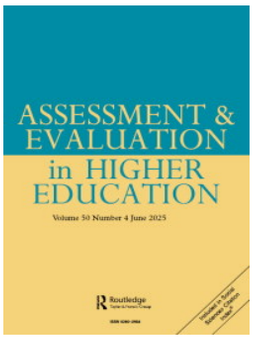

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Format-specific expectations explain preferences and perceived fairness: a study on four common examination formats

Maximilian C. Fink, Larissa J. Kaltefleiter, Isabell Reis & Bernhard Ertl

To cite this article: Maximilian C. Fink, Larissa J. Kaltefleiter, Isabell Reis & Bernhard Ertl (23 May 2025): Format-specific expectations explain preferences and perceived fairness: a study on four common examination formats, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2025.2501689

To link to this article: https://doi.org/10.1080/02602938.2025.2501689

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2025.2501689

Format-specific expectations explain preferences and perceived fairness: a study on four common examination formats

Maximilian C. Fink[a], Larissa J. Kaltefleiter[b], Isabell Reis[c] and Bernhard Ertl[a]

[a]Learning and Teaching with Media, Department of Human Sciences, Universität der Bundeswehr München, Neubiberg, Germany; [b]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany; [c]Institute of Sociology, Ludwig-Maximilians-Universität München, Munich, Germany

ABSTRACT

This study examines students' format-specific expectations and their preferences toward (1) written multiple-choice examinations, (2) written constructed-response examinations, (3) oral examinations, and (4) standardized practical examinations. $N = 509$ medical students completed a web-based survey, rating all four examination formats. Preferences and perceived fairness were assessed as primary outcomes; data on several format-specific expectations—such as objectivity and the potential to show ability—were gathered. Written multiple-choice and standardized practical examinations achieved higher preference ratings than the other examination formats. Written multiple-choice examinations attained the highest perceived fairness rating. This paper discusses the extent to which domains and culture influence the ranking of preferences and fairness of examination formats. Furthermore, multiple regressions demonstrated that format-specific expectations explained preferences and perceived fairness across all examination formats. These results indicate that students' format-specific expectations are associated with their preferences and perception of fairness in a diverse range of examination formats. To conclude, the paper highlights important research questions for future research.

KEYWORDS

Examination formats; expectations; preference; perceived fairness

Introduction

Examination formats and their use in higher education

Assessment fulfills important objectives, such as evaluating students and providing feedback to bolster their learning (Wiliam and Black 1996; Wormald et al. 2009). Moreover, assessment also contributes to the goal-oriented development of the educational system and safeguards the community from graduates who do not reach the minimum performance standards (Boulet and Durning 2019; Schleicher and Zoido 2016). A variety of examination formats have emerged and are used in different subjects and institutions to fulfill these objectives. Below, we provide an overview of various examination formats and their use in higher education.

CONTACT Maximilian Fink Maximilian.Fink@unibw.de Learning and Teaching with Media, Department of Education, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, Neubiberg 85577, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02602938.2025.2501689.

The assessment literature frequently distinguishes among written, spoken, and practical assessment types (Walzik 2012). Written assessment types contain both open and closed assignments. *Multiple choice* (MC) examinations, *constructed response* (CR) examinations, essays, and portfolios are important variations of this assessment type. Within written assessments, subordinate examination formats serve different functions: MC and CR examinations are mainly used to measure students' knowledge of concepts, theories, and mechanisms (Chamorro-Premuzic et al. 2005; Miller 1990); essays are often assigned to students when more complex cognitive processes, like argumentation, are to be tested (Brown 2010); and portfolios are mainly utilized to document active participation in the course (Reeves 2000). Spoken assessment types include *oral examinations* conducted by examiners with examinees, talks, and (group) discussions. These examination formats are mainly employed to assess the knowledge, argumentation, and presentation skills of students (Akimov and Malin 2020; Doherty et al. 2011; Kehm 2001). Practical examinations usually comprise standardized formats, such as demonstrations of observable practical skills in role-play and work-based assessment (Norcini 2003; Preston et al. 2020). Practical examinations are, thus, mainly employed for assessing diagnostic competences, physiomotor skills, and communication skills. This overview of the most important written, spoken, and practical assessment types already shows that the assessment types and their specific examination formats are used for different purposes and objectives.

Written MC examinations, written CR examinations, oral examinations, and standardized practical examinations

This study investigates examination formats in the domain of medical education. In this context, written MC examinations, written CR examinations, oral examinations, and standardized practical examinations are frequently used (Boulet and Durning 2019; Holzinger et al. 2020). These specific manifestations of examination formats are also common in teacher education, economics, and psychology (Harter, Chambers, and Asarta 2022; Kaufman and Ireland 2016; Kehm 2001; Lindner, Strobel, and Köller 2015), where they may also appear under different names. Next, we define the specific examination formats mentioned.

