frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/feduc.2022.931841

OPEN ACCESS

EDITED BY
Nicolas Becker,
University of Greifswald, Germany

REVIEWED BY
Matthias Stadler,
Ludwig Maximilian University of Munich, Germany
Florian Krieger,
Technical University Dortmund, Germany

*CORRESPONDENCE
Marek Denker
✉ denker.marek@gmail.com

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 29 April 2022
ACCEPTED 28 November 2022
PUBLISHED 05 January 2023

CITATION
Denker M, Schütte C, Kersting M, Weppert D and Stegt SJ (2023) How can applicants' reactions to scholastic aptitude tests be improved? A closer look at specific and general tests. *Front. Educ.* 7:931841. doi: 10.3389/feduc.2022.931841

How can applicants' reactions to scholastic aptitude tests be improved? A closer look at specific and general tests

Marek Denker[1]*, Clara Schütte[2], Martin Kersting[3], Daniel Weppert[4] and Stephan Josef Stegt[4]

[1]Institute of Psychology, University of Bonn, Bonn, Germany, [2]Institute of Psychology, University Heidelberg, Heidelberg, Germany, [3]Institute of Psychology, Justus Liebig University Giessen, Giessen, Germany, [4]Institute for Test Development and Talent Research, ITB Consulting GmbH, Bonn, Germany

Since the 1980's, scientific interest in applicants' reactions to admission procedures has been growing. Several theoretical frameworks and evaluation questionnaires were developed. Many researchers have asked potential participants about their attitudes—what is missing, however, are studies of applicants who have actually taken part in the selection procedures. In addition, applicants' reactions to student selection procedures receive less attention than applicants' reactions to personnel selection procedures. Furthermore, tests and testing conditions continue to develop, e.g., through online-based testing and test supervision at home ("proctoring"). Therefore, we used a standardized questionnaire for measuring the overall test evaluation and different dimensions of acceptance including face validity, controllability, and the absence of strain to examine six scholastic aptitude tests. We added items to get deeper insights into situational aspects. According to the results of 2,052 test participants, applicants prefer specific tests and shorter tests. Situational aspects, such as privacy, working conditions, and prior information, have an influence on the perceptions of applicants. Proctored testing is evaluated positively, but taking the test at a test center is still rated more favorably. This study discusses the practical implications.

KEYWORDS

applicants' reactions, admission test, scholastic aptitude test, university admission, test evaluation, face validity

1. Introduction

The topic of applicants' reactions to the selection procedures has been increasingly yielding interest in research and practice since the 1980's (Schuler and Stehle, 1983; Anderson et al., 2004, 2010; Ryan and Huth, 2008; Truxillo et al., 2017). In the field of employee selection, there is already a body of literature on applicants' reactions to selection procedures (Gilliland and Steiner, 2012; McCarthy et al., 2017). Numerous results were summarized meta-analytically by Anderson et al. (2010).

Although there is a large number of empirical studies on applicants' reactions, there is a need for further research on at least three points: (1) Most of them deal with the issues related to the image of personnel selection procedures (e.g., König et al., 2010; Krumm et al., 2011; Benit and Soellner, 2012). The majority of cases involve interviewing people for whom it is unclear whether they have ever taken part in a similar procedure. According to the instruction, they are supposed to imagine that they have been invited to an interview or test as part of the selection process. Marcus (2003) asked research participants twice to give their opinion on the acceptability of selection procedures: once based on a short description of the procedure (as in most applicants' reaction studies) and once after they had actually participated in the procedure (in a shortened version). He showed that pretest and posttest measurements were only modestly related. Many participants changed their minds after taking part in the procedure. Indeed, the image of a procedure is important but in order to find out how selection procedures are actually perceived and to identify improvement needs, it is necessary to interview the participants that actually underwent such a procedure. These kinds of studies are rare. In addition, in most of the few studies that include evaluations by participants who attended a certain procedure, it was not a high-stakes situation—the procedure had no relevant consequences (Speer et al., 2016; Watrin et al., 2019; Gnambs, 2022). We aim to fill this gap with the current research.

(2) Only a few studies deal with the selection of students for universities. In university admission procedures, scholastic aptitude tests play an important role. Many universities are not exclusively relying on final grades in the selection process. This is partly due to the low comparability of final grades in some contexts (e.g., differing educational and grading standards in different schools, regions, and countries). Moreover, in different European countries, the increasing interest in the selection of students is partially due to changing legislation and increasing internationalization (Becker and Kolster, 2012). Among the alternative or additional selection procedures, aptitude tests are widely used, mainly because they are characterized by a high prognostic validity (Mauger and Kolmodin, 1975; Bejar and Blew, 1981; House and Keeley, 1997; Kuncel et al., 2005; Hell et al., 2007; Schult et al., 2019). The importance of aptitude tests in student selection requires a closer analysis, also considering applicants' reactions. Applicants' reactions to university admission procedures received less attention than applicants' reactions to personnel selection procedures (Hell and Schuler, 2005), which is another gap of knowledge to be filled.

The few studies on applicants' reactions to university admission procedures provide an investigation of attitudes toward selection tools (Hell and Schuler, 2005; Herde et al., 2016; Stegt et al., 2018). Anderson et al. (2010) found that tests had less acceptance than interviews, while Stegt et al. (2018) found that tests were assessed more positively than interviews. This indicates that specific acceptance conditions exist in the context of student selection. It is particularly interesting that in the study by Stegt et al., subject-specific tests were assessed significantly more positively than general intelligence tests. This raises the research question of whether this finding is also evident when the respondents have actually taken the tests.

Final grades were rated favorably in the study of Herde et al. (2016), whereas in Hell and Schuler (2005) and Stegt et al. (2018), they were rated rather low.

There are multiple reasons to focus more on applicants' reactions to university admission procedures. Applicants' evaluations can have an impact on a university's reputation. A decent reputation elicits the interest of potential applicants and facilitates the selection of eligible applicants (Heine et al., 2006; Hell and Haehnel, 2008). An attractive test can increase the motivation to take it. Tests with a high face validity give a preview of the requirements of the program and support self-selection. In addition, quality criteria for diagnostic tools include aspects such as fairness, transparency, acceptance, and face validity that should also be assessed.

(3) Previous findings focus on comparing reactions to different diagnostic procedures. However, research on the specific design and implementation conditions of a procedure is limited. What difference does it make whether a test is short or long, abstract or specific to a particular field of study, whether it is presented at home or a test center?

Thus, the present study aims to fill three research gaps: (1) interviewing applicants after a high-stakes test, (2) investigating university admission, and (3) comparing different testing conditions and different tests. In addition, we aim to provide an evaluation tool to be used not only for research but also for improving scholastic aptitude tests and testing conditions.

2. Research, models, and questionnaires on applicants' reactions

2.1. Research on applicants' reactions

Early studies considering applicants' reactions to selection procedures label the concept as social validity or social quality. Schuler and Stehle (1983) described the construct of social validity as subjective judgments of applicants to selection procedures. According to Schuler (1990)[1], the concept encompasses several aspects that can serve to improve the process in terms of fairness and acceptance of candidates. Thus, social validity is a collective term capturing everything that turns a psychometric process into an overall acceptable, social situation. The author highlights four key features accounting for a fair and acceptable overall situation: prior

1 Hans Schuler, an essential founder of research on applicant reactions, born on 01.02.1923, died on 06.03.2021. This article is dedicated to him.

information, participation, transparency, and feedback (Schuler, 1990). Prior information includes information regarding the job, the organization, and the requirements. Participation is related to opportunities to display relevant knowledge and skills. Transparency is related to the whole selection process, including the assessment and selection tools. The component feedback involves the provision of feedback regarding the result and the content of the information exchanged. For instance, the social validity of feedback is given if it is restricted to the description of behavior and if the privacy of the applicant is taken into account.

Arvey et al. (1990), devoted themselves to motivational components in selection tests and developed the Test Attitude Survey (TAS), which has been used in several studies (Neuman and Baydoun, 1998; Sanchez et al., 2000; Rogelberg et al., 2001; Combs et al., 2007; Visser and Schaap, 2017; Liu and Hau, 2020). The TAS consists of nine dimensions, namely, motivation, lack of concentration, belief in tests, comparative anxiety, test ease, external attribution, general need (for achievement), future effects, and preparation.

An influential model of applicants' reactions was developed by Gilliland (1993), based on the concept of organizational fairness. In his model, the perceived procedural and distributive fairness of the application process influences applicants' reactions. Gilliland (1993) established 10 rules of procedural fairness in three categories: formal characteristics, explanation, and interpersonal treatment. Distributive fairness is defined by three rules: equity, equality, and needs.

