interviewees' empathy translates to higher scores awarded by the panel members in the face-to-face panel interviews, which ultimately lead to better chances of admission. This includes but is not limited to further analyses of the SPF-IRI subscale scores rather than the overall score as well as correlation and regression analyses of

the primary interview domains and empathy (subscale) scores. The reliability of our interview procedure needs to be addressed as well. A revision of our primary domains or the interview score sheet may have to be considered. Moreover, we will investigate longitudinally whether empathy scores remain stable throughout medical education and whether empathy at selection predicts performance in clinically-relevant exam formats, such as Objective Standardized Clinical Examinations (OSCE), at later stages of medical training or other clinically relevant outcome parameters to further validate the selection procedure.

7   Conclusions

Our results have implications for practice, particularly as most medical schools face transformation in admissions mandated by the Ruling of the German Constitutional Court (Bundesverfassungsgericht, 2017) and in curriculum development requested by the *Master Plan for Medical Studies 2020* (BMBF, 2017*)*. These structural changes open up a window of opportunity to direct more research at a wider range of selection methods that assess non-academic competencies in order to assess applicants' aptitude to become competent physicians holistically. In that regard, this study contributes to the body of knowledge on admission procedures in general and in the German context specifically. It could be informative for medical schools facing these transformative processes.

With a shift in research focus in the past 15 years on highly standardised selection methods such as MMIs and their ability to identify favourable candidates, the investigation of panel interviews has receded. Patterson et al. (2016) have called the reliability and validity of "traditional interviews" into question; they concede, however, that the mixed findings in their systematic review most likely stem from a broad range in interview methodology. The authors advocate for the utilisation of MMIs and Situational Judgement Tests (SJT) instead. MMIs are indeed intriguing as they purport to objectify the measurement of non-cognitive traits and to reduce various forms of bias (Eva et al., 2004). In theory, for any given trait, such as communication, integrity or empathy, one or more stations may be developed which then produce accurate measures of that construct. Vice versa, one would not expect to find aspects of a certain construct when it has not been intentionally included. However, construct validity and dimensionality are oftentimes insufficient in MMIs (Patterson et al., 2016; Sebok et al., 2014). In Germany, SJTs that also purport to measure non-cognitive traits in a standardised and less costly manner compared to interviews (Patterson et al., 2016) are still in their infancy (Schwibbe et al., 2018). Compared to face-to-face panel interviews, which leave room to probing a candidate's response, merely choosing one from a set of listed responses in SJTs seems to create a low-fidelity atmosphere in a high stakes situation.

Most concerning, SJTs are potentially susceptible to social desirability effects and faking, thus jeopardising fairness and validity in admission (Peeters & Lievens, 2005).

The intention of this article was to investigate whether we could detect a difference in empathy levels between admitted and non-admitted students in the standardised and structured face-to-face panel interviews conducted at LMS. Our results suggest that using such interviews can indeed be helpful in selecting more empathetic interviewees into medical school. This selection effect persisted even when we controlled for interviewees' age, sex and pu-GPA and is unlikely to have been distorted by faking or social desirability effects. Our finding of higher empathy scores in female participants is in line with previous research. Latest findings on the LMS selection procedure are encouraging. A pilot study by Kötter et al. (2020) recently found that general practitioners rated the overall suitability to become a good doctor most favourably for students admitted via face-to-face panel interviews, while Mommert et al. (2020) compared academic outcomes at the first federal examination of various admission quotas, in which selected students displayed high temporal continuity and examination success. Taken together with the findings of this study, consistent research is emerging that suggests that standardised and structured face-to-face panel interviews are an effective selection method to discern aspects of an applicant's aptitude for medical studies and the medical profession after all – contrary to zeitgeist beliefs.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing Personal Qualities in Medical School Admissions. *Academic Medicine*, *78*(3), 313–321.

Banja, J. D. (2006). Empathy in the physician's pain practice: Benefits, barriers, and recommendations. *Pain Medicine*, *7*(3), 265–275.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394–1411.

Brazeau, C. M. L. R., Schroeder, R., Rovi, S., & Boyd, L. (2011). Relationship Between Medical Student Service and Empathy. *Academic Medicine*, *86*(10), 42–45.

Brüheim, L., Sievers, K., & Westermann, J. (2012). Nicht allein die Abiturnote. *Forschung und Lehre*, *19*(11), 912–913.

Bundesministerium für Bildung und Forschung [BMBF]. (2017). *Beschlusstext „Masterplan Medizinstudium 2020"*. Retrieved on 5 September 2020 from https://www.bmbf.de/files/2017-03-31_Masterplan%20Beschlusstext.pdf.