Written MC examinations consist of a set of test questions from which examinees select responses (Lindner, Mayntz, and Schult 2018). This examination format, of course, includes questions in which several answers can be correct. In addition, single-choice (SC) questions with only one correct answer can also be considered a form of MC examination (Epstein 2007; Lindner, Mayntz, and Schult 2018). The reason for this is that the structure of MC and SC examinations are very similar, and they target—in general—similar cognitive processes. Written MC examinations are often considered to be very objective and reliable (Lindner, Strobel, and Köller 2015). However, it can be difficult to use MC items to assess higher cognitive processes, for which careful, deliberate item construction is needed (Schuwirth and van der Vleuten 2004). Careful item construction is also necessary to avoid the common issue of having few plausible distractors available (Gierl et al. 2017).

Written CR examinations contain questions for which the examinee must write or type answers (Lindner, Mayntz, and Schult 2018). There are several types of written CR examinations, and they include responses of varying lengths (Taha 2023). In medical education in Germany, however, the licensing body typically expects short, content-related answers (Jünger and Just 2014), which resemble so-called short-answer questions. Argumentation and longer answers usually play a minor role in German written CR examinations in medical education. The answers to the written CR examinations are graded through comparison with sample solutions or are rated with descriptive scoring rubrics. Written CR examinations are thought to be particularly suitable in situations where students should be able to generate answers freely (Schuwirth and van der Vleuten 2004). They can also be valid for measuring higher cognitive processes, such as evaluating decisions

(Epstein 2007). Nevertheless, written CR examinations have also been criticized for being time-consuming to score and susceptible to potential rater biases in marking.

Oral examinations denote assessments in which the examiners ask the examinee (preformulated) test questions, and the examinee responds verbally (Kehm 2001). This term excludes talks and other spoken formats in which the examinee actively steers the examination. In the context of our study, this term does not encompass formats in which simultaneous assessment of multiple examinees occurs. Oral examinations can be used to assess higher-order cognitive processes, and examiners can provide constructive feedback (Epstein 2007). However, this type of examination format is susceptible to biases and is time-consuming to conduct for large groups (Epstein 2007).

In *standardized practical examinations* (SP examinations), the examinee must demonstrate a skill. This examination format can include (multi-station) role-play and (digital) simulations (Harden and Cairncross 1980; Jones 1998; Norcini 2003). The demonstrated skill can be cognitive, physical, or communicative. While this examination format is frequently credited with testing higher cognitive skills adequately, it can be expensive to conduct (Epstein 2007). SP examinations can also suffer from low reliability if students are not tested with a representative sample of the content. Moving on, we discuss why students' expectations toward examination formats are relevant to educational practice.

Format-specific expectations and their educational relevance

First, format-specific expectations affect the self-regulated examination preparation of students. For instance, when students perceive an examination format as having high assessment quality, they are more likely to engage in deep learning strategies that enhance their learning outcomes (Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, and Kester 2019). Additionally, when students view the format as fair, their motivation increases (Chory-Assad, 2002)—a factor that has been linked to improved long-term performance.

Second, certain format-specific expectations can contribute to adverse emotional and motivational consequences. For example, expectations regarding the difficulty and consequences of examinations—factors that have been empirically studied (e.g. Holzinger et al. 2020)—can play a role in the development of test anxiety (von der Embse et al. 2018). The connection between these expectations and negative emotional–motivational impacts can also be well explained by the expectancy-value theory of achievement emotions (Pekrun et al. 2007).

Third, students' format-specific expectations provide valuable feedback to instructors and faculty at higher education institutions. In many European countries, including Germany, the examination experience of students is often not included in evaluations (Braun, Oberschelp, and Schwabe 2020). The evaluation takes place before the examinations, so that the students' view is not influenced by examination results. Gathering insights into students' expectations toward examination formats can thus help achieve an alignment between instruction and assessment (Blair and Valdez Noel 2014) without evaluating the teacher based on their examination.

Preference and perceived fairness

The notion of *preference* stands for two interconnected aspects. Preference refers to the choosing of an entity from several alternatives (Pfister, Jungermann, and Fischer 2017), but can also denote the fondness (i.e. valence or value) ascribed to a particular entity (Lindner, Mayntz, and Schult 2018; Struyven, Dochy, and Janssens 2005). Due to individual differences, the variability of examination formats, and prior experiences, it seems likely that a significant number of university students develop preferences for specific examination formats.