Hausknecht et al. (2004) built upon Gilliland's (1993) model and its extension by Ryan and Ployhart (2000). In their updated theoretical model of applicants' reactions to selection (Hausknecht et al., 2004), applicant perceptions are influenced by personality traits and perceptions of the procedure, as well as job characteristics and organizational characteristics.

In the meta-analysis by Anderson et al. (2010), seven key dimensions of overall favorability were derived, and the framework by Steiner and Gilliland (1996) served as a foundation: scientific evidence, employer's right, an opportunity to perform, interpersonal warmth, face validity, widely used, and respectful of privacy. Those dimensions were applied to 10 different assessment methods, among those cognitive tests. This study indicates the generalizability of the dimensions used.

2.2. The Akzept-questionnaire

Kersting (1998, 2008) took into account the models of Gilliland (1993) and Schuler (1990). Kersting prefers the term social acceptance or social quality to the term social validity, as these terms do not suggest a contextual connection to the quality criterion of validity.

In order to investigate applicant's reactions, Kersting (1998, 2008, 2010) developed the Akzept-questionnaire, aiming to provide a psychometrically sound instrument that would make results comparable across different studies. From Schuler's (1990) dimensions, Kersting (1998) refers to the dimensions of transparency and participation. In addition to Gilliland's (1993) model, he includes personal variables such as perceived and actual test performances. The objective of this study was to determine the moderation of the judgment of social acceptance by personal variables. There are four versions of the Akzept-questionnaire: for assessment centers (AC), for interviews (I), for performance tests (L), and for personality tests (P). The differences among the four versions are mainly in the choice of words. In addition, some procedure-specific items complete each version (such as "good organization at the assessment center"). The main scales are face validity, controllability, the absence of strain, and measurement quality. In addition, a single item is included to assess the overall satisfaction with the test.

The various forms of the Akzept have been used in diverse contexts. In the following, we report some findings on applicants' reactions to tests (Akzept-L-questionnaire). On the use of the Akzept-questionnaire in assessment centers, refer to Kersting (2010), König et al. (2015), and Melchers and Annen (2010). On the use of the Akzept-P-questionnaire to explore the acceptance of personality questionnaires, refer to Beermann et al. (2013) as well as Watrin et al. (2019).

Beermann et al. (2013) compared applicants' reactions to achievement tests in direct comparison with the corresponding acceptability assessment of personality questionnaires using Akzept-L and Akzept-P. In comparison, the intelligence test was rated better regarding the measurement quality; the job-specific personality test was rated better in terms of face validity. The ratings of the Akzept-dimensions were related to the overall rating in the study. In the case of the intelligence test, the absence of strain was particularly important for the overall rating. Thus, face validity and the absence of strain seem to be particularly significant for acceptance.

Kersting (2008) compared applicants' reactions to six different achievement tests. The overall evaluation of achievement tests was positive. Considering all the tests, the measurement quality and the controllability of all achievement tests were rated particularly positive, while the face validity was assessed critically. Even if a test situation is stressful for participants, this does not have a negative effect on the overall evaluation rating; apparently, they perceive the burden as appropriate. In addition to these similarities in applicants' reactions to the various tests, there are also clear differences in applicants' reactions. The highs and lows of applicants' reactions to the Raven test (matrices) were particularly extreme: the permanent repetition of one and the same item type had a favorable effect on the evaluation of measurement quality and controllability, but the participants could hardly establish a connection between the abstract test items on the one hand and real-life tasks on the other, resulting in a poor evaluation of face validity, which again proved to be of great importance.

Benit and Soellner (2012) and Krumm et al. (2011) showed that the tests have high face validity if they are job-specific. This was shown in both studies for the tests with job-specific content. The examination of whether this also applies to the tests with specific relation to the field of academic studies is still pending.

Besides the test content, the medium of presentation (paper-pencil or computer-based) also plays an important role, as university selection tests are presented in both formats. A study on the acceptance of an achievement test (measured with the Akzept-L questionnaire) that differs regarding the medium is presented by Gnambs (2022). In terms of face validity, the computer-based version of a test was rated better than the paper-based version.

While the Akzept-questionnaire was mostly based on the models primarily used in the context of personnel selection (Schuler, 1990; Gilliland, 1993) and has often been used in that context (Kersting, 2010; Melchers and Annen, 2010; König et al., 2015), it can notably be applied in a variety of other contexts. As such, it is also a useful tool for researching applicants' reactions to university admission tests.

2.3. Adding situational aspects to the Akzept-questionnaire

The Akzept-questionnaire has turned out to be valuable in comparing different tests with respect to applicants' reactions. But when it comes to the evaluation of a certain test with the objective of improving it, more information is needed. If a dimension is rated lower than expected, what can be done to improve it? How can we influence the feelings of controllability, face validity, or the absence of strain? Indeed, we can draw some plausible assumptions: a long test could have lower ratings on the absence of strain than a short test. Providing information to participants (e.g., preparation material, explaining the purpose of the test, and explaining the implementation conditions) could enhance face validity and feelings of controllability. But using only the Akzept-questionnaire, we are not in the position to test these assumptions empirically or evaluate these situational aspects.

In addition, admission tests and their situational conditions have been further developed in the recent years. They are often implemented as computer-based or online tests. Specifically, during the COVID-19 pandemic, tests with the so-called proctoring appeared as an attractive alternative to tests in test centers or on campus (Stegt and Hofmann, 2020). During proctored tests, participants can take a test at home and be supervised *via* webcam and screen-sharing. This way, proctoring embodies a strongly technologized kind of testing. There could be some drawbacks associated with this alternative of testing. Proctored tests could be perceived as violations of privacy. Moreover, due to a highly technologized test situation, they involve a stronger risk of technical complications. Participants are required to ensure that the testing procedure runs smoothly as well as to take more individual, direct responsibility. According to our knowledge, there is no research that investigates how applicants perceive a proctoring test.

Therefore, we decided to add several items to the original Akzept-questionnaire. For this purpose, we analyzed complaints and concerns of test takers and universities over the last decade of university admission testing of the Institute for Test Development and Talent Research ("Institut für Test- und Begabungsforschung" in German, ITB). The ITB develops and implements several scholastic aptitude tests in Germany, Switzerland, and Austria, such as the Test for Medical Studies (TMS) or, together with the universities of Heidelberg, Freiburg, and Tübingen, the Pharmaceutical Studies Aptitude Test (PhaST). About 40,000 participants per year take 1 of 15 scholastic aptitude tests by ITB. Based on their feedback, the following aspects were added to the Akzept-questionnaire:

2.3.1. Concentration

Some candidates complained that it was difficult to concentrate due to the testing conditions (e.g., "It was difficult to focus on the items because… of heat in the room/noise of a nearby construction site/another candidate coughing all the time," or with proctoring, "I had issues with my webcam and the proctor asked me to refresh the page several times," and "I had a weak internet connection, the test was interrupted and I had to restart"). We assume that the feeling of not being able to concentrate will increase strain and affect the overall satisfaction with the test. The lack of concentration is also a dimension in the TAS questionnaire, concentration is a precondition to the participation component in the model of Schuler (1990), and it should contribute to procedural fairness in the model of Gilliland (1993).

2.3.2. Privacy

Data protection is a very important topic, especially in Europe. The protection of privacy needs to be ensured according to the European Convention on Human Rights EMRK (1950) and the Charter of Fundamental Rights of the European Union. The General Data Protection Regulation (GDPR) strengthens the protection of privacy in the EU. Specifically, when introducing proctoring, there were concerns about universities and their data protection officers. It was discussed whether filming participants at home constituted an intrusion into their private life. We did not receive any complaints or questions from participants regarding their privacy but decided nevertheless to ask them how they felt about data protection and privacy. We think that perceived privacy could influence the dimensions such as the absence of strain and controllability. Privacy is also

taken into account by the model of Schuler (1990) and the study of Anderson et al. (2010).

2.3.3. Prior information

Participants want to be informed about the test itself, about how to prepare, and, especially with proctoring, about how to prepare for the test environment at home (e.g., "Where can I find more training material?" and "What can I do if there are problems with my internet connection?"). We think that transparency and information will not only affect the perceptions of face validity but also reduce the stress level, and improve the controllability and the overall evaluation. Information provided prior to the test and transparency were also important in the model of Schuler (1990) and influence the perception of procedural fairness in Gilliland's (1993) model.

