Measuring personal characteristics in applicants to German medical schools: Piloting an online Situational Judgement Test with an open-ended response format

Abstract

Objectives: Situational Judgement Tests (SJT) are a cost-efficient method for the assessment of personal characteristics (e.g., empathy, professionalism, ethical thinking) in medical school admission. Recently, complex open-ended response format SJTs have become more feasible to conduct. However, research on their applicability to a German context is missing. This pilot study tests the acceptability, reliability, subgroup differences, and validity of an online SJT with open-ended response format developed in Canada ("Casper").

Methods: German medical school applicants and students from Hamburg were invited to take Casper in 2020 and 2021. The test consisted of 12 video- and text-based scenarios, each followed by three open-ended questions. Participants subsequently evaluated their test experience in an online survey. Data on sociodemographic characteristics, other admission criteria (Abitur, TMS, HAM-Nat, HAM-SJT) and study success (OSCE) was available in a central research database (stav).

Results: The full sample consisted of 582 participants. Test-takers' global perception of Casper was positive. Internal consistency was satisfactory in both years (α=0.73; 0.82) while interrater agreement was moderate (ICC(1,2)=0.54). Participants who were female (d=0.37) or did not have a migration background (d=0.40) received higher scores. Casper scores correlated with HAM-SJT (r=.18) but not with OSCE communication stations performance. The test was also related to Abitur grades (r=-.15), the TMS (r=.18), and HAM-Nat logical reasoning scores (r=.23).

Conclusion: This study provides positive evidence for the acceptability, internal consistency, and convergent validity of Casper. The selection and training of raters as well as the scenario content require further observation and adjustments to a German context to improve interrater reliability and predictive validity.

Keywords: admission, situational judgement test, personal characteristics, Casper

Mirjana Knorr[1]
Ina Mielke[1]
Dorothee Amelung[2]
Mahla Safari[2]
Oana R. Gröne[1]
Simon M. Breil[3]
Alexander MacIntosh[4]

1 University Medical Center Hamburg-Eppendorf, Arbeitsgruppe Auswahlverfahren, Hamburg, Germany

2 University of Heidelberg, Heidelberg, Germany

3 University of Münster, Münster, Germany

4 Acuity Insights, Toronto, Canada

1. Introduction

1.1. Background

Personal characteristics of future physicians such as ethical thinking, professionalism, and social skills, have gained increased importance in competency frameworks for medical education [1], [2], [3]. Likewise, these characteristics were emphasized in the "Masterplan Medizinstudium 2020", a 2017 resolution by the federal and regional governments of Germany to regulate the reformation of medical curricula [4]. One of the directives in the resolution was to not exclusively focus on high-school grades or results of aptitude tests [5] but to attach more importance to personal characteristics in the admission process [4]. The current main methods used to evaluate such characteristics are traditional or multiple mini-interviews (MMIs) [6] and professional pre-qualifications (i.e. completed vocational training, volunteer work). However, both methods have limitations. Interviews are considered inefficient and resource-intensive for the assessment of an entire pool of multiple thousand applicants, especially considering the amount of interviewer time needed [7]. Although preliminary supporting evidence exists that (when controlling for Abitur grade and cognitive test performance) a vocational training can predict study success [8] it is yet unclear to what extent professional pre-qualifications are indicative of personal characteristics or clinical skills [9]. The fairness of professional pre-qualifications as selection criteria can also be questioned as not every applicant has the opportunity to volunteer or to complete a three-year vocational training.

Therefore, we suggest Situational Judgement Tests [10] as promising cost-efficient and evidence-based alternatives to interviews and professional pre-qualifications. SJTs present candidates with several short situation descriptions (scenarios) in a text or video format followed by instructions to identify what one would or should do in the described situation. Internationally, SJTs used for medical selection demonstrate good psychometric properties [11] with a recent meta-analysis reporting a pooled estimate of r=.32 for predicting interpersonal performance evaluations [12]. Traditionally, SJTs use a closed-ended response format (i.e., choosing from, rating, or ranking a list of response alternatives). Due to technological advances, open-ended response format SJTs (i.e., applicants provide their response to an SJT scenario in a written text or in an audio/video format) have recently become more feasible [13]. Research indicates that these types of response formats might reduce minority-majority differences (i.e., performance differences between natives and immigrants) because multiple choice formats require more cognitive resources to understand and compare each of the provided response options whereas open-ended questions can be responded to when the core dilemma of a scenario is understood [13]. In addition, it is assumed that open-ended response formats are less prone to faking [14]. In health-care selection, research on open-ended response format SJTs has focused on Casper (formerly known as: Computer-based Assessment for Sampling Personal Characteristics), a digitally administered SJT which is currently offered in English and French. In these studies, Casper demonstrated good acceptability and reliability [15], [16], fewer minority-majority performance differences compared to cognitive tests [17], and a correlation with later performance at licensure exam subtests which focus on communicational and ethical aspects [18].

Despite their potential benefits compared to interviews or professional pre-qualifications, SJTs currently play a minor role in German medical admission and supporting evidence is limited. The University of Heidelberg developed a video-based SJT for self-assessment purposes [19] and the University of Hamburg recently introduced a paper-pencil SJT (Hamburger Situational Judgement Test, HAM-SJT) for their undergraduate admission process [20]. Both SJTs use a closed-ended response format and to our knowledge, an SJT with an open-ended format has not yet been tested in a medical selection process in Germany.

1.2. Aim of the study

In this study, we piloted Casper as an online-SJT with an open-ended response format that could potentially be administered for high-stakes testing in Germany in the future. Our goal was to analyze the acceptability, reliability, subgroup performance differences as well as the convergent (i.e., relationship to other measures of personal characteristics) and discriminant (i.e., relationship to cognitive admission criteria) validity in comparison to the international evidence on Casper.

2. Methods

2.1. Procedure

This study took place on five test dates over the summers of 2020 and 2021. Applicants were invited to sign up for one of the test dates if they had registered for any of the major German medical school admission tests (Test für medizinische Studiengänge (TMS), Hamburger Naturwissenschaftstest (HAM-Nat), Hamburger Situational Judgement Test (HAM-SJT), see table 1) and had indicated their interest to participate in research studies on student selection. In addition, all medical students at the University of Hamburg, irrespective of study year, received an invitation to take part in this study via an electronic student newsletter. To incentivize study participation, all participants received feedback on their Casper performance and had the chance to win vouchers over 50€ for an online store. Test fees were not charged in this study but can roughly be estimated to range between 46 and 95 EUR based on the current pricing (2024) in North-America.

2.2. Casper

Casper focuses on assessing inter-individual differences in ten personal characteristics including collaboration, communication, empathy, equity, ethics, motivation, problem solving, professionalism, resilience, and self-awareness. Each scenario is usually designed to measure more than one characteristic and for each participant the composition of different scenarios ensures all ten characteristics are covered. In line with findings that such characteristics cannot be reliably discriminated within SJTs [21], [22], Casper only provides one overall score.

In this study, the assessment consisted of eight video and four text scenarios. Each scenario was accompanied by three questions and participants were asked to provide their responses in an open text format within a 5-minute time limit per scenario. English language scenarios were selected from an existing pool of which six were used both in 2020 and 2021 while the six other scenarios varied between years to include a broader variety of scenarios. Video dialogues and questions were translated into German by the German research team: A linguist and public health scientist fluent in English wrote the transcripts of the video dialogues, which then were translated into German by a German-native psychologist. This translation was reviewed by a third person (German-native psychologist). Discrepancies were discussed and solved within the team. Videos were either subtitled (2020) or provided with a voice-over (2021). Participants took the test via the Casper online-platform. English language examples of typical Casper scenarios and questions can be

Table 1: Overview of instruments and their reliability, descriptive statistics in Casper study population, and correlation between Casper score and each of the instruments

Instrument	Description	Scale	Reliability	*n*	*M*	*SD*	r_{Casper}
Casper	Casper performance at latest test date	z-standardized mean over 12 scenarios	α_{2020} = 0.73 α_{2021} = 0.82	582	-0,02	0.99	1
Abitur	Self-reported Abitur grade, comparable to school-leaving grade point average (GPA)	scale from 1 (highest performance) to 6 (lowest performance)	NA	354	1.72	0.44	-.15**
TMS	Test für medizinische Studiengänge: Subject-specific admission test for medicine and other healthcare studies	standardized score with a mean of 100 and a standard deviation of 15	.91≤ α ≤ .92 [49], [50]	371	103.1	9.52	.18**
HAM-Nat	Hamburger Naturwissenschaftstest: Multiple-choice test with three subtests	ability parameters based on Item Response Theory (IRT) model					
	1) natural sciences (60 items)		.88 ≤ α^1 ≤ .91	270	0.57	1.22	.04
	2) arithmetic problem solving (15 items)		.36 ≤ α^1 ≤ .55	270	0.02	0.59	.08
	3) logical reasoning (15 items)		.40 ≤ α^1 ≤ .62	270	0.14	0.60	.23**
HAM-SJT	Hamburger Situational Judgement Test: Contextualized paper-pencil SJT with a should do instruction and rate the effectiveness item format (75 to 80 items)	mean over squared deviations between intra-individually z-standardized applicants' responses and mean intra-individually z-standardized expert panel response for each item, multiplied by -1 (i.e., higher values = better performance)	.71 ≤ α^1 ≤ .81	263	-0.41	0.14	.18**
OSCE	Objective structured clinical exam taken after one and a half years of medical school with 12 stations (10 during COVID pandemic) including:	percentage of achieved points	α = .75				
	1) History taking station			94	79.18	9.31	-.09
	2) Communication skills station			55	81.45	10.83	.08

NA=not available, *n*=number of participants, *M*=Mean, *SD*=Standard deviation, r_{Casper}=Pearson correlation with Casper
*p<.05, **p<.01, [1]range for different test dates

found on the official website [https://acuityinsights.app/test-prep-casper/].

In 2020, 52 faculty staff members and student assistants from different German universities rated participants' responses. Of these, 15 provided their ratings again the following year. In line with widening participation policies it is recommended to include raters that reflect patient diversity and promote inclusivity in medicine within rater-based selection tools [23], [24], [25] in order to reduce bias and enhance fairness by considering different perspectives and backgrounds in the evaluation process. Thus, to diversify the rater pool for the 2021 study, we recruited 11 additional community raters via online platforms for temporary job offers and e-mail lists of associations for people with a migration background. All raters completed an online on-demand training offered in English (2020) or German (2021) language. On average, raters needed 46.19 seconds (*SD*=22.72) for the rating of one response with a mean count of 125.60 words (*SD*=38.05). Faculty raters completed their ratings within their working hours while community raters were compensated with a voucher for an online store (0.50 EUR per rated response). After completing their ratings, raters in the 2021 study were asked to provide sociodemographic data in a voluntary survey.

Each response to a scenario was evaluated by one (2020) or two raters (2021) on a 9-point global rating scale ranging from 1="poor" to 9="excellent" with no specific behavioral anchors. For each scenario, raters received guidelines on how to consider the specific construct(s) the scenario was designed to measure in their ratings. They were instructed to rate the quality of each response relative to the corresponding ones provided by other participants.

Raters were assigned responses through an online rating platform. After a certain number of ratings, they were able to switch to a new scenario to avoid fatigue. For each individual candidate, an algorithm of the online platform ensured that each scenario was rated by a different rater. In case of two raters, both ratings were averaged to generate a scenario score. The overall Casper score is delivered as a mean over twelve scenarios z-standardized within a cohort.

2.3. Other measures

All study participants had previously agreed to take part in an ongoing research project (Studierendenauswahlverbund, "stav", [https://www.projekt-stav.de/index.php] where admission data, study performance data of admitted students, and data from other research studies and a sociodemographic questionnaire (see attachment 1) are matched and stored in a central database. Casper data could thereby be matched to the following data sources available in this database. A summary of all instruments can also be found in table 1.

2.3.1. Acceptability

Upon completion of the Casper test, participants were directed to an online survey about their test experience. In addition to an overall evaluation of Casper on a 10-point scale, candidates were asked, for example, to indicate their perception of the fairness and difficulty of Casper on 7-point scales (the higher the evaluation, the more favorable; see attachment 2). Survey data was only available for the 2020 test dates.

2.3.2. Sociodemographic characteristics

To compare this study to previous findings on subgroup differences in SJTs [17], [26], [27], we included gender, parents' highest level of education (i.e., at least one of the parents holds an academic degree) as indicator for socio-economic status (SES) as well as "migration background" as indicator for ethnicity/nationality. Following the definition of the German census [28] a migration background was considered if at least one of the following was true: the person was not born in Germany, has a non-German citizenship, or one of the parents was not born in Germany.

2.3.2. Validity

To study convergent validity, two additional measures were included: the *HAM-SJT* and communication performance in an Objective Structured Clinical Exam (*OSCE*). The *HAM-SJT* is a paper-pencil SJT with a closed-ended response format that was added to the admissions process to medical school at the University of Hamburg in 2020 [20], [29]. Students at the University of Hamburg typically take their first OSCE, an exam that consists of several short standardized interactions (stations) evaluated by raters [30], after one and a half years of studies. Since medical students from all cohorts were invited to take part in this study, our participants took this OSCE between 2016 and 2022. Between these years the twelve stations of this OSCE were comparable in terms of content and rating checklists. We used the results (in percent) of two stations with simulated patients designed to target communication skills (communication skills station, history taking station) [31]. Data for the communication skills station was only available for students who took the OSCE before the summer of 2020 because this station could not take place during the COVID-19 pandemic. For the analysis of discriminant validity we compared the Casper results to cognitive admission criteria including the German *Abitur grade* (equivalent to school-leaving grade point average), performance at the cognitive admission test *HAM-Nat*, a multiple-choice test with subtests on knowledge in natural sciences [32], arithmetic problem solving, and logical reasoning, and performance at the *Test für medizinische Studiengänge (TMS)*, a subject-specific admission test for medicine and other healthcare studies [33].