In addition to preferences, perceived fairness is also an important format-specific expectation students have toward examination formats. The assessment literature contends that *perceived fairness* is related to equality, equity, and justice (Tierney 2016). *Equality* means that examinations are conducted and graded equally for all examinees. *Equity* ensures that each examinee is provided with impartial, 'fair treatment and opportunity to be successful' (Ling and Nasri 2019, 3163). *Justice*, whereas, relates to the morally correct and proportionate treatment of examinees (Tierney 2016). This conceptualization is useful for educators and researchers, but it is possible that university students—especially those less familiar with assessment concepts—may not always have such a comprehensive understanding of fairness. Instead, some students might focus on one of the three aspects—equality, equity, or justice—when reflecting on perceived fairness in examinations.

Differences in preference and perceived fairness across examination formats

Several quantitative studies have examined preferences for and the perceived fairness of examination formats. For instance, Lazarus and Kent (1983) investigated preferences for MC examinations, essays, oral examinations, and practical examinations in medical education. The 89 participating South African medical students favored practical examinations the most. Oral examinations and essays reached a relatively comparable second preference, followed by written MC examinations as the least preferred examination format. Concerning perceived fairness, practical examinations were rated as the fairest, followed by written MC examinations. Essays and oral examinations achieved comparably low perceived fairness ratings (Lazarus and Kent 1983). The study had a strong research design comparing several different examination formats. However, the data were sourced from only one institution in South Africa and were collected after the joint administration of practical examinations in one module. This raises the question of the extent to which the results also apply to higher education in today's Western countries and to practical examinations in general.

Zeidner (1987) compared in two studies the attitudes of 275 Israeli middle and high school students toward MC and essay examinations. The assessed attitude components included difficulty, complexity, value, and comfort level with the examination. In both studies, students provided more favorable ratings to written MC examinations for difficulty, complexity, and comfort level. However, only minor differences between MC and essay examinations were found in value and perceived fairness ratings (Zeidner 1987). This study elicited interesting findings on perceived fairness, which relate to two examination formats and were obtained from school students. Preferences were not measured directly, but the study contains data on the related construct of examination value.

Further, Chamorro-Premuzic et al. (2005) compared preferences for written MC examinations, essays, and oral examinations in a sample of 125 Australian university students with different majors. Written MC examinations reached a higher preference than essays. Oral examinations were the least preferred. However, the sample studied was small, and participants' majors were not specified, making it difficult to generalize the findings to other majors.

Similarly, Furnham, Batey, and Martin (2011) investigated students' preferences for written MC examinations, essays, and oral examinations in a sample of 410 social science college students from four British universities. Written MC examinations were preferred over essay examinations, while oral examinations garnered the lowest preference of these three formats (Furnham, Batey, and Martin 2011). The strengths of this study lie in the large sample size and the fact that it was conducted across several institutions. However, more practical examination formats—such as SP examinations—were not investigated.

Furthermore, Lindner, Mayntz, and Schult (2018) measured format-specific expectations toward MC and written CR examinations. Their sample consisted of 350 psychology students from Germany and Austria. Written MC and written CR examinations achieved a relatively similar

preference, but written MC examinations were perceived as fairer than written CR examinations. The study questionnaire operationalized several interesting format-specific expectations, like the potential to show ability. Moreover, a large sample consisting of students from multiple universities and two countries was examined in this study.

In another study, Holzinger et al. (2020) investigated students' preferences for written MC examinations, written CR examinations, and oral examinations in a sample of 459 Austrian medical students. Most participants favored written MC examinations; written CR and oral examinations reached lower and relatively comparable scores. However, all data in the study came from just one university. Moreover, the participants were in their second year of study and, therefore, had no extensive experience with university examinations. In addition, more practical examination formats were not surveyed.

Finally, Neto, Neto, and Furnham (2023) evaluated preferences for and the perceived fairness of written MC examinations, essays, and oral examinations. This quantitative study had a sample of 270 Portuguese psychology students. Written MC examinations were preferred over essays, and oral examinations achieved the lowest preference value. Essay examinations achieved higher perceived fairness scores than did written MC examinations, which were perceived as fairer than oral examinations (Neto, Neto, and Furnham 2023). The study also contains interesting results on preferences for group work, dissertations, and continuous assessment. However, results on the preference for and perceived fairness of practical examinations and shorter written CR examinations are not included.

Overall, the literature suggests medium to large differences in preferences and perceived fairness across examination formats. One pattern that becomes clear is that practical examinations reached high preference scores, while written MC examinations were perceived as fair in the majority of the studies reviewed.