In addition to these situational aspects, we had a look at the characteristics of the tests, which may play a role in applicants' reactions: test taker fees, test duration, implementation in test centers vs. proctoring, and the level of specificity for the study programs. Speer et al. (2016) showed in the context of personality and cognitive tests, a longer test format caused more fatigue and was perceived as causing more effort, and overall, the duration did not worsen applicants' reactions when applying for job offers.

2.4. The present study

The objective of the present study was to explore participants' reactions to university admission tests in a high-stake situation. We investigated the overall evaluation and three established factors of acceptancy: face validity, controllability, and the absence of strain. We examined different aspects of the tests and the testing conditions and explored how they influence the evaluations. These aspects were derived from the discussions with universities and participants over the last decade, and they are also related to the theories and models about applicants' reactions. They include test taker fees, the information given prior to the test, privacy, test duration, and test locations (in test centers vs. at home).

2.5. Hypotheses and research questions

First, we intend to reveal whether the three Akzept-dimensions contribute to the overall evaluation of scholastic aptitude tests.

H1. Face validity, controllability, and the absence of strain contribute to the overall evaluation of scholastic aptitude tests. This holds for the entire sample (H1a) as well as for subgroups (H1b).

As scholastic aptitude tests are the preferred selection criteria for university admission by the applicants (Stegt et al., 2018), we think that the tests in question will be evaluated positively.

H2. Scholastic aptitude tests are evaluated positively on average across all participants and by the majority of applicants. This holds for the entire sample (H2a) as well as for each of the six tests (H2b).

Due to evidence suggesting that specific aptitude tests elicit more favorable applicants' reactions than general tests and are evaluated better in terms of face validity (Krumm et al., 2011; Benit and Soellner, 2012; Beermann et al., 2013), we assume that the degree of specificity of a test improves the evaluation by the applicants.

H3. The higher the degree of specificity of a test for the study program, the better the evaluation by applicants. Specificity contributes to face validity (H3a) and to the overall evaluation (H3b).

The longer participants have to focus on a mentally challenging and tiring test, the more they will feel exhausted or depleted. Speer et al. (2016) showed in the context of a general mental ability test that a longer test version required more effort from the participants. This suggests that the test length will also affect the evaluation by the participants.

H4. Test length affects the dimension absence of strain negatively.

In addition, we want to assess the impact of situational testing conditions. The availability of prior information is not a characteristic of the test itself. The amount and quality of information can be influenced by the organization(s) using or offering the test. By providing sufficient information regarding the testing situation and the items, participants will have a better feeling of control and a better understanding of the utility of the test and the type of tasks used. Transparency and pre-information are required by several standards of psychometric testing.

H5. The better the prior information is evaluated by applicants, the better their evaluation of the test is. This holds for the dimensions such as controllability (H5a), face validity (H5b), and overall evaluation (H5c).

The possibility to work in a concentrated manner without being distracted or interrupted is important for the quality criterion objectivity and also for applicants' reactions. While concentration should positively affect the performance on a task, perceived incapability to focus on the items will cause strain and participants will think that they were not able to show their abilities, which will lead to a less favorable overall evaluation.

H6. The feeling of staying concentrated during the test will positively affect the evaluation of the test. This holds for the

dimensions such as the absence of strain (H6a) and the overall evaluation (H6b).

Besides these six hypotheses, we want to explore three questions.

Proctoring was used by several test providers since 2020 in order to make testing possible during the pandemic situation. We are curious to understand how university applicants react to this new method of testing. Given that there might be not only several advantages but also problems from the participant's point of view, we do not formulate hypotheses but compare the ratings of participants who took an admission test either in a test center or with proctoring. One of our tests (TM-WISO) was administered in both conditions and we are therefore able to compare their ratings.

Q1. How is proctoring evaluated by applicants?

In many countries, especially in Europe, providing equal opportunities in the educational system is an important value and objective. The opportunity to attend valuable study programs should not depend on the financial resources of the person (or his/her family). Therefore, any kind of fee in the educational system can lead to criticism in the public and political discussions, because this might improve the chances for wealthy people. On the other hand, products and services that are offered free or at very low prices might lead to the suspicion of bad quality. As the test taker fees in our sample range from 0 to 300 euros, we want to know whether the fees affect the evaluation by the participants.

Q2. How do test taker fees affect the evaluation of the tests?

Data protection represents another topic of interest. The protection of personal data is an important issue in many countries. Several reports of personal data being stolen and/or misused led to strict data protection rules and high sensitivity of persons and organizations. Even if the test providers act according to the strict regulations in Europe and inform the participants about their privacy policy, participants could feel uncomfortable about the security of their data. Participants taking a proctoring test and being filmed at home could have more concerns regarding this topic. We, therefore, want to explore what participants think about privacy and how they evaluate the different tests and testing conditions in this respect.

Q3. How do participants feel about privacy and data protection?

3. Methods

3.1. Scholastic aptitude tests

3.1.1. PhaST

The PhaST is a scholastic aptitude test for pharmaceutical studies. Participants have to work precisely under time pressure, deal with complex rules (e.g., naming of multi-unit polygon systems), understand structural chemical formulas, memorize and recall pharmaceutical information (e.g., pharmaceutical agents, biological targets, and structural chemical formulas), understand subject-specific texts, graphs, and tables, perform mental rotation with molecules, and interpret pharmaceutical experiments. Additional modules assess knowledge from mathematics, physics, biology, and chemistry.

3.1.2. TM-WISO

TM-WISO is a test for master's programs in business administration, economics, and social sciences. It measures the quantitative and verbal abilities, reasoning ability, and ability to plan and organize the projects with four task groups.

3.1.3. ITB-Business

ITB-Business is a scholastic aptitude test for business administration, economics, and social sciences. It measures quantitative and general figural mental abilities with unspecific tasks, as well as comprehension and interpretation of texts, graphs, and tables from business administration, economics, and social sciences with four task groups.

3.1.4. ITB-Science

ITB-Science is a scholastic aptitude test for STEM (Science, Technology, Engineering, and Mathematics) programs. It measures the quantitative and general figural mental abilities with unspecific tasks, as well as comprehension and interpretation of texts, graphs, and tables from different STEM subjects with four task groups.

3.1.5. GSAT

GSAT is a general scholastic aptitude test. It measures the quantitative and general figural mental abilities with unspecific tasks, as well as comprehension and interpretation of texts, graphs, and tables from different subjects (STEM, but also economics and social sciences) with four task groups. A German federal state applied GSAT to provide full university access for prospective students with a high school degree that is insufficient for regular university admission.

3.1.6. ICOS

ICOS is a short general mental ability screening used in the context of personnel selection. A university used it as a screening tool for preselection. It includes eight different task formats with verbal, figural, and numerical tasks.

An overview of all tests used in this study and their respective implementation conditions is shown in Table 1.

TABLE 1 Overview of the six scholastic aptitude tests.

Test	Specificity level	Implementation conditions	Duration (min)	Participation fee (€)
PhaST	4	Test center	240	75
TM-WISO	3	Both	230	100
ITB-Business	3	Proctoring	120	50
ITB-Science	3	Proctoring	120	300
GSAT	2	Proctoring	120	200
ICOS	1	Proctoring	30	0

Specificity levels: 1 = general mental ability test, 2 = general scholastic aptitude test, 3 = field-specific scholastic aptitude test, 4 = subject-specific scholastic aptitude test.

3.2. Variables and measurement instruments

As test characteristics, we investigate test taker fees (in Euro), test duration (in minutes), implementation condition (proctoring vs. test center), and level of specificity (one for the general mental ability screening, two for the general scholastic aptitude test, three for the field-specific tests, and four for the subject-specific test). The level of specificity is derived from the test contents: PhaST is designed only for the study program "Pharmazie, Staatsexamen," therefore it is subject-specific. TM-WISO and ITB-Business are designed for master's and bachelor's programs in business administration as well as economics and social sciences, therefore it is field-specific. ITB-Science is designed for different study programs in the field of STEM (Science, Technology, Engineering, and Mathematics), therefore it is also field-specific. GSAT is a general scholastic aptitude test without any reference to a specific program or field. In addition, ICOS is a screening for general mental abilities, which was developed for the preselection of the applicants by the companies.

In order to assess the situational conditions, we use three items for concentration, three items for privacy, and two items for prior information measured on a six-point Likert scale (1 = strongly disagree, 6 = strongly agree).