Table 2: Characteristics of study participants

Sample description		Casper performance
N	**582**	
Age at time of Casper		r = 0.09, p = 0.03
M	21.26	
SD	3.31	
Gender		d = 0.37
NA (percentage of N)	186 (32%)	
n male (percentage male in non-NA cases)	77 (19%)	M = -0.36, SD = 1.08
n female (percentage female in non-NA cases)	319 (81%)	M = 0.01, SD = 0.96
Migration background		d = 0.40
NA (percentage of N)	195 (34%)	
n yes (percentage yes in non-NA cases)	141 (36%)	M = -0.31, SD = 1.05
n no (percentage no in non-NA cases)	246 (64%)	M = 0.08, SD = 0.93
Parents' highest level of education		d = 0.15
NA (percentage of N)	209 (36%)	
n university degree (percentage academic in non-NA cases)	264 (71%)	M = -0.07, SD = 0.99
n no university degree (percentage non-academic in non-NA cases)	109 (29 %)	M = 0.07, SD = 0.98

N=overall number of participants, *n*=number of participants in subsample, *M*=Mean, *SD*=Standard deviation, *r*=Pearson correlation, *d*=Cohen's *d*, NA=not available

2.4. Data analysis

All analyses were conducted in R-4.2.1 [https://www.r-project.org/]. For the analysis of participants' responses to the acceptability questionnaire, we calculated basic descriptive statistics for quantitative evaluations and counted the frequencies of commonly mentioned topics in open text format questions using MAXQDA 2022 [https://www.maxqda.com/de/]. Reliability of Casper was analyzed in terms of internal consistency over 12 scenarios (Cronbach's alpha). For responses that were rated by two independent raters (2021 sample), we analyzed interrater agreement by means of intra class correlation (ICC(1,2)). We investigated individual subgroup differences in mean performance with Welch *t*-tests for independent samples; effect sizes were reported as Cohen's *d*. Convergent and discriminant validity was analyzed using Pearson correlations.

We based analyses of subgroup differences and validity on the overall sample. For cases in which participants took part in both years, the z-score of the more recent Casper date (2021) was used. Unpaired Welch *t*-Tests and Mann-Whitney-U-Tests were conducted to ensure that performance on study variables was comparable between study cohorts. The level of significance for all analyses was α=.05. The R code, a full data analysis report, all appendices, and information on how to request the original data can be found at [https://osf.io/9daz3/].

3. Results

3.1. Participants and raters

Overall, 582 individuals participated in this pilot study including 74 medical students and 508 applicants. Twenty participants took the Casper in both 2020 and 2021. Participants' mean age was 21 years (*SD*=3.30). Further sociodemographic information was available for around 64% of the participants. In this subsample, 19% identified as male, 36% had a migration background, and 71% had at least one parent holding a university degree (see table 2). Age, performance on Casper and other study variables were largely comparable between study cohorts (see attachment 3, p.1-2). Only HAM-SJT performance was significantly better in the 2021 cohort compared to the 2020 cohort (W=3773.5, p<.001, d=0.62). Applicants and medical students did not differ in their average Casper performance (t(91.226)=-1.16, p=0.25, d=0.16). Average performance in six video scenarios that were used both in 2020 (subtitles) and 2021 (voice-over) did not differ between years (t(465.16)=-0.48, p=0.63, d=0.04).

Of the 26 raters in 2021, 15 of the faculty and 6 of the community raters provided demographic data (see table 3). Most notably, community raters had a more diverse educational background as compared to faculty raters (33% vs. 83% holding a university degree).

Table 3: Characteristics of community and faculty raters in 2021

	Community raters	Faculty raters
N	11	15
n survey (percentage of *N*)	6 (55%)	12 (80%)
Gender		
male (percentage)	0 (0%)	3 (25%)
female	6	9
Age (in years)		
18 – 30 (percentage)	3 (50%)	6 (50%)
31 – 40	1	3
41 – 50	1	3
51 – 60	1	0
Highest level of education		
University degree* (percentage)	2 (33%)	10 (83%)
Vocational training	2	0
Secondary school level: Abitur	1	2**
Secondary school level: Mittlere Reife	1	0

*includes: Bachelor, Master, Diplom, Magister, PhD, **Currently studying medicine or psychology

Table 4: Multiple regression analyses predicting Casper by sociodemographic variables (model 1) controlling for native language (model 2) and cognitive ability (model 3) (n=227)

	Model 1			Model 2			Model 3		
	b	***SE***	***p***	***b***	***SE***	***p***	***b***	***SE***	***p***
Male gender	-0.33	0.16	.042	-0.36	0.16	.027	-0.37	0.16	.018
Migration background	-0.40	0.14	.004	-0.18	0.16	.271	-0.08	0.16	.614
German as native language				0.59	0.23	.012	0.54	0.23	.019
Abitur grade							0.02	0.16	.899
TMS							0.02	0.01	.003
R^2 / Δ R^2 (*p*)	0.06			0.09 / 0.03 (.010)			0.13 / 0.04 (.005)		

b=unstandardized regression weight, *SE*=standard error

3.2. Acceptability

Overall, participants of the 2020 study evaluated Casper favorably with a mean rating of 7.55 (*SD*=1.64, *n*=368) on a 10-point scale. On 7-point scales, participants indicated that they were satisfied with their overall test experience (*M*=5.40, *SD*=1.19, *n*=367) and perceived Casper as rather fair (*M*=5.24, *SD*=1.26, *n*=354). Participants evaluated Casper as a bit less stressful when asked to compare it to other exams in general (*M*=3.24, *SD*=1.50, *n*=359) and perceived it as neither difficult nor easy (*M*=4.08, *SD*=1.21, *n*=356). In the open text format questions, the most frequently criticized aspect regarding the test experience was the short response time which made some participants feel that the test could systematically disadvantage applicants with less typing experience (*n*=24) (see attachment 2 for full results).

3.3. Reliability

The internal consistency for Casper scenario scores was α=0.73, 95% CI [0.69, 0.77] in 2020 and α=0.82, 95% CI [0.79, 0.86] in 2021. For responses evaluated by two raters in 2021, overall interrater agreement was ICC(1,2)=0.54. Re-test reliability for twenty participants who completed Casper in both years was *p*=0.29 (Spearman's rank correlation).

3.4. Subgroup differences

Single group comparisons revealed that female participants (*t*(107.16)=2.73, *p*=0.01, *d*=0.37) and participants without a migration background (*t*(263.09)=3.65, *p*<.001, *d*=0.40) showed a better mean Casper performance compared to male participants and participants with a migration background, respectively. Casper performance did not significantly differ depending on parents' level of education (*t*(203.67)=1.30, *p*=0.19, *d*=0.15). Follow-up regression analyses with Casper performance as outcome variable revealed that adding native language as predictor explained the effect of migration background while gender and language remained significant predictors when controlling for cognitive criteria (see table 4).

3.5. Convergent and discriminant validity

With respect to other measures of personal characteristics, Casper had a significant relationship with HAM-SJT performance (r=.18, p=.004, n=263) but was neither related to performance at the OSCE history taking station (r=-.09, p=.37, n=94) nor to the communication skills station (r=.08, p=.57, n=55).
Regarding cognitive admission measures, Casper performance had significant correlations with the Abitur grade (r=-.15, p=.01, n=354; i.e. the better the Abitur grade, the better Casper performance), TMS performance (r=.18, p=.001, n=371), and the logical reasoning subtest of the HAM-Nat (r=.23, p<.001, n=270). On the other hand, it did not correlate with the HAM-Nat science (r=.04, p=.46, n=270) nor with the arithmetic problem solving subtest (r=.08, p=.18, n=270) (see table 1). Attachment 3 includes a full correlation table for all study variables.

4. Discussion

In German medical education, text-based and video-based SJTs have been developed and suggested for the (self-)assessment, teaching and monitoring of relevant skills such as communication or professional behavior of medical school applicants and students [19], [20], [34], [35], [36]. While all these examples rely on a closed-ended response format, this is the first study piloting an online-SJT with open-ended response format in a German medical admission context.
Similar to Canadian reports on Casper [16], participants' perception of Casper was favorable and internal consistency was good. These results also align with positive perceptions as well as satisfactory internal consistency values for the Heidelberg video-SJT ($0.81 \leq \alpha \leq .83$) [19] and HAM-SJT ($0.62 \leq \alpha \leq .82$) [37]. On the other hand, interrater agreement in our study was only moderate and diverged from the high rater agreement (0.95) found in the Canadian pilot study of Casper [15]. In the small subsample of participants who sat the test twice, test-retest reliability was low. This might be explained by individual differences in participants' personal development within the one-year time span between the two test applications but also by changes to the test format between both test applications (i.e. use of different scenarios, voice-over, inclusion of community raters). Nevertheless, the subsample in our study was too small (n=20) to draw definite conclusions and a follow-up study with a targeted test-retest design would be necessary.
Our study revealed significant performance differences in favor of females and participants without a migration background that are in line with a North-American study on Casper [17]. Our follow-up analyses suggest that native language rather than migration background was related to performance differences which diverges from findings in a U.S. study where differences in ethnicity remained when controlled for language use [38]. The open-ended response format did therefore not provide an advantage over the HAM-SJT which similarly showed performance differences depending on native language (d=0.24) [37] or the Heidelberg video-SJT which did not show any significant differences [19].
In support of the convergent and discriminant validity of the test, Casper performance was related to HAM-SJT performance but not to the HAM-Nat science and arithmetic problem solving subtests. Likewise, the Canadian Casper had not been found to be related to the MCAT science subtests [15]. On the other hand, we found weak correlations with the Abitur grade, TMS performance, and the HAM-Nat logical reasoning subtests. The weak reliability values of the HAM-Nat logical reasoning and arithmetic problem solving subtests might have affected the significance and magnitude of the correlation with Casper. Nevertheless, we found a similarly small significant correlation between TMS and Casper pointing in a similar direction and results are also in line with findings that Casper correlates with the verbal reasoning part of the MCAT [15]. This suggests that the cognitive but also non-cognitive competencies reflected in these measures (such as motivation, flexibility, or self-management in Abitur grades [39]) could be beneficial for Casper performance. The results also point to a somewhat higher cognitive load in Casper compared to the HAM-SJT or Heidelberg video-SJT which were either negatively related to Abitur grade, TMS and HAM-Nat or not at all [19], [20].
Finally, we did not find any relationship between Casper and two OSCE stations that address communication skills. Thus, we could not replicate positive evidence of predictive validity for the North-American Casper where Casper was related to MMI performance as well as to national licensure exams [15], [18]. HAM-SJT pilot studies, on the other hand, could demonstrate small but significant correlations with subsequent MMI (r=0.22) [20] and OSCE performance (r=0.20) [37].

Limitations

We applied different measures of quality assurance during rater training and the rating process including repeated training rounds if statistics from test ratings fall below pre-determined benchmarks, or temporary retention of raters if they submit their ratings within less time than it needs to read a candidate's response. However, in this pilot these measures were not employed to the same degree as they are in the high-stakes application of Casper. The moderate interrater agreement found in this study highlights the importance of continuously monitoring the rating process and providing feedback to raters.
In the 2021 study, we recruited additional community raters with the aim to diversify the rater pool. Although demographic data somewhat suggest that community raters differed from faculty raters in terms of their level of education, the lower participation rate of community raters in the follow-up survey (55%) makes it difficult to draw definite conclusions about the diversity of our rater pool. Future studies on rater-based selection tools would benefit from a systematic assessment and variation of

raters' sociodemographic characteristics to be able to explore how diverse rater backgrounds impact outcomes of high-stakes selection.

For this pilot, we used scenarios that were developed and previously tested in a North-American high-stakes context. However, it remains unclear whether any cultural differences related to scenario content had an impact on study results. In addition, the participants in our study were volunteers and their motivation to perform will differ from that in a high-stakes selection context. Lastly, we only invited applicants to this study who registered for the TMS and/or HAM-Nat and aimed at improving their chances of gaining a study place. Our sample is therefore not representative of the population of all those interested in studying medicine and likely excludes applicants with a high Abitur grade as well as those who are discouraged by the current selection system and do not apply. However, the latter group might potentially benefit from a non-cognitive test like Casper. For future assessments, it is advised to develop the test content in the culture and language where the test is administered and to confirm the psychometric properties within an actual selection procedure.

Implications for practice and research

A recent study revealed that physicians and medical students in Hamburg do not represent the general population especially in terms of their socio-economic and ethnic background [40]. Medical schools that adopt a widening participation policy need to pay attention to how underrepresented groups perform on a selection criterion when compiling and weighting their selection criteria to minimize adverse impact. Participants' performance in our study did not differ depending on socio-economic background. However, we could only use parents' level of education as indicator. The use of additional indicators such as parents' income or living conditions [40] in future studies might provide a more comprehensive picture. Although our results suggest a potential disadvantage for applicants whose first language is not German, it has been argued internationally that SJTs like Casper can mitigate the often more severe subgroup differences in cognitive tests and thereby potentially widen access to medical school [17], [27]. While preliminary data on the HAM-Nat suggests that applicants without a migration background perform better on the two reasoning subtests ($0.24 \leq d \leq 0.32$) and applicants with a higher socio-economic background perform better on all three HAM-Nat subtest ($0.06 \leq d \leq 0.25$), the magnitude of the effects is small [41]. Currently, to our knowledge, no such data is published for the TMS. Large education studies and reports regularly point to weaker secondary school performance [42], [43] and Abitur grades among students with low socio-economic status (e.g. mean Abitur grade of 2.27 vs. 2.48 in students transitioning to university with a high vs. low socio-economic background [44]) and a migration background (e.g. mean Abitur grade of 2.5 vs. 2.9 in students with a German vs. Turkish background [45]). Nevertheless, the exact statistical magnitude of these subgroup differences in current Abitur grades for those interested in studying medicine is unclear. Systematic studies and comparisons of subgroup differences in German selection criteria depending on applicants' ethnicity and socio-economic background are therefore necessary to evaluate the potential of SJTs to increase or decrease access for these groups and to inform decision makers in their selection strategies.

Since some participants voiced concern that the 5-minute time frame might disadvantage non-native speakers and those with less typing experience, a study of systematic variation of the time limit might shed more light on whether it has the potential to minimize performance differences. An audiovisual response, which seems to further reduce subgroup differences [13], has recently been added to Casper and could be explored in follow-up studies in their potential for a German test version.