Explaining preference and perceived fairness through format-specific expectations

Few studies have examined how format-specific expectations explain preferences for examination formats. Lindner, Mayntz, and Schult (2018) shed light on this topic for written MC and CR examinations. Their study revealed several format-specific expectations that may be critical determinants of examination preferences. More specifically, perceived fairness, anticipated low effort, and the potential to show ability were associated with preferences for written MC and CR examinations. Further, Holzinger et al. (2020) study addressed the same topic through examining medical students' preferences for written MC, written CR, and oral examinations. Judgments on satisfaction, objectivity, and expertise development due to the examination format explained the variance in the preference ratings. These findings corroborate the findings by Lindner, Mayntz, and Schult (2018) and suggest that format-specific expectations might also be related to preferences on a broader range of examination formats than written MC and CR examinations.

Likewise, the link between format-specific expectations and perceived fairness has only been investigated in a handful of studies to date. In the study by Lindner, Mayntz, and Schult (2018), students' format-specific expectations explained a small-to-medium proportion of variance in the perceived fairness ratings of the written MC and CR examinations. The variables of anticipated low effort, the potential to show ability, and objectivity were predictors for both examination formats. Moreover, the expected success of test-wiseness strategies was an additional predictor in written CR examinations. Test-wiseness strategies refer to 'a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score' (Millman, Bishop, and Ebel 1965, 707). These standalone, knowledge-independent strategies encompass tactics such as managing time effectively and eliminating incorrect responses (Lindner, Mayntz, and Schult 2018).

Two qualitative studies have also explored the association between format-specific expectations toward and the perceived fairness of examination formats. Sambell, McDowell, and Brown

(1997) conducted interviews with a small sample of participants from several domains, following a case-study approach. Based on their transcripts, they conclude that the perceived fairness of examinations could depend on the difficulty of the examinations, the transparency of demands, and the comprehensive measurement of competence (Sambell, McDowell, and Brown 1997). Though this conclusion was derived from a small number of interviews, it emphasizes the possible importance of variables similar to some of Lindner, Mayntz, and Schult (2018) predictors. Further, Murillo and Hidalgo (2017) interviewed 32 students from Spanish primary and high schools about their fairness conceptions. The students stated that perceived fairness was affected by the objectivity of examinations, transparency in marking rules, and the availability of multiple opportunities to showcase the knowledge and skills developed through instruction. Thus, the variables that the students mentioned bear similarity to the predictors of objectivity and potential to show ability from Lindner, Mayntz, and Schult (2018).

Overall, the reviewed studies indicate that several variables—including anticipated low effort, the potential to show ability, and objectivity—determine the perceived fairness of examination formats. Conceptually related variables were relevant in contexts with different examination formats.

Research gap

The literature already contains many findings on preferences for and the perceived fairness of written examinations, including written MC and CR examinations. However, there is a lack of comparative data on oral examinations and practical examination types (Holzinger et al. 2020; Lindner, Mayntz, and Schult 2018), which are becoming increasingly popular as examination formats in many domains. Another gap in the literature concerns the analysis of predictors that explain preferences and perceived fairness. The study by Lindner, Mayntz, and Schult (2018) provides important insights into this topic, which are reported and discussed for written MC and CR examinations in a format-specific manner. In our study, we seek to explain preferences and perceived fairness more universally across multiple examination formats through applying a parallel-slopes regression. This analysis will allow us to report the average relationships between the predictors and the outcomes across several examination formats.

Research questions and hypotheses

RQ1: To what extent do students' preference and perceived fairness ratings differ with regard to the examination formats of written MC examinations, written CR examinations, SP examinations, and oral examinations? We expected at least medium differences ($\eta^2 \geq 0.13$) in preferences across the examination formats (H1.1). Moreover, we suspected that there would be at least medium differences ($\eta^2 \geq 0.13$) in perceived fairness across the examination formats (H1.2).

RQ2: To what extent do format-specific expectations explain preferences for examination formats? We hypothesized that anticipated low effort, potential to show ability, and perceived fairness would be important predictors ($\beta \geq 0.10$) of preferences for examination formats (H2.1–H2.3).

RQ3: To what extent do format-specific expectations explain the perceived fairness of examination formats? We assumed that anticipated low effort, potential to show ability, and objectivity would be important predictors ($\beta \geq 0.10$) of perceived fairness for examination formats (H3.1–H3.3).

The thresholds used in the hypotheses above are based on established guidelines for effect sizes in variance analyses (Cohen 1988, 1992) and regression analyses (Fey, Hu, and Delios 2023).

Materials and methods

Data collection, sample, and additional analyses

Data were collected from medical students at the local and national levels. Locally, the study was aimed at medical students from two universities, LMU Munich and TU Munich, in southern Germany. We also included medical students from other German universities to cross-validate our data at the national level. Only medical students from the fifth semester onward were included—to ensure that the participants had sufficient experience with all examination formats. Due to extensive advertising in our own events and via cooperation partners, the sample comprised—to a large extent—local participants from LMU Munich and TU Munich (380 students, 75%) and, to a lesser extent, participants from other German universities (129 students, 25%).