Regarding the applicants' reactions, we use the Akzept-questionnaire for performance tests (Kersting, 2008) with the dimensions such as controllability, the absence of strain, and face validity. Every dimension has four items on a six-point Likert scale (1 = strongly disagree, 6 = strongly agree).

From the initial Akzept-scales, the scale measurement quality was left out. We decided to eliminate this scale because it proved to correlate highly with face validity: Kersting (2008) found these two scales to correlate with $r = 0.67$, Beermann et al. (2013) even found a correlation of $r = 0.84$. In order to keep the questionnaire as short as possible, this scale was left out.

Table 2 shows the items and scales as well as the reliability of our study ($N = 2{,}052$).

We assume that all the values of 3.5 or higher are positive ratings due to the semantic anchors associated with each response category. After recoding negatively formulated items, 1, 2, and 3 are negative answers (affirmation of critical statements), 4, 5, and 6 are positive answers (affirmation of positive), and 3.5 is the middle of the 1–6 scale. As this is the first study to collect data on the acceptability of scholastic aptitude tests, we cannot say which acceptability scores are "average" or whether other tests achieve better or worse acceptability scores. However, we can state that values above 3.5 stand for positive acceptance in absolute terms.

3.3. Participants

In total, 2,052 participants filled out the questionnaire sufficiently (i.e., they provided an overall evaluation and evaluated the three Akzept-dimensions). Due to different privacy agreements with universities, data regarding age and gender were only available for 995 and 1,013 participants, respectively. The mean age was 21.3 ($SD = 3.28$), and 55% were women. In total, 256 applicants evaluated PhaST (of 392 test takers, so the response rate was 65.3%), 264 for the TM-WISO (1,251, 21.1%), 76 for the ITB-Business (99, 76.8%), 395 for the ITB-Science (436, 90.6%), 962 for the GSAT (1,050, 91.6%), and 99 for the ICOS (148, 66,9%).

PhaST participants completed the test in order to apply for a pharmaceutical study program at one or more German universities. At the time of the study, six universities were using this test for student selection. TM-WISO participants did the test in order to apply for a master's program in business administration or economics at one or more of the eight universities using this test for student selection. ITB-Business participants applied for a private German business school. ITB-Science participants applied for a private Austrian University of Health Sciences. GSAT is an examination that is offered by a German federal state for prospective students who want to upgrade their university entrance qualification. ICOS participants applied for a public Austrian university of applied sciences.

TABLE 2 Scales and items of the questionnaire and internal consistencies of the scales.

Scale	Item
Controllability ($\alpha = 0.77$)	The test items were clear and understandable.
	I did not understand the instructions (-).
	When processing the test items, I always knew precisely what I had to do.
	I did not understand the test items (-).
Face validity ($\alpha = 0.82$)	It is doubtful that the test can be applied to find suitable candidates for a study program (-).
	Whether someone scores well on the test items or is doing well in a study program are two entirely different things (-).
	The test items do not reflect reality well-enough to predict academic success (-).
	The test items reflect requirements that are also relevant to the study program.
Absence of strain ($\alpha = 0.76$)	When dealing with the test, I felt overstrained (-).
	The processing of the test items is stressful (-).
	The test items were mostly too difficult for me (-).
	The processing of the test items is exhausting (-).
Concentration ($\alpha = 0.70$)	I was able to work in a concentrated way.
	While processing the test, I was distracted by many things (-).
	Being supervised during the test distracted me and affected my concentration (-).
Privacy ($\alpha = 0.76$)	I felt uncomfortable being supervised while taking the test (-).
	I felt uncomfortable regarding the protection of personal data (-).
	I felt that being supervised during the test was an unpleasant interference into my privacy (-).
Prior information ($r = 0.38$)	The test provider should provide more information in advance about the type of tasks used (-).
	I felt adequately informed about the testing procedure prior to the testing.
Overall evaluation	Which school grade (following the German grade system) would you assign to the test? (-)

Negatively poled items (-) are recoded for the analyses; N = 2,052.

3.4. Procedure

Applicants for different study programs participated in their admission tests. The tests were implemented online in test centers or with proctoring at home.

For one of the six tests, TM-WISO, the situation proved to be irregular: Applicants had booked the appointments for test centers on specific test dates. During the pandemic, decisions about the implementation conditions depended on the current sanitary regulations and sometimes changed a few weeks before the test dates, forcing some applicants to postpone their test or switch to proctoring. On several test dates, both options were available, but the number of places in test centers was strongly restricted, so late registrations had no other option than to get tested with proctoring, which caused some complaints and frustration.

Participants filled out the evaluation questionnaire either paper-based directly after the test in test centers or they received an invitation[2] providing a link to an online questionnaire 2 days after the test. As participants received their test results around 2–3 weeks after the test, they did not know their results when they filled out the questionnaire. The online questionnaire was implemented by the software IONA (ITB online assessment). After sending the link, the online questionnaire was open for about 2 weeks. As participants in test centers wrote an individual code on the sheet and participants in the online version received individual links, the questionnaire was not anonymous.

3.5. Analyses

The negatively poled items were recorded, and then, mean scores were calculated with 6, being the best score.

2 The invitation mail had the following wording: Dear test taker! You have completed a study aptitude test and we now ask you to participate in a short survey. Participation is voluntary and your answers will not affect your test score. The data from this acceptance survey will be used for research purposes only and will not be shared with third parties. Please rate on a scale of 1 (strongly disagree) to 6 (strongly agree) whether a statement is true or not. For some questions you will be asked to give a school grade [German grading system from 1 (very good) to 6 (insufficient)]. Thank you very much for your support!

First, we used multigroup-structural equation modeling (MSEM) to test whether the dimensions of the Akzept-questionnaire have the same meaning in different subgroups of our study and whether it is, therefore, appropriate to compare the Akzept-ratings across the groups in subsequent analyses. MSEM was also used to evaluate whether the Akzept-dimensions contribute to the overall evaluation of scholastic aptitude tests (H1). For these analyses, we formed two groups in order to achieve sample sizes that are large enough to obtain accurate parameter estimates and model fit statistics. The first group consists of applicants who evaluated PhaST, TM-WISO, ITB-Business, or ITB-Science and was labeled as a specific test group, whereas the general test group consists of applicants who evaluated GSAT or ICOS. We used list-wise deletion to treat missing data for the MSEM analyses, which resulted in sample sizes of $n = 902$ for general tests and $n = 974$ for specific tests.

Due to the data being slightly skewed, we used the robust maximum likelihood (MLR) estimator and the Satorra-Bentler scaled χ^2-difference test (Satorra and Bentler, 2001). We proposed a model with three correlated exogenous factors (controllability, face validity, and the absence of strain) measured by their respective items and one endogenous factor (overall evaluation), which is influenced by all three exogenous factors. For identification purposes, all factor variances were fixed to 1. Because overall evaluation is measured with one item only, we fixed the residual variance of the respective item to 0 to identify the model.

In order to test measurement invariance, we fit a set of models with increasing restrictions to the specific and general test groups: In step one, configural invariance was tested by fitting the same model with the aforementioned factor structure in both groups. In step two, metric invariance was tested by constraining the factor loadings to be equal across groups. In step three, scalar invariance was tested by constraining the loadings and intercepts for all observed variables to be equal across the groups. In step four, strict invariance was tested by constraining the loadings, the intercepts, and the residual variances for all observed variables in the model to be equal across groups. To test for structural invariance, we fit the strict invariance model and additionally constrained the regression path coefficients and factor covariances to be equal across the groups. The statistics used to evaluate general model fit were the robust χ^2, Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). To judge these indices, we oriented on Hu and Bentler (1999) with CFI values > 0.90 and 0.95 defined as "acceptable fit" and "good fit," respectively, and for RMSEA, values < 0.08 and 0.06 defined as "acceptable fit" and "good fit," respectively. However, these cutoffs are used with caution, as they may