Bundesverfassungsgericht. (2017). *Judgment of the First Senate of 19 December 2017 - 1 BvL 3/14 - paras. (1–253)*. Retrieved on 5 September 2020 from http://www.bverfg.de/e/ls20171219_1bvl000314en.html.

Chen, D. C. R., Kirshenbaum, D. S., Yan, J., Kirshenbaum, E., & Aseltine, R. H. (2012). Characterizing changes in student empathy throughout medical school. *Medical Teacher*, *34*(4), 305–311.

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience and Biobehavioral Reviews*, *46*(4), 604–627.

Cook, M. (2016). *Personnel selection: Adding value through people – a changing picture* (6th ed.). Chichester: Wiley-Blackwell.

Coulehan, J. L., Platt, F. W., Egener, B., Frankel, R., Lin, C.-T., Lown, B., & Salazar, W. H. (2001). "Let Me See If I Have This Right …": Words That Help Build Empathy. *Annals of Internal Medicine*, *135*(3), 221.

Davis, M. H. (1980). A Multidimensional Approach to Individual Differences in Empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85. Retrieved on 5 September 2020 from https://www.uv.es/~friasnav/Davis_1980.pdf.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126.

Davis, M. H., & Franzoi, S. L. (1991). Stability and change in adolescent self-consciousness and empathy. *Journal of Research in Personality*, *25*(1), 70–87.

Dohrenwend, A. M. (2018). Defining Empathy to Better Teach, Measure, and Understand Its Impact. *Academic Medicine*, *93*(12), 1754–1756.

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, *55*(4), 399–410.

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.

Ferreira-Valente, A., Monteiro, J. S., Barbosa, R. M., Salgueira, A., Costa, P., & Costa, M. J. (2017). Clarifying changes in student empathy throughout medical school: A scoping review. *Advances in Health Sciences Education*, *22*(5), 1293–1313.

Frank, J. R. (ed.). (2005). *Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada*. Retrieved on 5 September 2020 from http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf.

Graf, J., Smolka, R., Simoes, E., Zipfel, S., Junne, F., Holderried, F., Wosnik, A., Doherty, A. M., Menzel, K., & Herrmann-Werner, A. (2017). Communication skills of medical students during the OSCE: Gender-specific differences in a longitudinal trend study. *BMC Medical Education, 17*(1), 75.

Heintze, U., Radeborg, K., Bengtsson, H., & Stenlåås, A. (2004). Assessment and evaluation of individual prerequisites for dental education. *European Journal of Dental Education*, *8*(4), 152–160.

Hemmerdinger, J. M., Stoddart, S. D. R., & Lilford, R. J. (2007). A systematic review of tests of empathy in medicine. *BMC Medical Education*, *7*(1), 24.

Hochschulzulassungsgesetz. (2016). *Landesvorschriften und Landesrechtsprechung*. Retrieved on 5 September 2020 from http://www.gesetze-rechtsprechung.sh.juris.de/jportal/portal/t/ki3/page/bsshoprod.psml?pid=Dokumentanzeige&showdoccase=1&js_peid=Trefferliste&fromdoctodoc=yes&doc.id=jlr-HZGSH2016rahmen&doc.part=X&doc.price=0.0&doc.hl=0#focuspoint.

Hojat, M. (2014). Assessments of empathy in medical school admissions: What additional evidence is needed? *International Journal of Medical Education*, *5*, 7–10.

Hojat, M., Louis, D. Z., Markham, F. W., Wender, R., Rabinowitz, C., & Gonnella, J. S. (2011). Physicians' empathy and clinical outcomes for diabetic patients. *Academic Medicine*, *86*(3), 359–364.

Hojat, M., Mangione, S., Kane, G. C., & Gonnella, J. S. (2005). Relationships between scores of the Jefferson Scale of Physician Empathy (JSPE) and the Interpersonal Reactivity Index (IRI). *Medical Teacher*, *27*(7), 625–628.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J. M., Gonnella, J. S., Erdmann, J. B., Veloski, J., & Magee, M. (2001). The Jefferson Scale of Physician Empathy: Development and Preliminary Psychometric Data. *Educational and Psychological Measurement*, *61*(2), 349–365.

Hojat, M., Vergare, M. J., Maxwell, K., Brainard, G., Herrine, S. K., Isenberg, G. A., Veloski, J., & Gonnella, J. S. (2009). The devil is in the third year: A longitudinal study of erosion of empathy in medical school. *Academic Medicine*, *84*(9), 1182–1191.

Huang, Y.-R., & Chang, S.-M. (2004). Academic and Cocurricular Involvement: Their Relationship and the Best Combinations for Student Growth. *Journal of College Student Development*, *45*(4), 391–406.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment Interview Reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*(3), 264–276.