German medical schools are called to consider personal characteristics when selecting students [4] and to use selection criteria that indicate their suitability for medical school and the medical profession [46]. It is therefore essential to demonstrate construct and predictive validity. In our study, Casper correlated with non-cognitive selection criteria and cognitive selection criteria in similar magnitude. Thus, it seems that Casper does not merely measure the personal characteristics we aimed to assess but also cognitive characteristics. Therefore, the usefulness of Casper as a meaningful addition to existing selection criteria remains unclear. We could only consider two OSCE stations for a small subsample of study participants. The lack of reliability in a single OSCE station [30] and range restriction in OSCE scores (i.e. students' OSCE performance ranged between 52.5% and 100% of achievable points) are potentially limiting factors in our analysis. Future research should aim to look at different outcome measures of personal characteristics such as, for example, supervisor and peer ratings or a combination of relevant OSCE stations over the course of medical school [47]. Ideally, these should be compared to the predictive validity of other selection criteria that are currently used in conjunction with cognitive criteria: the completion of a vocational training, as well as work and volunteering experience [8].

Finally, from a practical point of view, medical schools need to weigh the costs of a test format like Casper in comparison to alternative selection tools and consider different stakeholders' perspectives. This study demonstrated that with an average rating time of 46 seconds per response, Casper requires less rater time in comparison to multiple mini-interviews with a station time of five to ten minutes [6] and compared to traditional interviews that are less cost efficient in terms of person hours [48]. Likewise, the estimated costs of a maximum of 95 EUR per applicant (2024) are much lower than 450 EUR per applicant (2014) in the Hamburg multiple mini-interview HAM-Int [7]. However, if costs are covered by test fees, the introduction of Casper would come with an additional financial burden for applicants who already pay to take

the TMS (100 EUR in 2024) and HAM-Nat (95 EUR in 2024). A vocational training, on the other hand, provides applicants with the opportunity to learn relevant skills and receive a salary but also requires applicants to invest three years into their training before being able to go to medical school.

5. Conclusions

Positive evaluations by test-takers, good internal consistency, and evidence for discriminant and convergent validity in this study confirm that the test format used in Casper is applicable to a German context. Based on the moderate interrater agreement in our study, the number, background, and training of raters need to be considered and carefully monitored if the test is applied in high-stakes selection. The potential adverse impact on the diversity of students selected by Casper and the current lack of correlation to OSCE performance require potential adjustments to the test and further investigation into the predictive validity of Casper considering a broader range of outcome criteria. It is important to ensure that the test content is relatable to test takers and that it aligns with the goals of German medical education in order to make the test fit for purpose in German medical school selection. In terms of subgroup differences and validity, our current results do not suggest that an open-ended response SJT like Casper is superior to available German SJTs with a closed-ended response format.

Ethics approval and informed consent

All participants gave their informed consent to data collection, storage and matching of the data. This study as part of the stav research project was approved by the local ethics committee at the Department of Medical Psychology, University Medical Center Hamburg-Eppendorf (LPEK-0042). All data was handled in accordance with European data protection laws (GDPR).

Acknowledgements

The authors would like to thank Dieter Münch-Harrach for creating the subtitles for the Casper videos. This study would not have been possible without the volunteer raters from the stav teams in Hamburg, Heidelberg, Münster, Saarbrücken, Berlin and Göttingen as well as members from the Eignung & Auswahl Baden-Wuerttemberg network at the Karlsruhe Institute of Technology, Heidelberg University, DHBW Mannheim, University of Education Weingarten and Pforzheim University.

Funding

This study was conducted as part of the larger stav research project funded by the Federal Ministry of Education and Research, Germany, project number: 01GK1801A-F.
We acknowledge financial support from the Open Access Publication Fund of UKE - Universitätsklinikum Hamburg-Eppendorf.

Authors' ORCIDs

- Mirjana Knorr: [0000-0002-0996-9286]
- Ina Mielke: [0000-0003-1764-5553]
- Dorothee Amelung: [0000-0002-9946-9073]
- Mahla Safari: [0000-0003-0976-8094]
- Oana R. Gröne: [0000-0002-6829-5365]
- Simon M. Breil: [0000-0001-5583-3884]
- Alexander MacIntosh: [0000-0002-5094-3774]

Competing interests

Alexander MacIntosh is a data scientist at Acuity Insights, the company that develops and distributes Casper. The other authors have no competing interests to declare.

Attachments

Available from https://doi.org/10.3205/zma001685

1. Attachment_1.pdf (160 KB)
   Sociodemographic questionnaire of the stav project (2019 version)
2. Attachment_2.pdf (203 KB)
   CASPer exit survey
3. Attachment_3.pdf (232 KB)
   Additional tables

References

1. Frank JR, Snell L, Sherbino J, editors. Can Meds 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015. Zugänglich unter/available from: https://canmeds.royalcollege.ca/en/framework
2. Medizinischer Fakultätentag. Nationaler Kompetenzbasierter Lernzielkatalog Medizin 2015. Berlin: MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V.; 2015. Zugänglich unter/available from: https://medizinische-fakultaeten.de/wp-content/uploads/2021/06/nklm_final_2015-12-04.pdf
3. Association of American Medical Colleges. The Core Competencies for Entering Medical Students. Washington, DC: Association of American Medical Colleges; 2022. Zugänglich unter/available from: https://students-residents.aamc.org/applying-medical-school/article/core-competencies

4. Bundesministerium für Gesundheit. Masterplan Medizinstudium 2020. Berlin: Bundesgesundheitsministerium; 2017. Zugänglich unter/available from: https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/masterplan-medizinstudium-2020.html

5. Schult J, Hofmann A, Stegt SJ. Leisten fachspezifische Studierfähigkeitstests im deutschsprachigen Raum eine valide Studienerfolgsprognose? Z Entwicklungspsychol Pädagog Psychol. 2019;51(1):16-30. DOI: 10.1026/0049-8637/a000204

6. Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME Guide No. 37. Med Teach. 2016;38(5):443-455. DOI: 10.3109/0142159X.2016.1158799

7. Hissbach JC, Sehner S, Harendza S, Hampe W. Cutting costs of multiple mini-interviews - changes in reliability and efficiency of the Hamburg medical school admission test between two applications. BMC Med Educ. 2014;14:54. DOI: 10.1186/1472-6920-14-54

8. Amelung D, Zegota S, Espe L, Wittenberg T, Raupach T, Kadmon M. Considering vocational training as selection criterion for medical students: evidence for predictive validity. Adv Health Sci Educ Theory Pract. 2022;27(4):933-948. DOI: 10.1007/s10459-022-10120-y

9. Erschens R, Herrmann-Werner A, Schaffland TF, Kelava A, Ambiel D, Zipfel S, Loda T. Association of professional pre-qualifications, study success in medical school and the eligibility for becoming a physician: A scoping review. PLoS One. 2021;16(11):e0258941. DOI: 10.1371/journal.pone.0258941

10. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med Teach. 2016;38(1):3-17. DOI: 10.3109/0142159X.2015.1072619

11. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60. DOI: 10.1111/medu.12817

12. Webster ES, Paton LW, Crampton PES, Tiffin PA. Situational judgement test validity for selection: A systematic review and meta-analysis. Med Educ. 2020;54(10):888-902. DOI: 10.1111/medu.14201

13. Lievens F, Sackett PR, Dahlke JA, Oostrom JK, De Soete B. Constructed response formats and their effects on minority–majority differences and validity. J Appl Psychol. 2019;104(5):715-726. DOI: 10.1037/apl0000367

14. Mortaz Hejri S, Ho JL, Pan X, Park YS, Sam AH, Mangardich H, MacIntosh A. Validity of constructed-response situational judgment tests in training programs for the health professions: A systematic review and meta-analysis protocol. PLoS One. 2023;18(1):e0280493. DOI: 10.1371/journal.pone.0280493

15. Dore KL, Reiter HI, Eva KW, Krueger S, Scriven E, Siu E, Hilsden S, Thomas J, Norman GR. Extending the interview to all medical school candidates-computer-based multiple sample evaluation of noncognitive skills (CMSENS). Acad Med. 2009;84:S9-S12. DOI: 10.1097/ACM.0b013e3181b3705a

16. Zou C, McConnell M, Leddy J, Antonacci P, Lemay G. Comparison of the English and French versions of the CASPer® Test in a bilingual population, version 1. MedEdPublish. 2018;7:281. DOI: 10.15694/mep.2018.0000281.1

17. Juster FR, Baum RC, Zou C, Risucci D, Ly A, Reiter H, Miller DD, Dore KL. Addressing the diversity-validity dilemma using situational judgment tests. Acad Med. 2019;94(8):1197-1203. DOI: 10.1097/ACM.0000000000002769

18. Dore KL, Reiter HI, Kreuger S, Norman GR. CASPer, an online pre-interview screen for personal/professional characteristics: prediction of national licensure scores. Adv Health Sci Educ Theory Pract. 2017;22(2):327-336. DOI: 10.1007/s10459-016-9739-9

19. Fröhlich M, Kahmann J, Kadmon M. Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. Int J Sel Assess. 2017;25(1):94-110. DOI: 10.1111/ijsa.12163

20. Schwibbe A, Lackamp J, Knorr M, Hissbach J, Kadmon M, Hampe W. Medizinstudierendenauswahl in Deutschland: Messung kognitiver Fähigkeiten und psychosozialer Kompetenzen [Selection of medical students: Measurement of cognitive abilities and psychosocial competencies]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61(2):178-186. DOI: 10.1007/s00103-017-2670-2

21. Jackson DJ, LoPilato AC, Hughes D, Guenole N, Shalfrooshan A. The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. J Occup Organ Psychol. 2017;90(1):1-27. DOI: 10.1111/joop.12151

22. Mielke I, Breil SM, Amelung D, Espe L, Knorr M. Assessing distinguishable social skills in medical admission: does construct-driven development solve validity issues of situational judgment tests? BMC Med Educ. 2022;22(1):293. DOI: 10.1186/s12909-022-03305-x

23. Melro CM, Pack R, MacLeod A, Rideout A, Watson-Creed G, Burm S. Front row seat: The role MMI assessors play in widening access to medical school. Med Teach. 2023:1-8. DOI: 10.1080/0142159X.2023.2289851

24. Robinett K, Kareem R, Reavis K, Quezada S. A multi-pronged, antiracist approach to optimize equity in medical school admissions. Med Educ. 2021;55(12):1376-1382. DOI: 10.1111/medu.14589

25. O'Sullivan L, Kagabo W, Prasad N, Laporte D, Aiyer A. Racial and Ethnic Bias in Medical School Clinical Grading: A Review. J Surg Educ. 2023;80(6):806-816. DOI: 10.1016/j.jsurg.2023.03.004

26. Herde CN, Lievens F, Jackson DJ, Shalfrooshan A, Roth PL. Subgroup differences in situational judgment test scores: Evidence from large applicant samples. Int J Sel Assess. 2020;28(1):45-54. DOI: 10.1111/ijsa.12269

27. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. Med Educ. 2016;50(6):624-636. DOI: 10.1111/medu.13060

28. Statistisches Bundesamt. Migrationshintergrund. Wiesbaden: Desatis; 2021. Zugänglich unter/available from: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/Glossar/migrationshintergrund.html

29. Corstjens J, Lievens F, Krumm S. Situational judgement tests for selection. In: Goldstein HW, Pulakos ED, Passmore J, Semedo C, editors. The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention. Hoboken (NJ): Blackwell Publ; 2017. p.226-246. DOI: 10.1002/9781118972472.ch11

30. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. Med Teach. 2013;35(9):e1437-e1446. DOI: 10.3109/0142159X.2013.818634

31. Knorr M, Schwibbe A, Ehrhardt M, Lackamp J, Zimmermann S, Hampe W. Validity evidence for the Hamburg multiple mini-interview. BMC Med Educ. 2018;18(1):106. DOI: 10.1186/s12909-018-1208-0

32. Hissbach JC, Klusmann D, Hampe W. Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. BMC Med Educ. 2011;11(1):83. DOI: 10.1186/1472-6920-11-83

33. Kadmon G, Kadmon M. Academic performance of students with the highest and mediocre school-leaving grades: Does the aptitude test for medical studies (TMS) balance their prognoses? GMS J Med Educ. 2016;33(1):Doc7. DOI: 10.3205/zma001006

34. Schick K, Reiser S, Janssen L, Schacht L, Pittroff SID, Dörfler E, Klein E, Roenneberg C, Dinkel A, Fleischmann A, Berberat PO, Bauer J, Gartmeier M. Training in medical communication competence through video-based e-learning: How effective are video modeling and video reflection? Patient Educ Couns. 2024 Apr;121:108132. DOI: 10.1016/j.pec.2023.108132

35. Schubert S, Ortwein H, Dumitsch A, Schwantes U, Wilhelm O, Kiessling C. A situational judgement test of professional behaviour: development and validation. Med Teach. 2008;30(5):528-533. DOI: 10.1080/01421590801952994

36. Institut für Kommunikations- und Prüfungsforschung gGmbH. Studentischer kompetenzorientierter Progresstest 2024. Heidelberg: Institut für Kommunikations- und Prüfungsforschung; 2024. Zugänglich unter/available from: https://www.komp-pt.de/

37. Knorr M, Rudloff A, Breil SM, Schwibbe A. Use of Situational Judgement Tests for Admission into Medical School: Experiences from the University Medical Centre Hamburg. In: 15th Conference of the Differential Psychology, Personality Psychology and Psychological Assessment (DPPD) of the German Psychological Society (DGPs); 2019 Sep 16-18; Dresden, Germany.

38. Pan X, Huang V, Laumbach S, Copeland HL, Akinola M, Rosenbaum D, MacIntosh A. Impact of patterns of language use and socio-economic status on a constructed response Situational Judgment Test (SJT). PLoS One. 2023;18(8):e0289420. DOI: 10.1371/journal.pone.0289420

39. Rindermann H, Oubaid V. Auswahl von Studienanfängern durch Universitäten - Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. J Individ Differ. 1999;20(3):172-191. DOI: 10.1024//0170-1789.20.3.172

40. Groene OR, Huelmann T, Hampe W, Emami P. German physicians and medical students do not represent the population they serve. Healthcare (Basel). 2023;11(12):1662. DOI: 10.3390/healthcare11121662

41. Finger C, Hampe W, Wittenberg T. Eignungstests für das Medizinstudium: Perspektiven auf Diversität und Fairness. In: Bundesministerium für Bildung und Forschung, editor. Vielfalt und Chancengerechtigkeit in Studium und Wissenschaft. Berlin: Bundesministerium für Bildung und Forschung (BMBF); 2023.