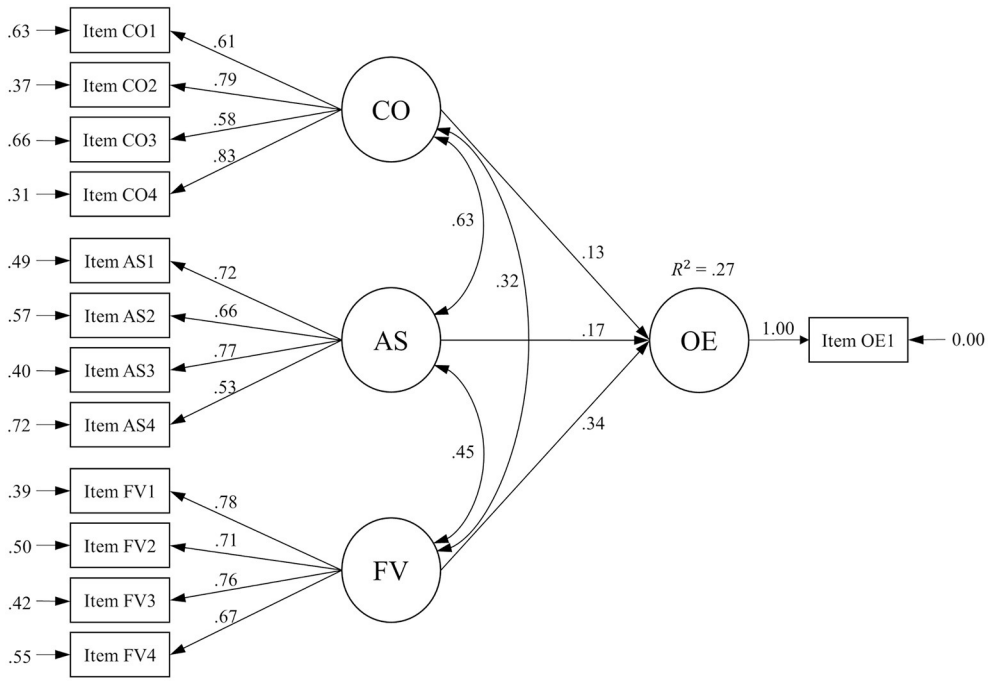

FIGURE 1
Proposed structural equation model. CO, controllability; AS, absence of strain; FV, face validity; OE, overall evaluation; standardized loading, structural and effect size estimates obtained from the complete sample ($N = 1,876$); the variances of all exogenous variables were set to 1 for model identification; all estimates are significant with $p < 0.001$.

TABLE 3 SEM results for total, individual, and invariance models across general and specific tests.

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
Total sample ($N = 1{,}876$)[a]	462.28**	60			0.948		0.063		0.045	
General tests ($n = 902$)	283.99**	60			0.932		0.067		0.053	
Specific tests ($n = 974$)	297.74**	60			0.946		0.067		0.049	
Configural invariance	581.76**	120			0.940		0.067		0.047	
Metric invariance	611.88**	129	29.77**	9	0.937	−0.003	0.066	−0.001	0.051	0.004
Scalar invariance	737.41**	138	132.72**	9	0.923	−0.014	0.071	0.005	0.055	0.004
Strict invariance	751.86**	150	25.43*	12	0.920	−0.003	0.069	−0.002	0.056	0.001
Structural invariance	784.75**	156	32.76**	6	0.917	−0.003	0.069	0.000	0.073	0.017

[a]Sample size is reduced for all analyses due to list-wise deletion; $\chi^2 = \chi^2$-value of absolute model fit; *df* = degrees of freedom; $\Delta\chi^2$ = value of the Satorra-Bentler scaled χ^2-difference test (note that since we used the Satorra-Bentler scaled χ^2-difference test, the $\Delta\chi^2$ cannot be calculated by subtracting the χ^2-value of the previous less restricted model from the χ^2-value of the actual model).

CFI, Comparative Fit Index; ΔCFI, CFI-difference between the actual model and the previous less restricted model; RMSEA, Root Mean Square Error of Approximation; ΔRMSEA, RMSEA-difference between the actual model and the previous less restricted model; SRMR, Standardized Root Mean Square Residual; ΔSRMR, SRMR-difference between the actual model and the previous less restricted model.

*Significant with $p < 0.05$.

**Significant with $p < 0.001$.

vary depending on the model complexity, sample size, or size of unique variances (e.g., Marsh et al., 2004; Heene et al., 2011). Since the χ^2-difference test is often overly sensitive to model rejection with large sample sizes (Saris et al., 2009), we evaluated fit differences for measurement invariance testing using the cutoffs by Chen (2007), who suggested that $\Delta CFI \leq -0.010$ coinciding with $\Delta RMSEA \geq 0.015$ or $\Delta SRMR \geq 0.030$ would indicate non-invariance for testing metric invariance. For all other levels of measurement invariance, a stricter cutoff for the ΔSRMR (≥ 0.010 instead of ≥ 0.030) was suggested. The assumption of invariance was therefore rejected as soon as the ΔCFI, and at least one of the two remaining indices exceeded their respective cutoff. It is important to note that the cutoffs proposed by Chen (2007) were originally developed for maximum likelihood (ML) rather than MLR estimation. However, Sass et al. (2014) demonstrated that the respective fit indices differ only slightly with MLR and ML estimation.

In order to test the assumption of a positive evaluation of scholastic aptitude tests, descriptive statistics are calculated (mean scores and proportion of participants who evaluate a scale positively, i.e., 3.5 or higher on the scale from 1 to 6). The hypotheses regarding the influence of dimensions, situational conditions, and test characteristics are tested with multiple linear regression models. The comparison of proctoring vs. test centers is done with *t*-tests for the TM-WISO participants. Statistical analyses were done with IBM-SPSS. The MSEM analyses were conducted in R (R Core Team, 2022) with the R package "lavaan" (Rosseel, 2012). Alpha was set to 0.05.

4. Results

4.1. Measurement and structural invariance and contributions of the dimensions to overall evaluations (H1)

We first tested the factorial validity in the total sample as well as in the two subgroups (general and specific tests) separately. The proposed model with standardized loading, structural, and effect size estimates for the total sample is depicted in Figure 1. All analyses provided evidence for factorial validity with consistently large loadings (λ between 0.53 and 0.83), which were all significant ($p < 0.001$). As shown in Table 3, the models showed an acceptable fit according to the conventions by Hu and Bentler (1999). As demonstrated by Heene et al. (2011) and McNeish et al. (2018), large sample sizes and high loadings can negatively affect model fit, even for correctly specified models or models with only minor misspecifications. In this context, we consider the model fit acceptable to proceed with the measurement invariance analyses.

The configural measurement invariance model also showed an acceptable fit. Despite the $\Delta\chi^2$ being significant for all model comparisons, strict measurement invariance can be found according to the cutoffs proposed by Chen (2007). The ΔCFI stayed below the threshold of −0.010 for the metric and strict invariance models. For the scalar invariance model, the ΔCFI exceeded the threshold but since the ΔRMSEA and the ΔSRMR stayed below the threshold of 0.015 and 0.010, respectively, for the metric, scalar, and strict invariance models,

TABLE 4 Average ratings by test: mean values and (in parentheses) standard deviations.

	PhaST	TM-WISO	ITB-Business	ITB-Science	GSAT	ICOS	Overall
Overall evaluation	4.57 (0.75)	3.52 (1.13)	4.58 (0.70)	4.16 (0.86)	4.02 (0.97)	4.06 (1.02)	4.08 (0.98)
Controllability	4.91 (0.85)	4.26 (1.01)	4.92 (0.78)	5.17 (0.79)	4.67 (0.97)	4.83 (1.00)	4.76 (0.96)
Face validity	3.90 (0.97)	3.01 (1.11)	3.58 (1.06)	3.14 (1.07)	2.65 (1.16)	2.57 (1.03)	2.98 (1.18)
Absence of strain	3.60 (0.97)	3.32 (1.08)	3.85 (0.86)	3.64 (0.99)	3.66 (1.06)	4.17 (1.16)	3.64 (1.05)
Concentration	5.37 (0.77)	4.87 (1.08)	4.97 (0.86)	4.78 (1.10)	5.11 (0.94)	5.10 (0.95)	5.04 (0.99)
Privacy	5.75 (0.49)	5.29 (1.10)	5.16 (0.99)	5.05 (1.14)	4.99 (1.12)	5.04 (1.06)	5.14 (1.08)
Prior information	3.42 (1.33)	3.86 (1.30)	3.78 (1.47)	4.30 (1.32)	3.52 (1.39)	3.71 (1.37)	3.72 (1.39)
n	256	264	76	395	962	99	2,052

Negatively poled items as the overall evaluation were recoded (1 = worst, 6 = best).

the (unstandardized) loadings, intercepts, and residual variances can be considered equal across both groups. We assume this as the first evidence that it is valid to make cross-group mean and regression comparisons.