Kadmon, G., & Kadmon, M. (2016). Academic Performance of Students with the Highest and Mediocre School-leaving Grades: Does the Aptitude Test for Medical Studies (TMS) Balance Their Prognoses? *GMS Journal for Medical Education*, *33*(1), 7.

Kadmon, G., Resch, F., Duelli, R., & Kadmon, M. (2014). Predictive Value of the School-leaving Grade and Prognosis of Different Admission Groups for Academic Performance and Continuity in the Medical Course – a Longitudinal Study. *GMS Journal for Medical Education, 31*(2), 21.

Kerasidou, A., & Horn, R. (2016). Making space for empathy: Supporting doctors in the emotional labour of clinical care. *BMC Medical Ethics*, *17*(1), 8.

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), 106.

Kötter, T., Obst, K., Brüheim, L., Eisemann, N., Voltmer, E., & Katalinic, A. (2017). Können psychometrische Tests den Erfolg im Auswahlgespräch zum Medizinstudium vorhersagen? Eine Querschnittsstudie an einer deutschen Hochschule. *Das Gesundheitswesen*, *79*(07), e40–e47.

Kötter, T., Rose, S. I., Waldmann, A., & Steinhäuser, J. (2020). Do Medical Students in Their Fifth Year of Undergraduate Training Differ in Their Suitability to Become a 'Good Doctor' Depending on Their Admission Criteria? A Pilot Study. *Advances in Medical Education and Practice*, *11*, 109–112.

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*(1), 241–293.

Menendez, M. E., Chen, N. C., Mudgal, C. S., Jupiter, J. B., & Ring, D. (2015). Physician Empathy as a Driver of Hand Surgery Patient Satisfaction. *The Journal of Hand Surgery*, *40*(9), 1860-1865.e2.

Mercer, S. W., & Reynolds, W. J. (2002). Empathy and quality of care. *The British Journal of General Practice*, *52*, 9–12.

Mommert, A., Wagner, J., Jünger, J., & Westermann, J. (2020). Exam performance of different admission quotas in the first part of the state examination in medicine: A cross-sectional study. *BMC Medical Education*, *20*(1), 169.

Neumann, M., Edelhäuser, F., Tauschel, D., Fischer, M. R., Wirtz, M., Woopen, C., Haramati, A., & Scheffer, C. (2011). Empathy Decline and Its Reasons: A Systematic Review of Studies with Medical Students and Residents. *Academic Medicine*, *86*(8), 996–1009.

Neumann, M., Scheffer, C., Tauschel, D., Lutz, G., Wirtz, M., & Edelhäuser, F. (2012). Physician empathy: Definition, outcome-relevance and its measurement in patient care and medical education. *GMS Journal for Medical Education*, *29*(1), 11.

Obst, K. U., Brüheim, L., Westermann, J., Katalinic, A., & Kötter, T. (2016). Are the results of questionnaires measuring non-cognitive characteristics during the selection procedure for medical school application biased by social desirability? *GMS Journal for Medical Education*, *33*(5), 75.

Ones, D. S., & Viswesvaran, C. (2011). The Effects of Social Desirability and Faking on Personality and Integrity Assessment for Personnel Selection. *Human Performance*, *11*(2–3), 245–269.

Papageorgiou, A., Miles, S., & Fromage, M. (2019). Does medical students' empathy change during their 5-year MBBS degree? *Education for Health*, *31*(3), 142–147.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, *50*(1), 36–60.

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. *Universität des Saarlandes*. Retrieved on 5 September 2020 from http://bildungswissenschaften.uni-saarland.de/personal/paulus/empathy/SPF_Artikel.pdf.

Paulus, C. (2012). *Ist die Bildung eines Empathiescores in der deutschen Fassung des IRI sinnvoll? Universität des Saarlandes*. Retrieved on 5 September 2020 from https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/23403/1/Empathiescore.pdf.

Paulus, C. (2019). Ist Empathie eine Lüge? Zur Frage der sozialen Erwünschtheit in der Empathiemessung. *Saarländische Universitäts- und Landesbibliothek*. Retrieved on 5 September 2020 from https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/27668/1/Empathie%20und%20soziale%20Erw%C3%BCnschtheit.pdf.

Pedersen, R. (2009). Empirical research on empathy in medicine – A critical review. *Patient Education and Counseling*, *76*(3), 307–322.

Peeters, H., & Lievens, F. (2005). Situational Judgment Tests and their Predictiveness of College Students' Success: The Influence of Faking. *Educational and Psychological Measurement*, *65*(1), 70–89.