42. Autor:innengruppe Bildungsberichterstattung. Bildung in Deutschland 2022. Bielefeld: wbv Publikation; 2022.

43. OECD. PISA 2022 Results (Volume I): The State of Learning and Equity in Education. Paris: OECD Publishing; 2023.

44. Neugebauer M, Schindler S. Early transitions and tertiary enrolment: The cumulative impact of primary and secondary effects on entering university in Germany. Acta Sociologica. 2012;55(1):19-36. DOI: 10.1177/0001699311427747

45. Kristen C, Reimer D, Kogan I. Higher Education Entry of Turkish Immigrant Youth in Germany. Int J Comp Soc. 2008;49(2-3):127-151. DOI: 10.1177/0020715208088909

46. Bundesverfassungsgericht. Leitsätze zum Urteil des Ersten Senats vom 19. Dezember 2017. 1 BvL 3/14, 1 BvL 4/14. Karlsruhe: Bundesverfassungsgericht; 2017. Zugänglich unter/available from: https://www.bverfg.de/e/ls20171219_1bvl000314.html

47. Groene O, Mielke I, Knorr M, Ehrhardt M, Bergelt C. Associations between communication OSCE performance and admission interviews in medical education. Patient Educ Couns. 2022;105(7):2270-2275. DOI: 10.1016/j.pec.2021.11.005

48. Rosenfeld JM, Reiter HI, Trinh K, Eva KW. A cost efficiency comparison between the multiple mini-interview and traditional admissions interviews. Adv Health Sci Educ Theory Pract. 2008;13(1):43-58. DOI: 10.1007/s10459-006-9029-z

49. Mallinger R, Holzbaur C, Mutz N, Prodinger WM, Heidegger M, Hänsgen KD, Spicher B. EMS: Eignungstest für das Medizinstudium in Österreich. Wien/Innsbruck: Medizinische Universität Innsbruck/Medizinische Universität Wien; 2011.

50. Spicher B, Hänsgen KD. EMS 2017 Bericht 24. Eignungstest für das Medizinstudium in der Schweiz. Bericht über Durchführung und Ergebnisse. Granges-Paccot: Zentrum für Testentwicklung und Diagnostik am Departement für Psychologie der Universität Freiburg; 2017.

Corresponding author:
Mirjana Knorr
University Medical Center Hamburg-Eppendorf, Arbeitsgruppe Auswahlverfahren, Martinistr. 52, D-20251 Hamburg, Germany
m.knorr@uke.de

Please cite as
Knorr M, Mielke I, Amelung D, Safari M, Gröne OR, Breil SM, MacIntosh A. Measuring personal characteristics in applicants to German medical schools: Piloting an online Situational Judgement Test with an open-ended response format. GMS J Med Educ. 2024;41(3):Doc30.
DOI: 10.3205/zma001685, URN: urn:nbn:de:0183-zma0016855

This article is freely available from
https://doi.org/10.3205/zma001685

Received: *2023-11-03*
Revised: *2024-03-19*
Accepted: *2024-04-17*
Published: *2024-06-17*

Erfassung von persönlichen Eigenschaften von Medizinstudiumbewerber*innen in Deutschland: Pilotierung eines Online Situational Judgement Tests mit Freitextformat

Mirjana Knorr[1]
Ina Mielke[1]
Dorothee Amelung[2]
Mahla Safari[2]
Oana R. Gröne[1]
Simon M. Breil[3]
Alexander MacIntosh[4]

1 Universitätsklinikum Hamburg-Eppendorf, Arbeitsgruppe Auswahlverfahren, Hamburg, Deutschland

2 Universität Heidelberg, Heidelberg, Deutschland

3 Universität Münster, Münster, Deutschland

4 Acuity Insights, Toronto, Kanada

Zusammenfassung

Ziele: Situational Judgement Tests (SJTs) sind eine kosteneffiziente Methode zur Beurteilung von persönlichen Eigenschaften (z.B. Empathie, Professionalität, ethisches Denken) bei der Zulassung zum Medizinstudium. Die Durchführung komplexer SJTs mit offenem Antwortformat ist in jüngster Zeit einfacher geworden. Es fehlen jedoch Untersuchungen zu ihrer Anwendbarkeit im deutschen Kontext. Diese Pilotstudie testet die Akzeptanz, Reliabilität, Subgruppenunterschiede und Validität eines in Kanada entwickelten Online-SJTs mit offenem Antwortformat („Casper“).
Methoden: Studienbewerber*innen aus Deutschland und Medizinstudierende aus Hamburg wurden eingeladen, Casper in den Jahren 2020 und 2021 zu absolvieren. Der Test bestand aus 12 video- und textbasierten Szenarien, auf die jeweils drei offene Fragen folgten. Die Teilnehmenden bewerteten anschließend ihre Testerfahrung in einer Online-Umfrage. Daten zu soziodemografischen Merkmalen, weiteren Zulassungskriterien (Abitur, TMS, HAM-Nat, HAM-SJT) und zum Studienerfolg (OSCE) waren in einer zentralen Forschungsdatenbank (stav) verfügbar.
Ergebnisse: Die Gesamtstichprobe bestand aus 582 Teilnehmenden. Die allgemeine Wahrnehmung von Casper durch die Testteilnehmenden war positiv. Die interne Konsistenz war in beiden Jahren zufriedenstellend (α=0,73; 0,82), während die Interrater-Übereinstimmung mäßig war (*ICC*(1,2)=0,54). Weibliche Teilnehmerinnen (d=0,37) oder Teilnehmende ohne Migrationshintergrund (d=0,40) erzielten höhere Testwerte. Die Casper Testwerte korrelierten mit dem HAM-SJT (r=.18), aber nicht mit der Leistung in OSCE-Kommunikationsstationen. Der Test zeigte auch Zusammenhänge mit der Abiturnote (r=-.15), dem TMS (r=.18) und dem HAM-Nat-Subtest für logisches Denken (r=.23).
Schlussfolgerungen: Die Studie liefert positive Belege für die Akzeptanz, interne Konsistenz und konvergente Validität von Casper. Die Auswahl und Schulung der Beurteiler*innen sowie die Inhalte der Szenarien bedürfen weiterer Untersuchungen und Anpassungen an den deutschen Kontext, um die Interrater-Reliabilität und prädiktive Validität zu verbessern.

Schlüsselwörter: Studierendenauswahl, Situational Judgement Test, persönliche Eigenschaften, Casper

1. Einleitung

1.1. Hintergrund

Persönliche Eigenschaften zukünftiger Ärzt*innen, wie beispielsweise ethisches Denken, Professionalität und soziale Kompetenz, haben in den Kompetenzrahmenwerken für die medizinische Ausbildung zunehmend an Bedeutung gewonnen [1], [2], [3]. Diese Eigenschaften wurden auch im "Masterplan Medizinstudium 2020" hervorgehoben, einem Beschluss von Bund und Ländern aus dem Jahr 2017, der die Reform der medizinischen Curricula regelt [4]. Eine der Richtlinien des Beschlusses war, sich nicht ausschließlich auf Schulnoten oder Ergebnisse von Eignungstests [5] zu konzentrieren, sondern den persönlichen Eigenschaften im Zulassungsverfahren mehr Bedeutung beizumessen [4]. Die derzeit wichtigsten Methoden zur Bewertung solcher Merkmale sind traditio-

nelle oder multiple Mini-Interviews (MMI) [6] und berufspraktische Vorerfahrungen (z. B. abgeschlossene Berufsausbildung, Dienst). Beide Methoden haben jedoch ihre Grenzen. Interviews gelten als ineffizient und ressourcenintensiv für die Beurteilung von mehreren tausend Bewerber*innen, insbesondere, wenn man den Zeitaufwand für Interviewer*innen berücksichtigt [7]. Obwohl es vorläufige Belege dafür gibt, dass (bei Kontrolle der Abiturnote und der kognitiven Testleistungen) eine Berufsausbildung den Studienerfolg vorhersagen kann [8], ist noch unklar, inwieweit berufspraktische Vorerfahrungen auf persönliche Eigenschaften oder klinische Kompetenzen schließen lassen [9]. Auch die Fairness berufspraktischer Vorerfahrungen als Auswahlkriterium kann in Frage gestellt werden, da nicht jede*r Bewerber*in die Möglichkeit hat, einen Freiwilligendienst oder eine dreijährige Berufsausbildung zu absolvieren.

Vor diesem Hintergrund schlagen wir Situational Judgement Tests [10] als vielversprechende kosteneffiziente und evidenzbasierte Alternative zu Interviews und berufspraktischen Vorerfahrungen vor. Bei SJTs werden den Bewerber:innen mehrere kurze Situationsbeschreibungen (Szenarien) in einem Text- oder Videoformat gezeigt, gefolgt von der Aufgabe, zu beschreiben, was man in der Situation tun würde oder sollte. International zeigen SJTs zur medizinischen Auswahl gute psychometrische Eigenschaften [11], wobei eine aktuelle Metaanalyse einen gepoolten Schätzwert von r=.32 für die Vorhersage von Leistungen in interpersonellen Kriterien ergab [12]. Typischerweise verwenden SJTs ein geschlossenes Antwortformat (d.h. vorgegebene Antwortalternativen werden ausgewählt, bewertet oder in eine Rangreihe gebracht). Aufgrund des technologischen Fortschritts sind SJTs mit offenem Antwortformat (d.h. Bewerber*innen antworten auf ein SJT-Szenario mit einem schriftlichen Text oder einer Audio-/Videoaufnahme) mittlerweile einfacher durchführbar [13]. Forschungsergebnisse deuten darauf hin, dass diese Art von Antwortformaten Subgruppenunterschiede (d.h. Leistungsunterschiede zwischen Personen mit und ohne Migrationshintergrund) verringern könnten, da Multiple-Choice-Formate mehr kognitive Ressourcen erfordern, um jede der angebotenen Antwortmöglichkeiten zu verstehen und zu vergleichen, während offene Fragen beantwortet werden können, wenn das Kerndilemma eines Szenarios verstanden wurde [13]. Darüber hinaus wird angenommen, dass offene Antwortformate weniger anfällig für Verfälschungen sind [14]. Bei der Auswahl von Bewerber*innen im Gesundheitswesen konzentrierte sich die Forschung zu SJTs mit offenem Antwortformat bisher auf Casper (Name ursprünglich abgeleitet von: Computer-based Assessment for Sampling Personal Characteristics), einem digital durchgeführten SJT, der derzeit in englischer und französischer Sprache angeboten wird. In den bisherigen Studien zeigte Casper eine gute Akzeptanz und Reliabilität [15], [16], geringere Subgruppenunterschiede im Vergleich zu kognitiven Tests [17], und eine Korrelation mit der späteren Leistung in Untertests von Approbationsprüfungen, welche den Schwerpunkt auf kommunikative und ethische Aspekte legen [18].

Trotz ihrer potenziellen Vorteile im Vergleich zu Interviews oder beruflichen Qualifikationen spielen SJTs derzeit eine untergeordnete Rolle bei der Zulassung zum Medizinstudium in Deutschland, und es gibt nur wenig wissenschaftliche Evidenz. Die Universität Heidelberg hat einen videobasierten SJT zur Selbsteinschätzung entwickelt [19] und die Universität Hamburg hat kürzlich einen Papier-Bleistift-SJT (Hamburger Situational Judgement Test, HAM-SJT) für ihr Auswahlverfahren eingeführt [20]. Beide SJTs verwenden ein geschlossenes Antwortformat und unseres Wissens nach wurde ein SJT mit einem offenen Format noch nicht in einem medizinischen Auswahlverfahren in Deutschland getestet.

1.2. Ziele der Studie

In dieser Studie haben wir Casper als Online-SJT mit einem offenen Antwortformat pilotiert, welcher potentiell zukünftig in Deutschland als Auswahltests eingesetzt werden könnte. Unser Ziel war die Analyse der Akzeptanz, Reliabilität, Leistungsunterschiede zwischen Subgruppen sowie konvergenten (d.h. Beziehung zu anderen Instrumenten zur Messung persönlicher Eigenschaften) und diskriminanten (d.h. Beziehung zu kognitiven Zulassungskriterien) Validität im Vergleich zu bisherigen internationalen Forschungsergebnissen zu Casper.

2. Methoden

2.1. Vorgehen

Die Studie fand an fünf Testterminen im Sommer 2020 und Sommer 2021 statt. Studienbewerber*innen wurden eingeladen, sich für einen der Testtermine zu registrieren, wenn sie sich für einen der großen deutschen Zulassungstests für medizinische Studiengänge (Test für medizinische Studiengänge (TMS), Hamburger Naturwissenschaftstest (HAM-Nat), Hamburger Situational Judgement Test (HAM-SJT), siehe Tabelle 1) angemeldet und ihr Interesse an der Teilnahme von Forschungsstudien zur Studierendenauswahl angegeben hatten. Darüber hinaus wurden alle Medizinstudierenden der Universität Hamburg, unabhängig vom Studienjahr, über einen elektronischen Studierenden-Newsletter zur Teilnahme an der Studie eingeladen. Um Anreize für die Studienteilnahme zu schaffen, erhielten alle Teilnehmer*innen Feedback zu ihrer Casper-Leistung und hatten die Chance, 50 €-Gutscheine für einen Online-Shop zu gewinnen. Testgebühren wurden in dieser Studie nicht erhoben, können aber basierend auf den aktuellen Preisen (2024) in Nordamerika grob auf 46 bis 95 EUR geschätzt werden.