Considering the total sample, all Akzept-factors contribute significantly to the overall evaluation. Standardized regression coefficients are $\gamma = 0.128$ ($p < 0.001$) for controllability, $\gamma = 0.172$ ($p < 0.001$) for the absence of strain, and $\gamma = 0.344$ ($p < 0.001$) for face validity (H1a). The results also indicate that structural invariance can be assumed for both groups, as only the ΔSRMR exceeds its cutoff while the ΔCFI and the ΔRMSEA stayed below their respective cutoffs. This indicates that the covariances and regression coefficients can be considered equal across both subgroups. Again, all Akzept-factors contribute significantly to the overall evaluation. The standardized regression coefficients for specific and general tests in the structural invariance model are $\gamma = 0.139$ ($p < 0.001$) for controllability, $\gamma = 0.136$ ($p = 0.001$) for absence of strain, and $\gamma = 0.350$ ($p < 0.001$) for face validity, which supports H1b.

4.2. Reliability of the questionnaire

All five scales achieve satisfactory internal consistencies with values of at least 0.70 (refer to Table 2). The two items regarding prior information had a correlation of $r = 0.38$, which shows that they cover two different aspects of prior information. One item addresses information regarding the tasks, and the other item is about information regarding the testing procedure. Applying the Spearman-Brown formula assuming a 4-item prior information scale, its estimated reliability would be $r = 0.55$.

4.3. Rating of university admission tests (H2)

Average ratings for each test are shown in Table 4 and the proportion of positive ratings is shown in Table 5. In accordance with H2a, overall evaluation, controllability, and the absence of strain were rated above the average of 3.5 and positively rated by the majority of the participants. Face validity, however, was rated below 3.5 and received positive ratings from a minority of participants. So, H2a is partly supported. Regarding H2b, this hypothesis is supported for the overall evaluation and controllability, but not for the absence of strain, which was rated below average for TM-WISO. A positive evaluation of face validity was only obtained for the subject-specific test PhaST, and, to a minor extent, for ITB-Business. The unspecific tests such as ICOS and GSAT were seen as very poor on face validity with mean scores below 3.0, and only about a quarter of participants provided good evaluations.

Concerning the additional aspects investigated, concentration and privacy were rated positively for all the tests, and again, the PhaST performed the best. Prior information was rated positively for all the tests, but there seems to be room for improvement in PhaST.

4.4. Test characteristics and situational conditions (H3–H6, Q1–Q3)

We calculated additional multiple linear regression models using the characteristics of the tests as independent and the Akzept-scales as well as the overall evaluation as dependent variables. The results are shown in Table 6. The predictors were able to explain a large amount of variance in controllability [$F_{(7, 2,045)} = 120.65$, $p < 0.001$], followed by face validity [$F_{(7, 2,045)} = 82.43$, $p < 0.001$], the absence of strain [$F_{(7, 2,045)} = 53.18$, $p < 0.001$], and overall evaluation [$F_{(7, 2,045)} = 52.99$, $p < 0.001$].

Subject specificity had a positive impact on face validity and on the overall evaluation, which supports H3. It also had a positive impact on controllability.

Test duration is negatively related to the absence of strain, as expected in H4. It is also negatively related to controllability, face validity, and overall evaluation.

TABLE 5 Proportion of positive ratings by test.

	PhaST	TM-WISO	ITB-Business	ITB-Science	GSAT	ICOS	Overall
Overall evaluation	91.8	52.3	94.7	81.3	75.0	76.8	76.2
Controllability	93.4	81.8	98.7	96.5	89.4	88.9	90.6
Face validity	70.0	35.6	48.7	43.0	27.2	24.2	37.3
Absence of strain	59.4	47.7	73.7	61.3	60.0	76.8	59.7
Concentration	96.5	88.3	96.1	87.3	93.6	92.3	92.1
Privacy	100	90.5	92.1	88.1	88.3	91.8	90.3
Prior information	53.1	67.1	63.2	80.0	57.7	61.6	63.0
n	256	264	76	395	962	99	2,171

TABLE 6 Standardized regression coefficients for multiple linear regression models with test characteristics as independent variables and Akzept-scales and overall evaluation as dependent variables.

	Overall evaluation	Controllability	Face validity	Absence of strain
Subject specificity (H3)	0.33**	0.35**	0.43**	0.07
Test duration (H4)	−0.48**	−0.40**	−0.12*	−0.21**
Prior information (H5)	0.18**	0.29**	0.27**	0.21**
Concentration (H6)	0.18**	0.27**	0.05*	0.23**
Proctoring (Q1)	−0.24**	−0.02	0.05	0.03
Fees (Q2)	−0.04	0.03	−0.16**	−0.10**
Privacy (Q3)	0.02	0.09**	0.06*	0.06*
F	52.99	120.65	82.43	53.18
p	<0.001	<0.001	<0.001	<0.001
Corr. R^2	0.15	0.29	0.22	0.15

*Significant with $p < 0.05$.
**Significant with $p < 0.01$; N = 2,052.

Prior information led to positive ratings on controllability, face validity, and the overall evaluation, which supports H5. It is also positively related to the absence of strain.

Concentration, e.g., participants' feelings to stay concentrated without distractions predicted all considered dependent variables—although only marginally for face validity—giving support to H6.

Proctoring had a negative effect on the overall evaluation but was not related to any of the three dimensions such as the absence of strain, controllability, or face validity.

Test taker fees did not affect the overall evaluation, but there are connections to face validity and the absence of strain.

Privacy ratings did not affect the overall evaluation, but they have small connections to controllability, face validity, and the absence of strain.

4.5. TM-WISO with proctoring vs. in test centers

Since the TM-WISO was the only test that was conducted both with proctoring and at the test centers, the following results will be limited to this subsample. Average ratings for overall evaluations, Akzept-scales, privacy, concentration, and prior information ratings by implementation condition (proctoring vs. test center) are shown in Table 7. Descriptively, the two implementation conditions seem to differ only regarding the overall evaluation, privacy, and concentration. We tested those differences using *t*-tests for independent samples. Because the Levene-test yielded statistically significant results for privacy, controllability, and concentration, variance homogeneity cannot be assumed, and we report the results of the Welch tests instead. There was a statistically significant difference between proctoring and test center participants in regard to the overall evaluation [$t_{(2,051)} = 7.71$, $p < 0.05$, $d = 0.97$], privacy [$t_{(2,051)}$

10.3389/feduc.2022.931841

TABLE 7 Average ratings by implementation condition.

	Overall evaluation	Controllability	Face validity	Absence of strain	Privacy	Concentration	Prior information
Proctoring ($n = 104$)	2.92 (1.05)	4.11 (1.13)	3.12 (1.17)	3.32 (1.09)	4.74 (1.35)	4.58 (1.16)	3.70 (1.35)
Test center ($n = 160$)	3.91 (1.00)	4.35 (0.92)	2.94 (1.06)	3.32 (1.07)	5.64 (0.70)	5.06 (0.99)	3.97 (1.26)
t-value	7.71*	1.82	−1.30	−0.017	6.20*	3.45*	1.63
Cohens d	0.97	0.23	−0.16	0.00	0.83	0.44	0.20

*Significant with $p < 0.05$; N = 2,052.

$= 6.20$, $p < 0.05$, $d = 0.83$], and concentration [$t_{(2,051)} = 3.45$, $p < 0.05$, $d = 0.44$]. Applicants gave a better rating to the tests at the test center.

5. Discussion

5.1. Summary of results

In this study, we investigated applicants' reactions to six scholastic aptitude tests of different degrees of specificity and under different implementation conditions as proctoring vs. test centers.

The results show that the dimensions of the Akzept-questionnaire and the additional questions predict the overall evaluation of tests by the participants.

Scholastic aptitude tests receive positive evaluations from the applicants. Applicants prefer a shorter duration and appreciate specific tests. Surprisingly, test-taker fees seem to have little impact on participants' evaluations. Regarding the situational factors, we found that prior information and concentration have a large influence on applicants' reactions. Privacy has only a slight impact on the Akzept-dimensions, which may be caused by a tendency of a ceiling effect.

Proctoring was evaluated positively, but less favorably than on-site tests. Most participants of the proctored tests felt comfortable concerning privacy and data protection, but a minority of about 10% turned out to be more skeptical.

5.2. Limitations and strengths

The featured study embraces the limitations and strengths. One strength is that the study involved different target groups and different selection situations: applicants for master's and bachelor's programs as well as applicants striving for an upgrade of their university entrance qualification. The data come from high-stakes situations. The study includes general and specific tests, admissions to private or public universities, different test durations, test taker fees, test center implementation, and proctoring. Thus, all the participants took the admission tests under quite divergent conditions. This variety may also be a limitation because it involves many confounding variables. Therefore, it is hard to claim the generalizability of certain findings.