Pohontsch, N. J., Stark, A., Ehrhardt, M., Kötter, T., & Scherer, M. (2018). Influences on students' empathy in medical education: An exploratory interview study with medical students in their third and last year. *BMC Medical Education*, *18*(1), 231.

Röding, K. (2005). Perceptions of admission committee members: Some aspects on individual admission to dental education at Karolinska Institutet. *European Journal of Dental Education*, *9*(3), 115–122.

Saguil, A., Dong, T., Gingerich, R. J., Swygert, K., LaRochelle, J. S., Artino, A. R., Cruess, D. F., & Durning, S. J. (2015). Does the MCAT predict medical school and PGY-1 performance? *Military Medicine*, *180*(4), 4–11.

Schwibbe, A., Lackamp, J., Knorr, M., Hissbach, J., Kadmon, M., & Hampe, W. (2018). Medizinstudierendenauswahl in Deutschland: Messung kognitiver Fähigkeiten und psychosozialer Kompetenzen. *Bundesgesundheitsblatt*, *61*(2), 178–186.

Sebok, S. S., Luu, K., & Klinger, D. A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and Rasch analyses. *Advances in Health Sciences Education*, *19*(1), 71–84.

Sladek, R.M., Bond, M.J., Frost, L.K., & Prior, K.N. (2016). Predicting success in medical school: A longitudinal study of common Australian student selection tools. *BMC Medical Education*, *16*(1), 187.

Staatsvertrag. (2019). *Staatsvertrag über die Hochschulzulassung*. Retrieved on 5 September 2020 from https://hochschulstart.de/fileadmin/downloads/gesetze/StV_2019.pdf.

Stiftung für Hochschulzulassung [SFH]. (2019). *Daten der bundesweit zulassungsbeschränkten Studiengänge an Hochschulen*. Retrieved on 16 March 2021 from https://hochschulstart.de/fileadmin/user_upload/bew_zv_ws19.pdf.

Swing, S.R. (2007). The ACGME outcome project: Retrospective and prospective. *Medical Teacher*, *29*(7), 648–654.

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, *21*(1), 11–27.

Wang, H., Kline, J.A., Jackson, B.E., Laureano-Phillips, J., Robinson, R.D., Cowden, C.D., d'Etienne, J.P., Arze, S.E., & Zenarosa, N.R. (2018). Association between emergency physician self-reported empathy and patient satisfaction. *PloS One*, *13*(9), e0204113.

Werwick, K., Winkler-Stuck, K., & Robra, B.-P. (2018). From HAM-Nat to the "Physikum" – Analysis of the study success parameters before and after the introduction of a science test in the approval procedure. *GMS Journal for Medical Education*, *35*(3), 30.

West, C.P., & Shanafelt, T.D. (2007). The influence of personal and environmental factors on professionalism in medical education. *BMC Medical Education*, *7*(1), 29.

Wilkinson, H., Whittington, R., Perry, L., & Eames, C. (2017). Examining the relationship between burnout and empathy in healthcare professionals: A systematic review. *Burnout Research*, *6*, 18–29.

Zachariae, R., Pedersen, C.G., Jensen, A.B., Ehrnrooth, E., Rossen, P.B., & von der Maase, H. (2003). Association of perceived physician communication style with patient satisfaction, distress, cancer-related self-efficacy, and perceived control over the disease. *British Journal of Cancer*, *88*(5), 658–665.

Information about the authors:

Josefin Wagner, M. Sc., EMPH
Dr. phil. Alex Mommert
Division of Study and Teaching, Faculty of Medicine

Prof. Dr. med. Jürgen Westermann
Institute of Anatomy
University of Lübeck
Ratzeburger Allee 160
23562 Lübeck, Germany

Email: josefin.wagner@uni-luebeck.de
        alex.mommert@uni-luebeck.de
        westermann@anat.uni-luebeck.de

Josefin Wagner is a scientific assistant at the Division of Study and Teaching, Faculty of Medicine, University of Lübeck. Since 2015, she has been focusing on student admission and concomitant research, evaluation and development of teaching.

Dr. Alex Mommert earned his doctoral degree in Psychology, addressing social recognition, social identity and motivation. As a former colleague at the Division of Study and Teaching, Faculty of Medicine, University of Lübeck, his focus between 2017 and 2019 has been on interprofessional education, student admission and concomitant research. He remains an associated researcher with the University of Lübeck.

Prof. Dr. Jürgen Westermann earned his doctoral degree in Medicine at the Medical School of Hannover, Germany. In 2001, he became the head of the Institute of Anatomy at the University of Lübeck, and has been the Dean of Student Affairs at the Faculty of Medicine since 2002.