2.2. Casper

Casper konzentriert sich auf die Beurteilung interindividueller Unterschiede in zehn persönlichen Eigenschaften,

Tabelle 1: Übersicht der Instrumente und ihrer Reliabilität, deskriptive Statistiken in der Casper-Stichprobe und Korrelation zwischen Casper-Score und jedem der Instrumente

Instrument	Beschreibung	Skala	Reliabilität	*n*	*M*	*SD*	r_{Casper}
Casper	Casper-Leistung zum letzten Testdatum	z-standardisierter Mittelwert über 12 Szenarien	$\alpha_{2020} = 0.73$ $\alpha_{2021} = 0.82$	582	-0,02	0.99	1
Abitur	Selbstberichtete Abiturnote, vergleichbar mit dem grade point average (GPA)	Skala von 1 (höchste Leistung) bis 6 (schlechteste Leistung)	NA	354	1.72	0.44	-.15**
TMS	Test für medizinische Studiengänge: Fachspezifischer Zulassungstest für Medizin und andere gesundheitswissenschaftliche Studiengänge	standardisierter Wert mit einem Mittelwert von 100 und einer Standardabweichung von 15	.91≤ α ≤ .92 [49], [50]	371	103.1	9.52	.18**
HAM-Nat	Hamburger Naturwissenschaftstest: Multiple-choice Test mit drei Untertests	Fähigkeitsparameter basierend auf Item Response Theory (IRT) Modell					
	1) Naturwissenschaften (60 Items)		.88 ≤ α^1 ≤ .91	270	0.57	1.22	.04
	2) arithemtisches Problemlösen (15 Items)		.36 ≤ α^1 ≤ .55	270	0.02	0.59	.08
	3) relationales Schließen (15 Items)		.40 ≤ α^1 ≤ .62	270	0.14	0.60	.23**
HAM-SJT	Hamburger Situational Judgement Test: Kontextualisierter Papier-Bleistift-SJT mit einer „Should-Do-Instruktion“ und Bewertung der Effektivität (75 bis 80 Items)	mittlere quadrierte Abweichungen zwischen intraindividuell z-standardisierten Bewerber*innenantworten und mittlerer intraindividuell z-standardisierter Expert*innenpanel-Antwort für jedes Item, multipliziert mit -1 (d. h. höhere Werte = bessere Leistung)	.71 ≤ α^1 ≤ .81	263	-0.41	0.14	.18**
OSCE	Objektive strukturierte klinische Prüfung nach eineinhalb Jahren Medizinstudium mit 12 Stationen (10 während der COVID-Pandemie), darunter:	Prozent erreichter Punkte	α = .75				
	1) Anamnesestation			94	79.18	9.31	-.09
	2) Kommunikationsfähigkeitsstation			55	81.45	10.83	.08

NA=nicht verfügbar, *n*=Stichprobengröße, *M*=Mittelwert, *SD*=Standardabweichung, r_{Casper}=Pearson Korrelation mit Casper
*p<.05, **p<.01, [1]Range über verschiedene Testversionen

darunter Zusammenarbeit, Kommunikation, Empathie, Gerechtigkeit, Ethik, Motivation, Problemlösung, Professionalität, Resilienz und Selbstbewusstsein. Jedes Szenario ist in der Regel darauf ausgelegt, mehr als ein Merkmal zu messen, und bei der Zusammenstellung verschiedener Szenarien ist für jede*n Teilnehmer*in sichergestellt, dass alle zehn Merkmale abgedeckt werden. Im Einklang mit den Erkenntnissen, dass diese Eigenschaften innerhalb von SJTs nicht zuverlässig unterschieden werden können [21], [22], liefert Casper nur eine Gesamtbewertung.

In dieser Studie bestand der Test aus acht Video- und vier Textszenarien. Auf jedes Szenario folgten drei Fragen und die Teilnehmer*innen wurden gebeten, ihre Antworten in einem offenen Textformat innerhalb von 5 Minuten pro Szenario einzugeben. Aus einem bestehenden Pool wurden englischsprachige Szenarien ausgewählt, von denen sechs sowohl im Jahr 2020 als auch im Jahr 2021 verwendet wurden, während die anderen sechs Szenarien zwischen den Jahren variierten, um eine breitere Vielfalt an Szenarien abzudecken. Die Videodialoge und Fragen wurden vom deutschen Forschungsteam ins Deutsche übersetzt: Eine Linguistin und Gesundheitswissenschaftlerin, die fließend Englisch spricht, verfasste die Transkripte der Videodialoge, die dann von einer deutschsprachigen Psychologin ins Deutsche übersetzt wurden. Diese Übersetzung wurde von einer dritten Person (deutschsprachige Psychologin) überprüft. Unstimmigkeiten wurden im Team besprochen und gelöst. Die Videos wurden entweder mit Untertiteln (2020) oder mit einem Voice-Over versehen (2021). Die Teilnehmer*innen absolvierten den Test über die Online-Plattform von Casper. Englischsprachige Beispiele typischer Casper-Szenarien und Fragen sind über die offizielle Website verfügbar [https://acuityinsights.app/test-prep-casper/].

Im Jahr 2020 bewerteten 52 Fakultätsangehörige und studentische Hilfskräfte verschiedener deutscher Hochschulen die Antworten der Teilnehmer*innen. Davon beteiligten sich 15 Personen im darauffolgenden Jahr erneut bei der Bewertung. Im Einklang mit Strategien zur Erhöhung der Chancengerechtigkeit in der Hochschulzulassung wird empfohlen, Beurteiler*innen einzubeziehen, die die Vielfalt der Patient*innen widerspiegeln und die Inklusivität in beurteilerbasierten Auswahlinstrumenten in der Medizin fördern [23], [24], [25], um Voreingenommenheit zu reduzieren und die Fairness durch die Berücksichtigung unterschiedlicher Perspektiven und Hintergründe im Bewertungsprozess zu erhöhen. Mit dem Ziel, den Beurteiler*innenpool für die Studie 2021 zu diversifizieren, haben wir daher 11 zusätzliche Beurteiler*innen aus der Allgemeinbevölkerung über Online-Plattformen für befristete Stellenangebote und E-Mail-Listen von Vereinen für Menschen mit Migrationshintergrund rekrutiert. Alle Beurteiler*innen absolvierten eine Online-On-Demand-Schulung, die in englischer (2020) oder deutscher (2021) Sprache angeboten wurde. Im Durchschnitt benötigten die Beurteiler*innen 46,19 Sekunden (SD=22,72) für die Bewertung einer Antwort mit einer durchschnittlichen Anzahl von 125,60 Wörtern (SD=38,05). Die Beurteiler*innen der Fakultät gaben ihre Bewertungen innerhalb ihrer Arbeitszeit ab, während die Beurteiler*innen der Allgemeinbevölkerung einen Gutschein für einen Online-Shop als Aufwandsentschädigung erhielten (0,50 EUR pro bewerteter Antwort). Nach Abschluss ihrer Bewertungen wurden die Beurteiler*innen im Jahr 2021 gebeten, im Rahmen einer freiwilligen Umfrage soziodemografische Daten anzugeben.

Jede Antwort auf ein Szenario wurde von einem*r (2020) oder zwei Beurteiler*innen (2021) auf einer 9-stufigen globalen Bewertungsskala von 1=„schlecht“ bis 9=„ausgezeichnet“ ohne spezifische Verhaltensanker bewertet. Für jedes Szenario erhielten die Beurteiler*innen einen Leitfaden, wie sie die spezifischen Konstrukte, die das Szenario messen sollte, in ihren Bewertungen berücksichtigen sollten. Die Instruktion lautete, die Qualität jeder Antwort im Vergleich zu den entsprechenden Antworten anderer Teilnehmer*innen zu bewerten. Den Beurteiler*innen wurden die Textantworten über eine Online-Bewertungsplattform zugewiesen. Nach einer bestimmten Anzahl an Bewertungen konnten sie auf ein neues Szenario umsteigen, um Ermüdungserscheinungen vorzubeugen. Über einen Algorithmus der Online-Plattform wurde für jede*n Teilnehmer*in sichergestellt, dass jedes Szenario von einem*r anderen Beurteiler*in bewertet wurde. Im Falle von zwei Beurteiler*innen wurden beide Bewertungen zur Berechnung eines Szenariowertes gemittelt. Der vom Anbieter übermittelte Casper-Gesamtwert ist der innerhalb einer Kohorte z-standardisierte Mittelwert über zwölf Szenarien.

2.3. Weitere Variablen

Alle Studienteilnehmer*innen hatten sich zuvor bereit erklärt, an einem laufenden Forschungsprojekt (Studierendenauswahlverbund, „stav“, [https://www.projekt-stav.de/index.php] teilzunehmen, in dem Zulassungsdaten, Studienleistungsdaten der zugelassenen Studierenden sowie Daten aus anderen Forschungsstudien und einem soziodemografischen Fragebogen (siehe Anhang 1) abgeglichen und in einer zentralen Datenbank gespeichert werden. Casper-Daten konnten dadurch mit den folgenden verfügbaren Datenquellen verknüpft werden. Eine Übersicht aller Instrumente befindet sich zudem in Tabelle 1.

2.3.1. Akzeptanz

Nach Abschluss des Casper-Tests wurden die Teilnehmer*innen zu einer Online-Umfrage über ihre Testerfahrung weitergeleitet. Zusätzlich zu einer Gesamteinschätzung von Casper auf einer 10-Punkte-Skala wurden die Teilnehmer*innen beispielsweise gebeten, ihre Wahrnehmung der Fairness und Schwierigkeit von Casper auf einer 7-Punkte-Skala anzugeben (je höher die Einschätzung, desto positiver; siehe Anhang 2). Daten aus dieser Umfrage waren nur für die Testtermine in 2020 verfügbar.

2.3.2. Soziodemographische Merkmale

Um diese Studie mit früheren Ergebnissen zu Subgruppenunterschieden bei SJTs vergleichen zu können [17], [26], [27], haben wir das Geschlecht, den höchsten Bildungsabschluss der Eltern (d. h. mindestens einer der Eltern hat einen akademischen Abschluss) als Indikator für den sozioökonomischen Status (SES) sowie Migrationshintergrund als Indikator für Ethnizität/Nationalität einbezogen. In Anlehnung an die Definition des Statistischen Bundesamtes [28] wurde von einem Migrationshintergrund ausgegangen, wenn mindestens eine der folgenden Bedingungen zutraf: die Person wurde nicht in Deutschland geboren, hatte eine nichtdeutsche Staatsangehörigkeit oder ein Elternteil wurde nicht in Deutschland geboren.

2.3.3. Validität

Um die konvergente Validität zu untersuchen, wurden zwei zusätzliche Messinstrumente einbezogen: der HAM-SJT sowie die Kommunikationsleistung in einer objektiven strukturierten klinischen Prüfung (OSCE). Der HAM-SJT ist ein Papier-Bleistift-SJT mit geschlossenem Antwortformat, der seit 2020 im Zulassungsverfahren für das Medizinstudium an der Universität Hamburg eingesetzt wird [20], [29]. Studierende der Universität Hamburg legen üblicherweise ihre erste OSCE-Prüfung nach eineinhalb Jahren Studium ab. Die Prüfung besteht aus mehreren kurzen standardisierten Interaktionen (Stationen), die von Beurteiler*innen bewertet werden [30]. Da Medizinstudierende aller Kohorten zur Studienteilnahme eingeladen wurden, absolvierten die Teilnehmer*innen ihre OSCE-Prüfung zwischen 2016 und 2022. Zwischen diesen Jahren waren die zwölf Stationen der OSCE-Prüfung hinsichtlich des Inhaltes und der Bewertungschecklisten vergleichbar. Wir haben die Ergebnisse (in Prozent) von zwei Stationen mit simulierten Patienten genutzt, die speziell auf Kommunikationsfähigkeiten abzielten (Kommunikationsstation, Anamnesestation) [31]. Daten zur Kommunikationsstation waren nur für Studierende verfügbar, die vor dem Sommer 2020 an der OSCE-Prüfung teilgenommen haben, da diese Station während der COVID-19-Pandemie nicht stattfinden konnte.
Für die Analyse der diskriminanten Validität haben wir die Casper-Ergebnisse mit kognitiven Zulassungskriterien verglichen, einschließlich der Abiturnote (entspricht dem Notendurchschnitt beim Schulabschluss), der Leistung beim Zulassungstest HAM-Nat, einem Multiple-Choice-Test mit Untertests zu Naturwissenschaften [32], arithmetischem Problemlösen und relationalem Schließen sowie der Leistung beim Test für medizinische Studiengänge (TMS), einem fachspezifischen Studierfähigkeitstest für Medizin und andere Gesundheitsstudiengänge [33].

2.4. Datenanalyse

Alle Analysen wurden in R-4.2.1 durchgeführt [https://www.r-project.org/]. Für die Analyse der Antworten der Teilnehmer*innen in der Umfrage zur Akzeptanz haben wir für quantitative Fragen deskriptive Statistiken berechnet und in offenen Fragen die Häufigkeit relevanter Themen mittels MAXQDA 2022 gezählt [https://www.maxqda.com/de/]. Die Reliabilität von Casper wurde anhand der internen Konsistenz über 12 Szenarien (Cronbachs Alpha) analysiert. Für Antworten, die von zwei unabhängigen Beurteiler*innen bewertet wurden (Stichprobe 2021), haben wir die Interrater-Übereinstimmung mittels Intraklassenkorrelation (ICC(1,2)) analysiert. Mittelwertsunterschiede in der Leistung zwischen einzelnen Subgruppen haben wir mit Welch-*t*-Tests für unabhängige Stichproben untersucht. Die Effektstärken wurden als Cohens d angegeben. Die konvergente und diskriminante Validität wurde mithilfe von Pearson-Korrelationen analysiert.
Analysen der Subgruppenunterschiede und der Reliabilität basieren auf der Gesamtstichprobe. Für Fälle, in denen Teilnehmer*innen in beiden Jahren teilnahmen, wurde der z-Wert des neueren Casper-Testdatums (2021) verwendet. Um sicherzustellen, dass die Leistung in den untersuchten Variablen zwischen den Studienkohorten vergleichbar war, wurden ungepaarte Welch-t-Tests und Mann-Whitney-U-Tests durchgeführt. Das Signifikanzniveau für alle Analysen betrug α=.05. Der R-Code, ein vollständiger Datenanalysebericht, alle Anhänge und Informationen zum Anfordern der Originaldaten sind unter [https://osf.io/9daz3/] einsehbar.