Concerning the measurement, the Akzept-L was used, which proved to be a valid, multidimensional instrument for different contexts before. The instrument was extended by further relevant criteria. This way, a psychometrically sound and valid instrument got contemporarily adapted, thereby fulfilling the challenge to science, to facilitate an equal, uniform, and comparable research of applicants' reactions.

It is important to note that we could only test for measurement invariance of the Akzept-questionnaire across the general and specific test groups due to sample size restrictions. However, the analyses showed that strict measurement invariance can be assumed for general and specific tests, providing evidence that cross-group comparisons of the Akzept-scores are appropriate.

A further limitation concerns the difference between proctored tests and tests in test centers. The research questions regarding proctoring vs. on-site tests were examined for one test only, which, due to the pandemic situation, was carried out under exceptional conditions (e.g., requests to switch to proctored testing on short notice for applicants who wanted to take an on-site test).

Items regarding "concentration" were not formulated in a sufficiently specific manner; they did not fully relate to situational characteristics, but rather to the factors inherent to the participants and independent of the test (e.g., one could be distracted by internal factors). Above all, the findings on the variable concentration can be interpreted in different ways. Either one assumes that the limitations in concentration are the reason for the critical rating, or one argues that the rating of concentration is the reaction to a subjectively experienced performance failure.

In order to better classify the findings on the effect of participation fees, further information is needed, especially on the socioeconomic status of the participants. Possibly, applicants with lower financial resources avoid admission tests with large fees, and therefore, their critical attitude might not become effective.

5.3. Practical implications

How can we improve applicants' reactions to scholastic aptitude tests? Based on our findings, it is possible to draw some preliminary conclusions. Applicants of six admission tests evaluated their tests rather positively, but there is room for improvement.

First, the more specific the tests are, the better they are rated in terms of face validity and overall evaluation. Therefore, test developers could try to increase the level of specificity. In order to achieve high face validity, test items should be designed to be less abstract and include as much concrete content as possible of the subject area for which the tests are being used. Content from different areas could be replaced (e.g., items with social science content in TM-WISO, as this test is only used for business administration and economics). Alternatively, one can check whether an explanation of the relevance of the task groups or a different description of the test and task groups also improves face validity *via* prior information.

Second, applicants prefer shorter tests. Of course, it is difficult to shorten a test without impairing its reliability. Specifically, complex items that are embedded in the field or subject of study (e.g., graphs and tables)—and thus items with high face validity—need more processing time than some general mental ability tasks such as matrices. Therefore, the objective of providing shorter tests may conflict with the objective of improving face validity and/or reliability. A compromise is needed to reconcile these objectives. As the longest test in our study—PhaST—obtained top scores in face validity and overall evaluation, other factors seem to be more important than the length of the test.

Third, the role of prior information is very important for all dimensions of applicants' reactions. Therefore, test providers should offer enough preparation material and transparently inform participants about the testing conditions. This is especially important for proctoring tests, where participants need to take responsibility for creating their own test environment. PhaST, considered the "top test" regarding face validity in our sample, received lower scores on prior information by the applicants than the other tests. At the time of the study, the PhaST had only recently been developed, and in fact, only a few preparation items were available. We recommend providing many sample items for applicants' preparation.

Fourth, even though a bit trivial, not being distracted during test processing ("concentration") is an important aspect not only for applicants' reactions but also for the quality criteria of objectivity and reliability. Most applicants in our sample evaluated the possibility to work in a concentrated manner positively, but a small proportion of around 5% indicated that they were distracted. Unfortunately, our items do not reveal the source of the distraction. There may be internal sources (e.g., thinking about the pandemic) and external sources, which are difficult to control by the test provider (e.g., a relative knocking at the door during a proctoring test at home; problems with the internet connection at home) as well as external sources which could be influenced by the test provider (e.g., supervisor behavior, test center located next to a noisy street). Thus, the self-evident conclusion remains that test providers should maximize their efforts in order to provide optimal working conditions without distractions. In the case of proctored testing at home, participants should be made aware of the importance of a concentrated working atmosphere in advance.

Fifth, applicants are accepting the proctored tests. The four tests that were implemented only with proctoring obtained good overall ratings. Nevertheless, a comparison of TM-WISO participants shows that the tests conducted in test centers received a better evaluation. It is unclear whether this result can be generalized or whether it was due to the exceptional conditions (as outlined under "procedure").

What about test taker fees? Our results suggest that fees play an almost negligible role in applicants' reactions—in the range of up to 300 € at maximum. However, some questions remain open and the impact of fees should certainly be investigated in further

studies. What is an acceptable maximum fee for which group of applicants? Do we introduce social discrimination by charging fees? At which amount does discrimination start? In our study, the highest fee was charged for a selection test at a private university that also charges tuition fees; therefore, applicants may be more willing to accept also a test taker fee. Thus, the generalizability of our findings seems to be very limited. But considering that scholastic aptitude tests generally have the aim to identify the applicants with the best cognitive aptitude and not those with the highest financial resources, fees should certainly be as low as possible. Most educational organizations using these tests will want to admit the best candidates independently of their financial or social background. They can ensure social fairness by providing grants for the fees to applicants who need them.

Finally, most participants felt comfortable regarding the protection of their personal data. However, a minority of participants in proctored testing—around 10%—did not. This leads to the conclusion that test providers should not only comply with data protection regulations when offering proctored testing, but they should also provide detailed and understandable explanations about how the personal data of applicants are protected. In addition, they should—if possible—offer alternative solutions, such as taking the test in a test center or on campus for the minority of applicants who do not feel comfortable with proctoring tests.

Our research shows that we have several options to further improve scholastic aptitude tests to make them even more acceptable for applicants. However, test developers and organizations using these tests should not forget that their primary objective is the correct assessment of aptitude. Therefore, the key quality criteria are still objectivity, reliability, and validity.

5.4. Further research

Future research on applicants' reactions should involve more comparisons of different variables such as the age or gender of participants. Also, the ethnical background or nationality could be considered. It is important to know whether applicants' reactions are moderated by these variables, in order to assure fairness for different groups.

All applicants' reactions given in this section were examined before the participants received their test results. Future research should be conducted to examine the reactions after the feedback of results, as they can affect applicants' reactions, as Kersting et al. (2019) have shown for assessment centers.

All of the six tests were conducted in 2021 under special restrictions related to the pandemic and governmental prescriptions. This special situation is likely to affect applicants' reactions, so further studies should be made under "normal" circumstances after the pandemic.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

MD, CS, MK, and SS contributed to conception and design of the study. MD and SS organized the database. MD, DW, and SS performed statistical analyses. MD, CS, DW, and SS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Conflict of interest

SS and DW are working for the ITB Consulting GmbH, an organization that is developing and/or implementing the scholastic aptitude tests evaluated in this article. MK is a member of the scientific advisory board of the ITB Consulting GmbH. MD was employed by ifp—Institut für Personal- und Unternehmensberatung, Will & Partner GmbH & Co. KG.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anderson, N., Lievens, F., Van Dam, K., and Ryan, A. M. (2004). Future perspectives on employee selection: Key directions for future research and practice. *Appl. Psychol.* 53, 487–501.

Anderson, N., Salgado, J. F., and Hülsheger, U. R. (2010). Applicant reactions in selection: comprehensive meta-analysis into reaction generalization versus situational specificity. *Int. J. Select. Assess.* 18, 291–304. doi: 10.1111/j.1468-2389.2010.00512.x

Arvey, R. D., Strickland, W., Drauden, G., and Martin, C. (1990). Motivational components of test taking. *Person. Psychol.* 43, 695–716. doi: 10.1111/j.1744-6570.1990.tb00679.x

Becker, R., and Kolster, R. (2012). *International Student Recruitment: Policies and Developments in Selected Countries*.

Beermann, D., Kersting, M., Stegt, S. J., and Zimmerhofer, A. (2013). Vorurteile und Urteile zur Akzeptanz von Persönlichkeitsfragebögen. *Personal Quart.* 65, 41–45.

Bejar, I. I., and Blew, E. O. (1981). Grade inflation and the validity of the Scholastic Aptitude Test. *Am. Educ. Res. J.* 18, 143–156. doi: 10.3102/00028312018002143

Benit, N., and Soellner, R. (2012). Misst gut, ist gut? Vergleich eines abstrakten und eines berufsbezogenen Matrizentests. *J. Bus. Media Psychol.* 3, 22–29.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Eq. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Combs, H. M., Michael, L., Fiore, B., and Poling, D. A. (2007). Easy test or hard test, does it matter? The impact of perceived test difficulty on study time and test anxiety. *Undergrad. Res. J. Hum. Sci.* 6. Available online at: https://publications.kon.org/urc//v6/combs.html

European Convention on Human Rights EMRK (1950). *Art. 8*. Rome.