In Germany, about 70% of medical universities—including LMU Munich and TU Munich—follow a standard curriculum in which theoretical content is taught first and practical learning takes place later. This curriculum includes direct instruction, seminars, and practical exercises in separate courses. The other 30% of all medical universities follow a model curriculum that combines theoretical and practical content from the beginning of the degree course. For example, problem-oriented learning and clinical practice are integrated into classes conveying theoretical concepts. Participants from foreign universities were not included in the sample.

Advertisements were posted online on social networks and in the campuses of LMU Munich and TU Munich. Participation in the study was voluntary, and all participants gave informed consent. Moreover, the study adhered to the National Ethical Review Board and university regulations for ethical research. Data protection regulations were followed strictly.

Data collection took place from February 2020 until May 2020. The time frame was the same for the data collection at LMU Munich, TU Munich, and all other German universities. During the data collection period, COVID-19 spread to Germany. From mid-March onward, universities were closed, and curfews were imposed. The first relaxations in the curfews were introduced mid-May. Most German universities, including LMU Munich and TU Munich, have a semester break from mid-February to mid-April. Thus, the survey was mainly conducted during the semester break; consequently, participants had little exposure to the new digital forms of teaching and assessment that emerged as a result of the pandemic.

The final sample consisted of $N = 509$ participants. Participants were, on average, $M = 24.44$, $SD = 3.67$ years old and had been enrolled in their degree for $M = 3.86$, $SD = 0.93$ years. Concerning gender, the sample included more females (69%, $n = 353$) than males (30%, $n = 155$) and one gender-nonconforming student (<1%). The described sample is representative in age and gender of the population of German university students of medicine in semester five and above (Kassenärztliche Bundesvereinigung 2022).

Procedure and questionnaire

Data were collected electronically through an electronic survey platform. At the start of the questionnaire, participants were familiarized with written CR examinations, written MC examinations, oral examinations, and SP examinations (see Supplementary Material 1, Table 1) through a definition provided for each type of examination format. Afterward, participants filled out two questions on how frequently they had encountered the different examination formats. The first question asked how often they had experienced each examination format. The second question asked how often each examination format had been conducted digitally (e.g. with tablets or online). Both questions used a scale with answer options ranging from (1) 'never' to (5) 'very frequently'. Following the repeated-measures design of the study, participants then provided ratings of format-specific expectations for all four examination formats.

We used Lindner, Mayntz, and Schult (2018) questionnaire on format-specific expectations (Table 1). This questionnaire surveys format-specific expectations with six subscales: preference, perceived fairness, anticipation of low effort, potential to show ability, anticipated success of test-wiseness strategies, and objectivity. To evaluate the questionnaire's factor structure, we conducted a multi-group confirmatory factor analysis using the Maximum Likelihood estimator with robust (Huber-White) standard errors (MLR), treating examination format as a grouping variable (Table 2). We compared a six-factor model, distinguishing perceived fairness and objectivity as separate factors, and a five-factor model, which combined the fairness and objectivity scales into a single 'fairness and objectivity' factor due to their high correlation ($r = 0.68$; see Table 4). A chi-squared difference test indicated no significant difference in model fit between the six-factor and five-factor solutions ($\Delta\chi^2(12) = 13.05$, $p = 0.37$). A follow-up analysis showed that the Cronbach's alpha of the combined subscale 'fairness and objectivity' had a very low value of $\alpha = 0.30$ for written MC examinations. Given the non-significant difference of the chi-squared test and the low reliability reached when combining the items to one subscale, we opted for the six-factor model. This decision improved interpretability and aligns theoretically with the original questionnaire by Lindner, Mayntz, and Schult (2018). In the survey, participants rated their agreement with each item on a Likert scale with anchors ranging from (1) 'strongly disagree' to (4) 'strongly agree'. This Likert scale also included an additional answer option, 'I don't know', which was scored as a missing value. The latter answer option enabled participants lacking personal experience to skip a question. For subscales consisting of multiple items, a mean score was calculated. Although we provide an English translation below, the questionnaire was given to our German-speaking participants in the original German version. We took the English translations of the German instrument's scale names from Stadler, Kolb, and Sailer (2021). Table 1 reports the subscales' reliabilities.

Table 1. Questionnaire on students' format-specific expectations from Lindner, Mayntz, and Schult (2018) .