3. Ergebnisse

3.1. Teilnehmer*innen und Beurteiler*innen

Insgesamt nahmen 582 Personen an dieser Pilotstudie teil, darunter 74 Medizinstudierende und 508 Bewerber*innen. Zwanzig Teilnehmer*innen nahmen sowohl 2020 als auch 2021 an Casper teil. Das Durchschnittsalter der Teilnehmer*innen betrug 21 Jahre (*SD*=3,30). Für rund 64% der Teilnehmer*innen lagen weitere soziodemografische Informationen vor. In dieser Teilstichprobe identifizierten sich 19% als männlich, 36% hatten einen Migrationshintergrund und 71% hatten mindestens einen Elternteil mit einem Hochschulabschluss (siehe Tabelle 2). Das Alter, die Casper-Leistung und andere Studienvariablen waren zwischen den Studienkohorten weitgehend vergleichbar (siehe Anhang 3, S. 1-2). Lediglich die HAM-SJT-Leistung war in der Kohorte 2021 im Vergleich zur Kohorte 2020 signifikant besser (W=3773,5, p<.001, d=0,62). Bewerber*innen und Medizinstudierende unterschieden sich nicht in ihrer durchschnittlichen Casper-Leistung (t(91,226)=-1,16, p=0,25, d=0,16). Die durchschnittliche Leistung in sechs Videoszenarien, die sowohl

Tabelle 2: Merkmale der Stichprobe

Stichprobenbeschreibung		**Casper Leistung**
N	**582**	
Alter bei Casper-Teilnahme		r = 0.09, p = 0.03
M	21.26	
SD	3.31	
Geschlecht		d = 0.37
NA (Prozent an N)	186 (32%)	
n männlich (Prozent männlich an verfügbaren Daten)	77 (19%)	M = -0.36, SD = 1.08
n weiblich (Prozent weiblich an verfügbaren Daten)	319 (81%)	M = 0.01, SD = 0.96
Migrationshintergrund		d = 0.40
NA (Prozent an N)	195 (34%)	
n ja (Prozent ja an verfügbaren Daten)	141 (36%)	M = -0.31, SD = 1.05
n nein (Prozent nein an verfügbaren Daten)	246 (64%)	M = 0.08, SD = 0.93
Höchster Bildungsabschluss der Eltern		d = 0.15
NA (Prozent an N)	209 (36%)	
n Hochschulabschluss (Prozent Hochschulabschluss an verfügbaren Daten)	264 (71%)	M = -0.07, SD = 0.99
n kein Hochschulabschluss (Prozent kein Hochschulabschluss an verfügbaren Daten)	109 (29 %)	M = 0.07, SD = 0.98

N=Anzahl Teilnehmer*innen in Gesamtstichprobe, *n*=Anzahl Teilnehmer*innen in Teilstichprobe, *M*=Mittelwert, *SD*=Standardabweichung, *r*=Pearson Korrelation, *d*=Cohen's d, *NA*=nicht verfügbar

im Jahr 2020 (Untertitel) als auch im Jahr 2021 (Voice-Over) verwendet wurden, unterschied sich zwischen den Jahren nicht (t(465,16)=-0,48, p=0,63, d=0,04).
Von den 26 Beurteiler*innen im Jahr 2021 stellten 15 der Fakultätsangehörigen und 6 der Beurteiler*innen aus der Allgemeinbevölkerung demografische Daten zur Verfügung (siehe Tabelle 3). Hierbei war am auffälligsten, dass die Beurteiler*innen aus der Allgemeinbevölkerung im Vergleich zu Beurteiler*innen der Fakultät einen vielfältigeren Bildungshintergrund hatten (33% gegenüber 83% mit Universitätsabschluss).

3.2. Akzeptanz

Insgesamt beurteilten die Teilnehmer*innen aus 2020 Casper positiv mit einer durchschnittlichen Bewertung von 7,55 (SD=1,64, n=368) auf einer 10-Punkte-Skala. Auf einer 7-Punkte-Skala gaben die Teilnehmer*innen an, dass sie mit ihrer allgemeinen Testerfahrung zufrieden waren (M=5,40, SD=1,19, n=367) und Casper als eher fair empfanden (M=5,24, SD=1,26, n=354). Die Teilnehmer*innen bewerteten Casper im Allgemeinen als etwas weniger stressig im Vergleich zu anderen Prüfungen (M=3,24, SD=1,50, n=359) und empfanden den Test weder als schwierig noch als einfach (M=4,08, SD=1,21, n=356). Bei den Fragen im offenen Textformat war der am häufigsten kritisierte Aspekt der Testdurchführung die kurze Antwortzeit, die bei einigen Teilnehmern*innen den Eindruck erweckte, dass Bewerber*innen mit geringerer Tipperfahrung systematisch benachteiligt werden könnten (n=24) (vollständige Ergebnisse siehe Anhang 2).

3.3. Reliabilität

Die interne Konsistenz für die Casper-Szenarien betrug im Jahr 2020 α=0,73, 95%-KI [0,69, 0,77] und im Jahr 2021 α=0,82, 95%-KI [0,79, 0,86]. Für Antworten, die im Jahr 2021 von zwei Beurteiler*innen bewertet wurden, betrug die Interrater-Übereinstimmung ICC(1,2)=0,54. Die Retest-Reliabilität für zwanzig Teilnehmer*innen, die Casper in beiden Jahren absolviert hatten, betrug ρ=0,29 (Rangkorrelation nach Spearman).

3.4. Subgruppenunterschiede

Gruppenvergleiche ergaben, dass weibliche Teilnehmerinnen (t(107,16)=2,73, p=0,01, d=0,37) und Teilnehmer*innen ohne Migrationshintergrund (t(263,09)=3,65, p<.001, d=0,40) im Mittel eine bessere Casper-Leistung hatten im Vergleich zu männlichen Teilnehmern bzw. Teilnehmer*innen mit Migrationshintergrund. Die Casper-Leistung unterschied sich nicht signifikant je nach Bildungsniveau der Eltern (t(203,67)=1,30, p=0,19, d=0,15). Nachfolgende Regressionsanalysen mit der Casper-Leistung als abhängige Variable ergaben, dass die Hinzunahme von Muttersprache als Prädiktor den Effekt des Migrationshintergrunds erklärte, während Ge-

Tabelle 3: Merkmale der Beurteiler*innen der Allgemeinbevölkerung und Fakultät in 2021

	Beurteiler*innen der Allgemeinbevölkerung	Beurteiler*innen der Fakultät
N	11	15
n Umfrage (Prozent an *N*)	6 (55%)	12 (80%)
Geschlecht		
männlich (Prozent)	0 (0%)	3 (25%)
weiblich	6	9
Alter (in Jahren)		
18 – 30 (Prozent)	3 (50%)	6 (50%)
31 – 40	1	3
41 – 50	1	3
51 – 60	1	0
Höchster Bildungsabschluss		
Hochschulabschluss* (Prozent)	2 (33%)	10 (83%)
Berufsausbildung	2	0
Weiterführende Schule: Abitur	1	2**
Weiterführende Schule: Mittlere Reife	1	0

*beinhaltet: Bachelor, Master, Diplom, Magister, PhD, **derzeit im Medizin- oder Psychologiestudium

Tabelle 4: Multiple Regressionsanalysen zur Vorhersage von Casper anhand soziodemografischer Variablen (Modell 1), unter Berücksichtigung der Muttersprache (Modell 2) und der kognitiven Fähigkeiten (Modell 3) (*n*=227)

	Modell 1			Modell 2			Modell 3		
	b	***SE***	***p***	***b***	***SE***	***p***	***b***	***SE***	***p***
Männliches Geschlecht	-0.33	0.16	.042	-0.36	0.16	.027	-0.37	0.16	.018
Migrationshintergrund	-0.40	0.14	.004	-0.18	0.16	.271	-0.08	0.16	.614
Deutsch als Muttersprache				0.59	0.23	.012	0.54	0.23	.019
Abiturnote							0.02	0.16	.899
TMS							0.02	0.01	.003
R^2 / Δ R^2 (*p*)	0.06			0.09 / 0.03 (.010)			0.13 / 0.04 (.005)		

b = unstandardisiertes Regressionsgewicht, *SE* = Standardfehler

schlecht und Sprache bei der Kontrolle kognitiver Kriterien weiterhin signifikante Prädiktoren blieben (siehe Tabelle 4).

3.5. Konvergente und diskriminante Validität

In Hinblick auf andere Messinstrumente persönlicher Eigenschaften hatte Casper einen signifikanten Zusammenhang mit der HAM-SJT-Leistung (*r*=.18, *p*=.004, *n*=263), jedoch keinen Zusammenhang mit der Leistung an der OSCE-Anamnesestation (*r*=-.09, *p*=.37, *n*=94) oder der Kommunikationsstation (*r*=.08, *p*=.57, *n*=55).

In Bezug auf kognitive Zulassungskriterien wies die Casper-Leistung signifikante Korrelationen auf mit der Abiturnote (*r*=-.15, *p*=.01, *n*=354; d. h. je besser die Abiturnote, desto besser die Casper-Leistung), der Leistung im TMS (*r*=.18, *p*=.001, *n*=371) und dem Untertest zum relationalen Schließen im HAM-Nat (*r*=.23, *p*<.001, *n*=270). Andererseits korrelierte er weder mit der HAM-Nat Untertest zur Naturwissenschaft (*r*=0,04, *p*=0,46, *n*=270) noch mit dem Untertest zum arithmetischen Problemlösen (*r*=0,08, *p*=0,18, *n*=270) (siehe Tabelle 1). Anhang 3 enthält eine vollständige Korrelationstabelle für alle Studienvariablen.

4. Diskussion

In der deutschen medizinischen Ausbildung wurden textbasierte und videobasierte SJTs bereits für die (Selbst-) Beurteilung, Vermittlung und Nachverfolgung relevanter Fähigkeiten wie Kommunikation oder professionellem Verhalten von Studienbewerber*innen und Studierenden der Medizin entwickelt und vorgeschlagen [19], [20], [34], [35], [36]. Während alle diese Beispiele auf einem geschlossenen Antwortformat basieren, ist dies die erste Studie, die einen Online-SJT mit offenem Antwortformat in einem deutschen Zulassungskontext erprobt.

Ähnlich wie in kanadischen Berichten über Casper [16] war die Wahrnehmung von Casper durch die Teilnehmer*innen positiv und die interne Konsistenz gut. Diese Ergebnisse stimmen auch mit positiven Wahrnehmungen sowie zufriedenstellenden internen Konsistenzwerten für

den Heidelberger Video-SJT (0,81≤α≤0,83) [19] und den HAM-SJT (0,62≤α≤0,82) [37] überein. Andererseits war die Interrater-Übereinstimmung in unserer Studie nur mäßig und weicht von der hohen Übereinstimmung (0,95) ab, die in der kanadischen Pilotstudie von Casper gefunden wurde [15]. In der kleinen Teilstichprobe von Teilnehmer*innen, die den Test zweimal absolvierten, war die Retest-Reliabilität gering. Dies könnte durch individuelle Unterschiede in der persönlichen Entwicklung der Teilnehmer*innen innerhalb der einjährigen Zeitspanne zwischen den beiden Testterminen erklärt werden, aber auch durch Änderungen im Format zwischen beiden Testungen (z. B. Verwendung unterschiedlicher Szenarien, Voice-Over, Einbeziehung der Beurteiler*innen aus der Allgemeinbevölkerung). Dennoch war die Teilstichprobe in unserer Studie zu klein (n=20), um eindeutige Schlussfolgerungen zu ziehen, und eine Folgestudie mit einem gezielten Test-Retest-Design wäre erforderlich.

Unsere Studie ergab signifikante Leistungsunterschiede zugunsten von Frauen und Teilnehmer*innen ohne Migrationshintergrund, die mit einer nordamerikanischen Studie zu Casper übereinstimmen [17]. Unsere Folgeanalysen deuten darauf hin, dass eher die Muttersprache als der Migrationshintergrund mit Leistungsunterschieden zusammenhängt, was von den Ergebnissen einer US-Studie abweicht, in der Unterschiede in Abhängigkeit von der ethnischen Zugehörigkeit nach Kontrolle der Sprache bestehen blieben [38]. Das offene Antwortformat bot daher keinen Vorteil gegenüber dem HAM-SJT, der ebenfalls Leistungsunterschiede je nach Muttersprache aufwies (d=0,24) [37] oder dem Heidelberger Video-SJT, der keine signifikanten Unterschiede aufwies [19].

Die konvergente und diskriminante Validität des Tests wird gestützt durch den Zusammenhang von Casper mit der HAM-SJT-Leistung und fehlenden Zusammenhängen mit den HAM-Nat-Untertests zu Naturwissenschaften und arithmetischem Problemlösen. Gleichermaßen wurde kein Zusammenhang zwischen dem kanadischen Casper und dem MCAT-Naturwissenschaftsteil festgestellt [15]. Andererseits fanden wir schwache Korrelationen mit der Abiturnote, der TMS-Leistung und dem HAM-Nat-Untertest zum relationalen Schließen. Die schwachen Reliabilitätswerte der HAM-Nat-Untertests zum relationalen Schließen und arithmetischem Problemlösen könnten die Signifikanz und das Ausmaß der Korrelation mit Casper beeinflusst haben. Allerdings gab es eine kleine signifikante Korrelation zwischen TMS und Casper, die in eine ähnliche Richtung weist, und die Ergebnisse stimmen auch mit den Erkenntnissen überein, dass Casper mit dem Untertest zum verbalen Schlussfolgern des MCAT korreliert [15]. Dies deutet darauf hin, dass kognitive, aber auch nicht-kognitive Kompetenzen, die in diesen Auswahlkriterien erfasst werden (wie z.B. Motivation, Flexibilität oder Selbstmanagement in Abiturnoten [39]), vorteilhaft für die Leistung bei Casper sein könnten. Die Ergebnisse deuten auch auf eine etwas höhere kognitive Belastung bei Casper hin im Vergleich zum HAM-SJT oder Heidelberger Video-SJT, welche entweder negativ oder gar nicht mit der Abiturnote, TMS und HAM-Nat zusammenhingen [19], [20].