Gilliland, S. W. (1993). The perceived fairness of selection systems: an organizational justice perspective. *Acad. Manag. Rev.* 18, 694–734. doi: 10.2307/258595

Gilliland, S. W., and Steiner, D. D. (2012). "Applicant reactions to testing and selection," in *The Oxford Handbook of Personnel Assessment and Selection*, ed N. Schmitt (Oxford: Oxford University Press), 629–666. doi: 10.1093/oxfordhb/9780199732579.013.0028

Gnambs, T. (2022). The web-based assessment of mental speed: an experimental study of testing mode effects for the trail-making test. *Eur. J. Psycholog. Assess.* 2022, a000711. doi: 10.1027/1015-5759/a000711

Hausknecht, J. P., Day, D. V., and Thomas, S. C. (2004). Applicant reactions to selection procedures: an updated model and meta-analysis. *Person. Psychol.* 57, 639–683. doi: 10.1111/j.1744-6570.2004.00003.x

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., and Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psycholog. Methods* 16, 319. doi: 10.1037/a0024917

Heine, C., Briedis, K., Didi, H. J., Haase, K., and Trost, G. (2006). *Auswahl- und Eignungsfeststellungsverfahren in Deutschland und ausgewählten Ländern*. Hochschul-Informations-System (HIS). Kurzinformation A, 3.

Hell, B., and Haehnel, C. (2008). Bewerbermarketing im tertiären Bildungsbereich unter Berücksichtigung des Entscheidungsverhaltens Studieninteressierter. *Beiträge zur Hochschulforschung* 30, 8–32.

Hell, B., and Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. *Empirische Pädagogik* 19, 361–376.

Hell, B., Trapmann, S., and Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik* 21, 251–270.

Herde, C. N., Stegt, S., and Preckel, F. (2016). Auswahlverfahren für Masterstudiengänge aus Sicht von Bachelorstudierenden. *Zeitschrift für Arbeits- und Organisationspsychologie* 60, 145–161. doi: 10.1026/0932-4089/a000216

House, J. D., and Keeley, E. J. (1997). Predictive validity of college admissions test scores for American Indian students. *J. Psychol.* 131, 572–574. doi: 10.1080/00223989709603548

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Eq. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie* 42, 61–75.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie* 33, 420–433.

Kersting, M. (2010). Tests und Persönlichkeitsfragebogen in der Personalarbeit. *Personalführung* 10, 20–31.

Kersting, M., Fellner, K., Schneider-Ströer, J., and Bianucci, D. E. (2019). Assessment Center. Wie Unternehmen einen guten Eindruck bei Bewerbern hinterlassen. *Personal Manager*. 2019, 38–40.

König, C. J., Fell, C. B., Steffen, V., and Vanderveken, S. (2015). Applicant reactions are similar across countries: a refined replication with assessment center data from the European Union. *J. Person. Psychol.* 14, 213–217. doi: 10.1027/1866-5888/a000142

König, C. J., Klehe, U. C., Berchtold, M., and Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *Int. J. Select. Assess.* 18, 17–27. doi: 10.1111/j.1468-2389.2010.00485.x

Krumm, S., Hüffmeier, J., Dietz, F., Findeisen, A., and Dries, C. (2011). Towards positive test takers' reactions to cognitive ability assessments: development and initial validation of the Reasoning Ability at Work Test. *J. Bus. Media Psychol.* 2, 11–18.

Kuncel, N. R., Credé, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: a meta-analysis and review of the literature. *Rev. Educ. Res.* 75, 63–82. doi: 10.3102/00346543075001063

Liu, Y., and Hau, K. T. (2020). Measuring motivation to take low-stakes large-scale test: new model based on analyses of "Participant-Own-Defined" missingness. *Educ. Psycholog. Measur.* 80, 1115–1144. doi: 10.1177/0013164420911972

Marcus, B. (2003). Attitudes towards personnel selection methods: a partial replication and extension in a German sample. *Appl. Psychol.* 52, 515–532. doi: 10.1111/1464-0597.00149

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Eq. Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Mauger, P. A., and Kolmodin, C. A. (1975). Long-term predictive validity of the Scholastic Aptitude Test. *J. Educ. Psychol.* 67, 847. doi: 10.1037/0022-0663.67.6.847

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., and Ahmed, S. M. (2017). Applicant perspectives during selection: a review addressing "so what?," "what's new?," and "where to next?" *J. Manag.* 43, 1693–1725. doi: 10.1177/0149206316681846

McNeish, D., An, J., and Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *J. Personal. Assess.* 100, 43–52. doi: 10.1080/00223891.2017.1281286

Melchers, K. G., and Annen, H. (2010). Officer selection for the Swiss armed forces. *Swiss J. Psychol.* 69, 105–115. doi: 10.1024/1421-0185/a000012

Neuman, G. A., and Baydoun, R. (1998). An empirical examination of overt and covert integrity tests. *J. Bus. Psychol.* 13, 65–79. doi: 10.1023/A:1022971016454

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rogelberg, S. G., Fisher, G. G., Maynard, D. C., Hakel, M. D., and Horvath, M. (2001). Attitudes toward surveys: development of a measure and its relationship to respondent behavior. *Org. Res. Methods* 4, 3–25. doi: 10.1177/109442810141001

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Statist. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Ryan, A. M., and Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. *Hum. Resour. Manag. Rev.* 18, 119–132. doi: 10.1016/j.hrmr.2008.07.004

Ryan, A. M., and Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: a critical review and agenda for the future. *J. Manag.* 26, 565–606. doi: 10.1177/014920630002600308

Sanchez, R. J., Truxillo, D. M., and Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *J. Appl. Psychol.* 85, 739. doi: 10.1037/0021-9010.85.5.739

Saris, W. E., Satorra, A., and van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Struct. Eq. Model.* 16, 561–582. doi: 10.1080/10705510903203433

Sass, D. A., Schmitt, T. A., and Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. *Struct. Eq. Model.* 21, 167–180. doi: 10.1080/10705511.2014.882658

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Schuler, H. (1990). Personenauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie* 34, 184–191.

Schuler, H., and Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes - beurteilt unter dem Aspekt der sozialen Validität. *Zeitschrift für Arbeits- und Organisationspsychologie* 27, 33–44.

Schult, J., Hofmann, A., and Stegt, S. J. (2019). Leisten fachspezifische Studierfähigkeitstests im deutschsprachigen Raum eine valide Studienerfolgsprognose? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 2019, a000204. doi: 10.1026/0049-8637/a000204

Speer, A. B., King, B. S., and Grossenbacher, M. (2016). Applicant reactions as a function of test length: is there reason to fret over using longer tests? *J. Person. Psychol.* 15, a000145. doi: 10.1027/1866-5888/a000145

Stegt, S. J., Didi, H. J., Zimmerhofer, A., and Seegers, P. K. (2018). Akzeptanz von Auswahlverfahren zur Studienplatzvergabe. *Zeitschrift für Hochschulentwicklung* 13, 15–35. doi: 10.3217/zfhe-13-04/02

Stegt, S. J., and Hofmann, A. (2020). Eignungstests in der Coronapandemie: Fachspezifische Onlinetests mit Proctoring zur Auswahl von Bachelor- und Masterstudierenden. *Qualität in der Wissenschaft* 14, 84–90.

Steiner, D. D., and Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *J. Appl. Psychol.* 81, 134–141.

Truxillo, D. M., Bauer, T. N., and Garcia, A. M. (2017). "Applicant reactions to hiring procedures," in *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention*, eds H. W. Goldstein, E. D. Pulakos, J. Passmore, and C. Semedo (Chichester: Wiley Blackwell), 53–70. doi: 10.1002/9781118972472.ch4

Visser, R., and Schaap, P. (2017). Job applicants' attitudes towards cognitive ability and personality testing. *SA J. Hum. Resour. Manag.* 15, 1–11. doi: 10.4102/sajhrm.v15i0.877

Watrin, L., Geiger, M., Spengler, M., and Wilhelm, O. (2019). Forced-choice vs. Likert responses on an occupational Big Five questionnaire. *J. Individ. Differ.* 40, 134. doi: 10.1027/1614-0001/a000285