Subscale	Item(s)	Cronbach's alpha range for the different examination formats
Preference	I like [examination format].	$\alpha = 0.84–0.90$
	[Examination format] should be used more frequently in examinations.	
Perceived fairness	[Examination format] are fair.	–
Low effort	I consider [examination format] to be suitable for achieving a good result even with little learning effort.	$\alpha = 0.66–0.81$
	[Examination format] are associated with little effort to prepare for me.	
	[Examination format] are particularly demanding in terms of preparation (R).	
	[Examination format] allow me to learn, leaving gaps in one's knowledge.	
Show ability	[Examination format] give me the opportunity to show that I really understood the material well.	$\alpha = 0.73–0.84$
	[Examination format] give me the opportunity to show that I know more than other students.	
	[Examination format] enable me to express my knowledge precisely.	
	I consider [examination format] to be an appropriate examination format for important assessments.	
Test-wiseness strategies	Golden rules for solving [examination format] tasks can be useful even if you are completely clueless.	$\alpha = 0.77–0.85$
	If you consider a few tips for solving [examination format] tasks, you can do much better in examinations.	
Objectivity	[Examination format] are assessed very objectively.	

Note. The questionnaire above is an English translation of the survey by Lindner, Mayntz, and Schult (2018). Permission to publish this translation was granted by Hogrefe. A Likert scale with the answer options (1) 'strongly disagree' to (4) 'strongly agree' and the additional answer option (NA) 'I don't know' was used. The Cronbach's alpha values are reported in ranges, since distinct results were obtained for the different examination formats. The item marked (R) was negatively formulated and reverse-coded before statistical analyses.

Table 2. Confirmatory factor analysis.

Fit index	Six-factor model	Five-factor model	Interpretation
Chi-squared (χ^2)	672.08	686.61	Lower is better (this score is sensitive to large samples).
Degrees of freedom (DF)	256	268	Higher DF indicates a simpler model.
Comparative fit index (CFI)	0.943	0.943	≥0.95 Is considered good.
Tucker-Lewis index (TLI)	0.919	0.922	≥0.90 Is regarded as acceptable.
Root mean square error of approximation (RMSEA)	0.067	0.066	≤0.06 Is ideal, ≤0.08 is acceptable.
Standardized root mean square residual (SRMR)	0.057	0.058	≤0.08 Is considered good.
Number of parameters	220	208	More parameters indicate a more complex model, but also more flexibility.

Note. The six-factor model treats fairness and objectivity as separate factors; the five-factor model combines them. Fit indices reported here are from the robust (MLR) estimator in Lavaan.

Statistical analyses and power analyses

To address RQ1, we compared the preferences for and perceived fairness of different examination formats using repeated-measures analysis of variance (ANOVA) with examination format as the within-subjects factor. Assumption checks, including Mauchly's tests for sphericity (Supplementary Material 2, Table 1), were conducted for the repeated-measures ANOVAs. A Greenhouse–Geisser correction had to be applied for both preferences and perceived fairness (Table 3) due to a violation of the assumption of sphericity. Regarding the sample source, independent-sample *t*-tests showed no significant differences in the outcome variables between the two samples of students from LMU and TU Munich and other German universities (all $p > 0.05$). Consequently, the samples were combined for RQ1.

RQ2 and RQ3, concerning the prediction of examination preferences and perceived fairness with format-specific expectations, were examined with parallel-slopes regression analyses. This type of regression analysis was selected because it allows for examining relationships between predictors and an outcome across different levels of a categorical variable, that is, the average relationships across several examination formats. For this purpose, the four examination formats were entered as dummy variables in the regression models. A model comparison showed that our parallel-slopes regressions explained a similar amount of variance as did more complex models with interaction effects of the examination formats. For RQ2, the variance explanation with the parallel-slopes model yielded $R^2 = 0.50$, and for the interaction effects model, $R^2 = 0.51$. The analyses for RQ3 produced similar results, with $R^2 = 0.52$ for the parallel-slopes model and $R^2 = 0.54$ for the interaction-effects model. The error probability α was set to 5% for all statistical analyses, and tests were conducted two-sided. R version 4.3.1 (R Core Team 2023) was used as the software for data analysis. When specifying and interpreting effect sizes, we followed well-established guidelines from the literature (Cohen 1988, 1992; Fey, Hu, and Delios 2023).

After data collection, we performed post-hoc power analyses with G*Power (Faul et al. 2009). Concerning the differences in the preferences for and perceived fairness of examination formats (RQ1), investigated with repeated-measures ANOVAs, we discovered medium to large effect sizes ($\eta^2 = 0.23$ for preference and $\eta^2 = 0.36$ for perceived fairness). The post-hoc power analyses showed that we achieved a power of 99% with our sample. Regarding the prediction of examination preferences and perceived fairness by format-specific expectations (RQ2 & RQ3), we carried out parallel-slopes regressions. For the predictors of preference, we found significant effect sizes ranging from $\beta = 0.14$ to $\beta = 0.46$. This resulted in an observed power of above 89%. With respect to the predictors of perceived fairness, we discovered effect sizes from $\beta = 0.11$ to $\beta = 0.60$. The *post-hoc* power analysis elicited an observed power of above 70%.