Schließlich konnten wir keinen Zusammenhang zwischen Casper und zwei OSCE-Stationen zur Kommunikationsfähigkeit feststellen. Somit konnten wir die positiven Ergebnisse zur prädiktiven Validität von Casper aus Nordamerika nicht replizieren. Hier gab es sowohl Zusammenhänge zwischen Casper und der MMI-Leistung sowie mit nationalen Approbationsprüfungen [15], [18]. HAM-SJT-Pilotstudien konnten dagegen kleine, aber signifikante Korrelationen mit den späteren Leistungen in einem MMI (r=0,22) [20] und OSCE (r=0.20) [37] nachweisen.

Limitationen

Wir haben während der Schulung der Beurteiler*innen und des Bewertungsprozesses verschiedene Maßnahmen zur Qualitätssicherung angewendet. Dazu zählen wiederholte Schulungsrunden, wenn die Statistiken von Probebewertungen nicht vorab festgelegten Richtwerten entsprechen oder die vorübergehenden Sperrung von Beurteiler*innen, wenn sie ihre Bewertung in einer kürzeren Zeitspanne abgeben als die erforderliche Lesezeit für die Antwort eines Bewerbers. Allerdings wurden diese Maßnahmen in dieser Pilotstudie nicht im gleichen Umfang umgesetzt wie bei einer Casper-Testung in einem tatsächlichen Auswahlsetting. Die in dieser Studie beobachtete moderate Interrater-Übereinstimmung unterstreicht, wie wichtig es ist, den Ratingprozess kontinuierlich zu überwachen und den Beurteiler*innen eine Rückmeldung zu geben.

Im Jahr 2021 haben wir zusätzlich Beurteiler*innen aus der Allgemeinbevölkerung rekrutiert mit dem Ziel, den Pool der Beurteiler*innen diverser zusammenzusetzen. Obwohl demografische Daten teilweise darauf hindeuten, dass sich die Beurteiler*innen aus der Allgemeinbevölkerung von Fakultätsangehörigen hinsichtlich ihres Bildungsniveaus unterschieden, ist es aufgrund der geringeren Beteiligungsquote von Beurteiler*innen der Allgemeinbevölkerung an der Folgebefragung (55%) schwierig, eindeutige Schlussfolgerungen über die Vielfalt unserer Beurteiler*innen zu ziehen. Zukünftige Studien zu urteilsbasierten Auswahlinstrumenten würden von einer systematischen Erhebung und Variation der soziodemografischen Merkmale der Beurteiler*innen profitieren, um untersuchen zu können, wie sich unterschiedliche Hintergründe der Beurteiler*innen auf die Ergebnisse im Auswahlverfahren auswirken.

Für diese Pilotstudie haben wir Szenarien verwendet, die zuvor in einem nordamerikanischen Auswahlkontext entwickelt und getestet wurden. Es bleibt jedoch unklar, ob kulturelle Unterschiede im Zusammenhang mit den Inhalten der Szenarien einen Einfluss auf die Studienergebnisse hatten. Darüber hinaus handelte es sich bei den Teilnehmer*innen unserer Studie um Freiwillige, die sich in ihrer Leistungsmotivation wahrscheinlich zu Teilnehmer*innen in einem tatsächlichen Auswahlkontext unterscheiden. Schließlich haben wir zu dieser Studie nur Bewerber*innen eingeladen, die sich für den TMS

und/oder HAM-Nat angemeldet haben und deren Ziel darin bestand, ihre Chancen auf einen Studienplatz zu verbessern. Unsere Stichprobe ist daher nicht repräsentativ für die Grundgesamtheit aller an einem Medizinstudium Interessierten und schließt vermutlich Bewerber*innen mit einer guten Abiturnote aus sowie Personen, die durch das aktuelle Auswahlsystem entmutigt sind und sich nicht bewerben. Allerdings könnte die letztgenannte Gruppe möglicherweise von einem nicht-kognitiven Test wie Casper profitieren. Für zukünftige Testungen wird empfohlen, die Testinhalte innerhalb der Kultur und Sprache zu entwickeln, in denen der Test durchgeführt werden soll, und die psychometrischen Eigenschaften im Rahmen eines tatsächlichen Auswahlverfahrens zu bestätigen.

Implikationen für Praxis und Forschung

Eine aktuelle Studie ergab, dass Ärzt*innen und Medizinstudierende in Hamburg insbesondere hinsichtlich ihres sozioökonomischen und ethnischen Hintergrunds nicht die Gesamtbevölkerung repräsentieren [40]. Medizinische Fakultäten, die eine Strategie der Chancengerechtigkeit verfolgen, müssen bei der Zusammenstellung und Gewichtung ihrer Auswahlkriterien auch darauf achten, wie unterrepräsentierte Gruppen in einem Auswahlkriterium abschneiden, um nachteilige Auswirkungen zu minimieren. Die Leistung unserer Studienteilnehmer*innen unterschied sich nicht in Abhängigkeit vom sozioökonomischen Hintergrund. Als Indikator konnten wir jedoch nur den Bildungsstand der Eltern heranziehen. Die Verwendung zusätzlicher Indikatoren wie etwa des Einkommens der Eltern oder der Lebensumstände [40] könnte in zukünftigen Studien ein umfassenderes Bild liefern. Obwohl unsere Ergebnisse auf einen potenziellen Nachteil für Bewerber*innen schließen lassen, deren Muttersprache nicht Deutsch ist, wird international argumentiert, dass SJTs wie Casper die oft schwerwiegenderen Subgruppenunterschiede in kognitiven Tests abmildern und dadurch möglicherweise den Zugang zum Medizinstudium erweitern können [17], [27]. Während vorläufige Daten zum HAM-Nat darauf hindeuten, dass Bewerber*innen ohne Migrationshintergrund in den beiden Untertests zum logischen Denken besser abschneiden ($0{,}24 \leq d \leq 0{,}32$) und Bewerber*innen mit einem höheren sozioökonomischen Hintergrund in allen drei Untertests des HAM-Nat bessere Leistungen zeigen ($0{,}06 \leq d \leq 0{,}25$) ist die Größe der Effekte gering [41]. Aktuell gibt es unseres Wissens nach keine vergleichbaren veröffentlichten Daten zum TMS. Große Bildungsstudien und Berichte weisen regelmäßig auf schwächere Sekundarschulleistungen [42], [43] und Abiturnoten bei Schüler*innen mit niedrigem sozioökonomischen Status hin (z. B. mittlere Abiturnote von 2,27 vs. 2,48 bei Schüler*innen vor dem Übergang zur Hochschule mit einem hohen vs. niedrigen sozioökonomischer Hintergrund [44]) und einem Migrationshintergrund (z. B. mittlere Abiturnote von 2,5 vs. 2,9 bei Schüler*innen mit deutschem vs. türkischem Hintergrund [45]). Nichtsdestotrotz ist das genaue statistische Ausmaß dieser Subgruppenunterschiede in den aktuellen Abiturnoten bei Interessierten an einem Medizinstudium unklar. Systematische Studien und Vergleiche der Subgruppenunterschiede für die deutschen Auswahlkriterien in Abhängigkeit von der ethnischen Zugehörigkeit und dem sozioökonomischen Hintergrund der Bewerber*innen sind daher erforderlich, um das Potenzial von SJTs zur Verbesserung oder Verringerung des Zugangs für diese Gruppen zu bewerten und Entscheidungsträger*innen mehr Informationen für ihre Auswahlstrategien bereitzustellen.

Da einige Teilnehmer*innen Bedenken äußerten, dass der 5-Minuten-Zeitrahmen Nicht-Muttersprachler*innen und Personen mit weniger Tipperfahrung potentiell benachteiligt, könnte eine Studie zur systematischen Variation der Zeitbegrenzung mehr Aufschluss darüber geben, ob eine solche Änderung das Potenzial hat, Leistungsunterschiede zu minimieren. Ein audio-visuelles Antwortformat, welches Subgruppenunterschiede weiter zu reduzieren scheint [13], wurde kürzlich für Casper eingeführt und könnte in Folgestudien auf das Potenzial für eine deutsche Testversion untersucht werden.

Medizinische Fakultäten in Deutschland sind aufgerufen, bei der Auswahl der Studierenden persönliche Eigenschaften zu berücksichtigen [4] und Auswahlkriterien zu verwenden, die die Eignung für das Medizinstudium und den Arztberuf belegen [46]. Aus diesem Grund ist es wichtig, die Konstrukt- und prädiktive Validität nachzuweisen. In unserer Studie korrelierte Casper in einem ähnlichen Ausmaß mit nicht-kognitiven und kognitiven Auswahlkriterien. Es scheint also, dass Casper nicht nur die persönlichen Eigenschaften erfasst, die wir messen wollten, sondern auch kognitive Fähigkeiten. Daher bleibt der Nutzen von Casper als sinnvolle Ergänzung zu den bestehenden Auswahlkriterien unklar. Wir konnten nur zwei OSCE-Stationen für eine kleine Teilstichprobe der Studienteilnehmer*innen einbeziehen. Die mangelnde Reliabilität einer einzelnen OSCE-Station [30] und die eingeschränkte Streuung bei den OSCE-Werten (d. h. die OSCE-Leistung der Studierenden lag zwischen 52,5% und 100% der erreichbaren Punkte) sind potenziell limitierende Faktoren in unserer Analyse. Zukünftige Forschung sollte darauf abzielen, verschiedene Ergebniskriterien zu persönlichen Eigenschaften zu untersuchen, wie beispielsweise Einschätzungen durch Lehrende und Mitstudierende oder eine Kombination relevanter OSCE-Stationen im Verlauf des Medizinstudiums [47]. Idealerweise sollten diese mit der prädiktiven Validität anderer Auswahlkriterien verglichen werden, die derzeit in Verbindung mit kognitiven Kriterien verwendet werden: eine abgeschlossenen Berufsausbildung sowie Berufserfahrung und ein Freiwilligendienst [8].

Aus praktischer Sicht müssen medizinische Fakultäten schließlich die Kosten eines Testformats wie Casper im Vergleich zu alternativen Auswahlinstrumenten abwägen und die Perspektiven verschiedener Interessengruppen berücksichtigen. Diese Studie konnte zeigen, dass Casper mit einer durchschnittlichen Bewertungszeit von 46 Sekunden pro Antwort weniger Beurteilungszeit benötigt als multiple Mini-Interviews mit einer Stationszeit von fünf

bis zehn Minuten [6] und als klassische Interviews, die im Hinblick auf die Arbeitsstunden noch weniger kosteneffizient sind [48]. Ebenso liegen die geschätzten Kosten mit maximal 95 EUR pro Bewerber*in (2024) deutlich unter den 450 EUR pro Bewerber*in (2014) im Hamburger multiplen Mini-Interview HAM-Int [7]. Wenn jedoch die Kosten durch Testgebühren gedeckt werden, wäre die Einführung von Casper mit einer zusätzlichen finanziellen Belastung für Bewerber*innen verbunden, die bereits für die Teilnahme am TMS (100 EUR im Jahr 2024) und HAM-Nat (95 EUR im Jahr 2024) zahlen. Eine Berufsausbildung hingegen bietet Bewerber*innen die Möglichkeit, relevante Fähigkeiten zu erlernen und ein Gehalt zu beziehen, erfordert aber auch, dass Bewerber*innen drei Jahre in ihre Ausbildung investieren, bevor sie ein Medizinstudium beginnen können.

5. Schlussfolgerungen

Positive Bewertungen der Testteilnehmer*innen, eine gute interne Konsistenz und Hinweise auf diskriminante und konvergente Validität in dieser Studie bestätigen, dass das in Casper verwendete Testformat auf einen deutschen Kontext anwendbar ist. Basierend auf der moderaten Interrater-Übereinstimmung in unserer Studie müssen die Anzahl, der Hintergrund und die Schulung der Beurteiler*innen berücksichtigt und sorgfältig überprüft werden, falls der Test zur tatsächlichen Auswahl eingesetzt wird. Die potentiell negativen Auswirkungen auf die Vielfalt der von Casper ausgewählten Studierenden und die derzeit fehlende Korrelation zur OSCE-Leistung erfordern mögliche Anpassungen des Tests sowie weitere Untersuchungen zur prädiktiven Validität von Casper unter Berücksichtigung eines breiteren Spektrums an Ergebniskriterien. Es ist wichtig sicherzustellen, dass der Testinhalt für die Testteilnehmer*innen nachvollziehbar und im Einklang mit den Zielen der deutschen medizinischen Ausbildung ist, damit der Test für die Medizinstudierendenauswahl in Deutschland geeignet ist. Im Hinblick auf Subgruppenunterschiede und Validität deuten unsere aktuellen Ergebnisse nicht darauf hin, dass ein SJT mit offenem Antwortformat wie Casper den verfügbaren deutschen SJTs mit einem geschlossenen Antwortformat überlegen ist.

Ethikvotum und Einverständniserklärung

Alle Teilnehmer*innen gaben ihr Einverständnis zur Erhebung, Speicherung und Verknüpfung der Daten. Diese Studie wurde im Rahmen des stav-Forschungsprojekts von der örtlichen Ethikkommission der Abteilung für Medizinische Psychologie des Universitätsklinikums Hamburg-Eppendorf (LPEK-0042) genehmigt. Alle Daten wurden im Einklang mit den europäischen Datenschutzgesetzen (DSGVO) verarbeitet.

Danksagungen

Die Autor*innen danken Dieter Münch-Harrach für die Erstellung der Untertitel für die Casper-Videos. Diese Studie wäre nicht möglich gewesen ohne die ehrenamtlichen Beurteiler*innen der stav-Teams in Hamburg, Heidelberg, Münster, Saarbrücken, Berlin und Göttingen sowie der Mitglieder*innen des Netzwerks Eignung & Auswahl Baden-Württemberg am Karlsruher Institut für Technologie, an der Universität Heidelberg, der DHBW Mannheim, der Pädagogische Hochschule Weingarten und der Hochschule Pforzheim.

Förderung

Diese Studie wurde im Rahmen des stav-Forschungsprojekts durchgeführt, das vom Bundesministerium für Bildung und Forschung gefördert wurde, Projektnummer: 01GK1801A-F.
Wir danken für die finanzielle Unterstützung durch den Open-Access-Publikationsfonds des UKE - Universitätsklinikum Hamburg-Eppendorf.