Results

Experience with the examination formats

All participants in the final sample had some experience with each examination format for which they provided information. We ensured this through including only student participants from the fifth semester onward and offering the option of not providing any information on an examination format in the survey. However, to obtain an overview of their experience, we also asked additional questions about the participants' experiences with offline and digital examination formats. The ratings of experience with the examination formats showed that participants had encountered written MC examinations, SP examinations, and oral examinations relatively often. In contrast, they had encountered written CR examinations less frequently (Supplementary Material 2, Table 2). The examinations participants faced could be digital or nondigital. Only in the written MC examinations was a significant amount of testing done digitally (Supplementary Material 2, Table 3).

Differences of preference and perceived fairness across the examination formats (RQ1)

Students' preference and perceived fairness scores are compared across the four examination formats in Table 3. This table includes descriptive statistics and results from repeated-measures ANOVAs. As can be seen, there were medium to large differences across the examination formats in both preferences and perceived fairness. Thus, hypotheses H1.1 and H1.2, which propose that there are at least medium differences in the preference and perceived fairness ratings across the examination formats, are supported.

As the descriptive statistics in Table 3 indicate, written MC examinations achieved the highest preference values, while SP examinations scored second; oral examinations achieved the third highest preference, while written CR examinations were the least preferred. We further conducted post-hoc tests, using the Bonferroni method for adjustment of p-values, to investigate the individual differences across the examination formats. We first compared the participants' preference for written MC examinations with their preferences for other examination formats. Written MC and SP examinations did not differ significantly by preference ($t(471) = 0.77$, $p = 0.99$, $d = 0.04$). However, written MC examinations achieved a moderately higher preference compared to oral examinations ($t(499) = 11.20$,

Table 3. Differences in format-specific expectations across examination formats.

Variable	Written MC examinations	Written CR examinations	Oral examinations	SP examinations	F	df	p	η^2
Preference[a]	3.03 (0.80)	2.05 (0.86)	2.33 (0.93)	2.98 (0.78)	132.44	2.56, 1150.99	<0.001	0.23
Perceived fairness[a]	3.27 (0.79)	2.53 (0.72)	1.98 (0.71)	2.85 (0.77)	249.85	2.75, 1195.32	<0.001	0.36
Objectivity	3.85 (0.51)	2.30 (0.76)	1.61 (0.69)	2.70 (0.82)	914.11	3, 1323	<0.001	0.68
Low effort[a]	2.61 (0.67)	1.65 (0.51)	1.51 (0.55)	2.38 (0.73)	392.91	2.70, 1210.44	<0.001	0.47
Show ability[a]	2.22 (0.71)	2.96 (0.63)	3.28 (0.69)	2.99 (0.74)	236.08	2.74, 1250.91	<0.001	0.34
Test-wiseness strategies[a]	2.80 (0.91)	1.77 (0.73)	2.23 (0.87)	2.39 (0.89)	136.74	2.54, 950.22	<0.001	0.27

Note. Means (and standard deviations) of the four different examination formats. Repeated-measures ANOVAs for the different variables were calculated using examination format as a within-subjects factor. The effect sizes (η^2) are based on the global model tests. All reported variables used an agreement scale with anchors from (1) 'low' to (4) 'high', which also included the answer option (NA) 'I don't know'. [a]Greenhouse–Geisser corrections were applied due to significant results from Mauchly's tests for sphericity. There were no significant differences in any outcome variables reported in this table between the two samples (all $p > 0.05$), so the samples were combined for RQ1.

$p < 0.001$, $d = 0.50$) and written CR examinations ($t(483) = 14.86$, $p < 0.001$, $d = 0.68$). SP examinations had a moderately higher preference score than oral examinations ($t(470) = 13.67$, $p < 0.001$, $d = 0.63$); they were also preferred more than written CR examinations with a large effect size ($t(453) = 17.63$, $p < 0.001$, $d = 0.83$). Oral examinations outperformed the lowest-rated format of written CR examinations with a small effect size ($t(481) = 5.58$, $p < 0.001$, $d = 0.25$).