ORCIDs der Autor*innen

- Mirjana Knorr: [0000-0002-0996-9286]
- Ina Mielke: [0000-0003-1764-5553]
- Dorothee Amelung: [0000-0002-9946-9073]
- Mahla Safari: [0000-0003-0976-8094]
- Oana R. Gröne: [0000-0002-6829-5365]
- Simon M. Breil: [0000-0001-5583-3884]
- Alexander MacIntosh: [0000-0002-5094-3774]

Interessenkonflikt

Alexander MacIntosh ist Data Scientist bei Acuity Insights, dem Unternehmen, das Casper entwickelt und vertreibt. Die anderen Autor*innen haben keine zu erklärenden Interessenkonflikte.

Anhänge

Verfügbar unter https://doi.org/10.3205/zma001685

1. Anhang_1.pdf (160 KB)
   Soziodemografischer Fragebogen des stav-Projektes (Version 2019)
2. Anhang_2.pdf (203 KB)
   Evaluationsfragebogen zu CASPer
3. Anhang_3.pdf (241 KB)
   Zusätzliche Tabellen

Literatur

1. Frank JR, Snell L, Sherbino J, editors. Can Meds 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015. Zugänglich unter/available from: https://canmeds.royalcollege.ca/en/framework
2. Medizinischer Fakultätentag. Nationaler Kompetenzbasierter Lernzielkatalog Medizin 2015. Berlin: MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V.; 2015. Zugänglich unter/available from: https://medizinische-fakultaeten.de/wp-content/uploads/2021/06/nklm_final_2015-12-04.pdf
3. Association of American Medical Colleges. The Core Competencies for Entering Medical Students. Washington, DC: Association of American Medical Colleges; 2022. Zugänglich unter/available from: https://students-residents.aamc.org/applying-medical-school/article/core-competencies
4. Bundesministerium für Gesundheit. Masterplan Medizinstudium 2020. Berlin: Bundesgesundheitsministerium; 2017. Zugänglich unter/available from: https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/masterplan-medizinstudium-2020.html
5. Schult J, Hofmann A, Stegt SJ. Leisten fachspezifische Studierfähigkeitstests im deutschsprachigen Raum eine valide Studienerfolgsprognose? Z Entwicklungspsychol Pädagog Psychol. 2019;51(1):16-30. DOI: 10.1026/0049-8637/a000204
6. Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME Guide No. 37. Med Teach. 2016;38(5):443-455. DOI: 10.3109/0142159X.2016.1158799
7. Hissbach JC, Sehner S, Harendza S, Hampe W. Cutting costs of multiple mini-interviews - changes in reliability and efficiency of the Hamburg medical school admission test between two applications. BMC Med Educ. 2014;14:54. DOI: 10.1186/1472-6920-14-54
8. Amelung D, Zegota S, Espe L, Wittenberg T, Raupach T, Kadmon M. Considering vocational training as selection criterion for medical students: evidence for predictive validity. Adv Health Sci Educ Theory Pract. 2022;27(4):933-948. DOI: 10.1007/s10459-022-10120-y
9. Erschens R, Herrmann-Werner A, Schaffland TF, Kelava A, Ambiel D, Zipfel S, Loda T. Association of professional pre-qualifications, study success in medical school and the eligibility for becoming a physician: A scoping review. PLoS One. 2021;16(11):e0258941. DOI: 10.1371/journal.pone.0258941
10. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med Teach. 2016;38(1):3-17. DOI: 10.3109/0142159X.2015.1072619
11. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60. DOI: 10.1111/medu.12817
12. Webster ES, Paton LW, Crampton PES, Tiffin PA. Situational judgement test validity for selection: A systematic review and meta-analysis. Med Educ. 2020;54(10):888-902. DOI: 10.1111/medu.14201
13. Lievens F, Sackett PR, Dahlke JA, Oostrom JK, De Soete B. Constructed response formats and their effects on minority–majority differences and validity. J Appl Psychol. 2019;104(5):715-726. DOI: 10.1037/apl0000367
14. Mortaz Hejri S, Ho JL, Pan X, Park YS, Sam AH, Mangardich H, MacIntosh A. Validity of constructed-response situational judgment tests in training programs for the health professions: A systematic review and meta-analysis protocol. PLoS One. 2023;18(1):e0280493. DOI: 10.1371/journal.pone.0280493
15. Dore KL, Reiter HI, Eva KW, Krueger S, Scriven E, Siu E, Hilsden S, Thomas J, Norman GR. Extending the interview to all medical school candidates-computer-based multiple sample evaluation of noncognitive skills (CMSENS). Acad Med. 2009;84:S9-S12. DOI: 10.1097/ACM.0b013e3181b3705a
16. Zou C, McConnell M, Leddy J, Antonacci P, Lemay G. Comparison of the English and French versions of the CASPer® Test in a bilingual population, version 1. MedEdPublish. 2018;7:281. DOI: 10.15694/mep.2018.0000281.1
17. Juster FR, Baum RC, Zou C, Risucci D, Ly A, Reiter H, Miller DD, Dore KL. Addressing the diversity-validity dilemma using situational judgment tests. Acad Med. 2019;94(8):1197-1203. DOI: 10.1097/ACM.0000000000002769
18. Dore KL, Reiter HI, Kreuger S, Norman GR. CASPer, an online pre-interview screen for personal/professional characteristics: prediction of national licensure scores. Adv Health Sci Educ Theory Pract. 2017;22(2):327-336. DOI: 10.1007/s10459-016-9739-9
19. Fröhlich M, Kahmann J, Kadmon M. Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. Int J Sel Assess. 2017;25(1):94-110. DOI: 10.1111/ijsa.12163
20. Schwibbe A, Lackamp J, Knorr M, Hissbach J, Kadmon M, Hampe W. Medizinstudierendenauswahl in Deutschland: Messung kognitiver Fähigkeiten und psychosozialer Kompetenzen [Selection of medical students: Measurement of cognitive abilities and psychosocial competencies]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61(2):178-186. DOI: 10.1007/s00103-017-2670-2
21. Jackson DJ, LoPilato AC, Hughes D, Guenole N, Shalfrooshan A. The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. J Occup Organ Psychol. 2017;90(1):1-27. DOI: 10.1111/joop.12151
22. Mielke I, Breil SM, Amelung D, Espe L, Knorr M. Assessing distinguishable social skills in medical admission: does construct-driven development solve validity issues of situational judgment tests? BMC Med Educ. 2022;22(1):293. DOI: 10.1186/s12909-022-03305-x
23. Melro CM, Pack R, MacLeod A, Rideout A, Watson-Creed G, Burm S. Front row seat: The role MMI assessors play in widening access to medical school. Med Teach. 2023:1-8. DOI: 10.1080/0142159X.2023.2289851
24. Robinett K, Kareem R, Reavis K, Quezada S. A multi-pronged, antiracist approach to optimize equity in medical school admissions. Med Educ. 2021;55(12):1376-1382. DOI: 10.1111/medu.14589
25. O'Sullivan L, Kagabo W, Prasad N, Laporte D, Aiyer A. Racial and Ethnic Bias in Medical School Clinical Grading: A Review. J Surg Educ. 2023;80(6):806-816. DOI: 10.1016/j.jsurg.2023.03.004
26. Herde CN, Lievens F, Jackson DJ, Shalfrooshan A, Roth PL. Subgroup differences in situational judgment test scores: Evidence from large applicant samples. Int J Sel Assess. 2020;28(1):45-54. DOI: 10.1111/ijsa.12269
27. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. Med Educ. 2016;50(6):624-636. DOI: 10.1111/medu.13060

28. Statistisches Bundesamt. Migrationshintergrund. Wiesbaden: Desatis; 2021. Zugänglich unter/available from: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/Glossar/migrationshintergrund.html

29. Corstjens J, Lievens F, Krumm S. Situational judgement tests for selection. In: Goldstein HW, Pulakos ED, Passmore J, Semedo C, editors. The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention. Hoboken (NJ): Blackwell Publ; 2017. p.226-246. DOI: 10.1002/9781118972472.ch11

30. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. Med Teach. 2013;35(9):e1437-e1446. DOI: 10.3109/0142159X.2013.818634

31. Knorr M, Schwibbe A, Ehrhardt M, Lackamp J, Zimmermann S, Hampe W. Validity evidence for the Hamburg multiple mini-interview. BMC Med Educ. 2018;18(1):106. DOI: 10.1186/s12909-018-1208-0

32. Hissbach JC, Klusmann D, Hampe W. Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. BMC Med Educ. 2011;11(1):83. DOI: 10.1186/1472-6920-11-83

33. Kadmon G, Kadmon M. Academic performance of students with the highest and mediocre school-leaving grades: Does the aptitude test for medical studies (TMS) balance their prognoses? GMS J Med Educ. 2016;33(1):Doc7. DOI: 10.3205/zma001006

34. Schick K, Reiser S, Janssen L, Schacht L, Pittroff SID, Dörfler E, Klein E, Roenneberg C, Dinkel A, Fleischmann A, Berberat PO, Bauer J, Gartmeier M. Training in medical communication competence through video-based e-learning: How effective are video modeling and video reflection? Patient Educ Couns. 2024 Apr;121:108132. DOI: 10.1016/j.pec.2023.108132

35. Schubert S, Ortwein H, Dumitsch A, Schwantes U, Wilhelm O, Kiessling C. A situational judgement test of professional behaviour: development and validation. Med Teach. 2008;30(5):528-533. DOI: 10.1080/01421590801952994

36. Institut für Kommunikations- und Prüfungsforschung gGmbH. Studentischer kompetenzorientierter Progresstest 2024. Heidelberg: Institut für Kommunikations- und Prüfungsforschung; 2024. Zugänglich unter/available from: https://www.komp-pt.de/

37. Knorr M, Rudloff A, Breil SM, Schwibbe A. Use of Situational Judgement Tests for Admission into Medical School: Experiences from the University Medical Centre Hamburg. In: 15th Conference of the Differential Psychology, Personality Psychology and Psychological Assessment (DPPD) of the German Psychological Society (DGPs); 2019 Sep 16-18; Dresden, Germany.

38. Pan X, Huang V, Laumbach S, Copeland HL, Akinola M, Rosenbaum D, MacIntosh A. Impact of patterns of language use and socio-economic status on a constructed response Situational Judgment Test (SJT). PLoS One. 2023;18(8):e0289420. DOI: 10.1371/journal.pone.0289420

39. Rindermann H, Oubaid V. Auswahl von Studienanfängern durch Universitäten - Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. J Individ Differ. 1999;20(3):172-191. DOI: 10.1024//0170-1789.20.3.172

40. Groene OR, Huelmann T, Hampe W, Emami P. German physicians and medical students do not represent the population they serve. Healthcare (Basel). 2023;11(12):1662. DOI: 10.3390/healthcare11121662

41. Finger C, Hampe W, Wittenberg T. Eignungstests für das Medizinstudium: Perspektiven auf Diversität und Fairness. In: Bundesministerium für Bildung und Forschung, editor. Vielfalt und Chancengerechtigkeit in Studium und Wissenschaft. Berlin: Bundesministerium für Bildung und Forschung (BMBF); 2023.

42. Autor:innengruppe Bildungsberichterstattung. Bildung in Deutschland 2022. Bielefeld: wbv Publikation; 2022.

43. OECD. PISA 2022 Results (Volume I): The State of Learning and Equity in Education. Paris: OECD Publishing; 2023.

44. Neugebauer M, Schindler S. Early transitions and tertiary enrolment: The cumulative impact of primary and secondary effects on entering university in Germany. Acta Sociologica. 2012;55(1):19-36. DOI: 10.1177/0001699311427747

45. Kristen C, Reimer D, Kogan I. Higher Education Entry of Turkish Immigrant Youth in Germany. Int J Comp Soc. 2008;49(2-3):127-151. DOI: 10.1177/0020715208088909

46. Bundesverfassungsgericht. Leitsätze zum Urteil des Ersten Senats vom 19. Dezember 2017. 1 BvL 3/14, 1 BvL 4/14. Karlsruhe: Bundesverfassungsgericht; 2017. Zugänglich unter/available from: https://www.bverfg.de/e/ls20171219_1bvl000314.html

47. Groene O, Mielke I, Knorr M, Ehrhardt M, Bergelt C. Associations between communication OSCE performance and admission interviews in medical education. Patient Educ Couns. 2022;105(7):2270-2275. DOI: 10.1016/j.pec.2021.11.005

48. Rosenfeld JM, Reiter HI, Trinh K, Eva KW. A cost efficiency comparison between the multiple mini-interview and traditional admissions interviews. Adv Health Sci Educ Theory Pract. 2008;13(1):43-58. DOI: 10.1007/s10459-006-9029-z

49. Mallinger R, Holzbaur C, Mutz N, Prodinger WM, Heidegger M, Hänsgen KD, Spicher B. EMS: Eignungstest für das Medizinstudium in Österreich. Wien/Innsbruck: Medizinische Universität Innsbruck/Medizinische Universität Wien; 2011.

50. Spicher B, Hänsgen KD. EMS 2017 Bericht 24. Eignungstest für das Medizinstudium in der Schweiz. Bericht über Durchführung und Ergebnisse. Granges-Paccot: Zentrum für Testentwicklung und Diagnostik am Departement für Psychologie der Universität Freiburg; 2017.

Korrespondenzadresse:
Mirjana Knorr
Universitätsklinikum Hamburg-Eppendorf, Arbeitsgruppe Auswahlverfahren, Martinistr. 52, 20251 Hamburg, Deutschland
m.knorr@uke.de

Bitte zitieren als
Knorr M, Mielke I, Amelung D, Safari M, Gröne OR, Breil SM, MacIntosh A. Measuring personal characteristics in applicants to German medical schools: Piloting an online Situational Judgement Test with an open-ended response format. GMS J Med Educ. 2024;41(3):Doc30.
DOI: 10.3205/zma001685, URN: urn:nbn:de:0183-zma0016855

Artikel online frei zugänglich unter
https://doi.org/10.3205/zma001685

Eingereicht: *03.11.2023*
Überarbeitet: *19.03.2024*
Angenommen: *17.04.2024*
Veröffentlicht: *17.06.2024*