Concerning perceived fairness, written MC examinations reached the highest score, surpassing SP examinations (Table 3). SP examinations achieved a higher perceived fairness rating than did written CR examinations, which exceeded the perceived fairness scores of oral examinations. The results of the post-hoc tests using the Bonferroni method to adjust p-values are as follows. Written MC examinations outperformed SP examinations with a small effect size ($t(476) = 7.61$, $p < 0.001$, $d = 0.35$). Further, written MC examinations surpassed written CR examinations with a moderate effect size ($t(464) = 15.27$, $p < 0.001$, $d = 0.71$). In addition, written MC examinations were judged as fairer than oral examinations, reaching a large effect size ($t(504) = 25.72$, $p < 0.001$, $d = 1.14$). SP examinations were perceived as fairer than written CR examinations, with a small effect size ($t(437) = 7.24$, $p < 0.001$, $d = 0.35$). Moreover, SP examinations were rated as fairer than oral examinations, demonstrating a large effect size ($t(474) = 21.02$, $p < 0.001$, $d = 0.96$). Written CR examinations outperformed the lowest-rated oral examinations with a moderate effect size ($t(463) = 13.12$, $p < 0.001$, $d = 0.61$).

The comparison of examination formats yielded significant differences in objectivity, low effort, the potential to show ability, and test-wiseness strategies (all p-values <0.001, all η^2 ≥0.27 in the global model test; see Table 3). These variables were primarily examined as predictors in this study, which is why the differences are not discussed here in more detail. *Post-hoc* t-tests on differences between these variables are reported in Supplementary Material 2, Tables 4–7.

Explanation of preference and fairness with format-specific expectations (RQ2 and RQ3)

Bivariate correlations (RQ2 & RQ3)

We calculated and inspected bivariate correlations across format-specific expectations, perceived fairness, and preference before fitting regression models for RQ2 and RQ3 (Table 4). Preference was positively related to low effort, the potential to show ability, and test-wiseness strategies. Moreover, preference and perceived fairness correlated positively with each other. Perceived fairness was found to have positive correlations with low effort and objectivity, but no association with the potential to show ability.

Table 4. Correlations among expectations, perceived fairness, and preference across the four examination formats.

Variables	1.	2.	3.	4.	5.
1. Low effort	—				
2. Show ability	−0.29***	—			
3. Test-wiseness strategies	0.41***	−0.19***	—		
4. Objectivity	0.49***	−0.31***	0.24***	—	
5. Perceived fairness	0.39***	−0.02	0.13***	0.68***	—
6. Preference	0.35***	0.26***	0.16***	0.40***	0.51***

Note. Two-tailed Pearson correlations of the variables. All reported variables used an agreement scale with anchors from (1) 'low' to (4) 'high'. ***$p < 0.001$.

Explanation of preferences for examination formats with format-specific expectations (RQ2)

We conducted a parallel-slopes regression analysis to examine the extent to which format-specific expectations explained preferences for examination formats (RQ2). The overall model was significant and explained a substantial share of variance, $F(9, 1750) = 190.00$, $p < 0.001$, $R^2 = 0.50$ (Table 5). Low effort, the potential to show ability, and perceived fairness were identified as significant predictors. As all three predictors also explained at least a small but meaningful amount of variance ($\beta \geq 0.10$), our hypotheses H2.1–H2.3 are supported. The variable objectivity, for which we had not formulated a hypothesis, was another significant predictor. Please note that the sample source was included in the regression analysis in Table 5 as an additional dummy variable, modeling whether participants came from Munich universities or other German universities. This variable was found not significant, suggesting that there were no differences between the results of the different samples.

The results of additional separate regression analyses for the different examination formats are available in Supplementary Material 3, Tables 1–4. All four regression models for the different examination formats demonstrated a substantial proportion of explained variance ($R^2 \geq 0.36$). The potential to show ability and perceived fairness were significant predictors in all examination formats. Low effort was a predictor in all examination formats except for SP examinations. Overall, the reported results align with the results reported for the parallel-slopes regression across all examination formats (Table 5).

Explanation of perceived fairness of examination formats with format-specific expectations (RQ3)

We conducted a parallel-slopes regression analysis to examine the extent to which format-specific expectations explained the perceived fairness of examination formats (RQ3). The overall model was significant and explained a substantial proportion of variance, $F(8, 1768) = 240.80$, $p < 0.001$, $R^2 = 0.52$ (Table 6). Low effort, the potential to show ability, and objectivity were significant predictors. Because all three predictors had at least a small, meaningful effect size ($\beta \geq 0.10$), our hypotheses H3.1–H3.3 stand substantiated. The sample source was entered as a dummy variable in the regression (Table 6), modeling whether participants came from Munich universities or other universities. The sample source was found not significant, indicating that there were no differences between the results of the different samples.

Table 5. Multiple regression with the criterion of preference.

Variable	*b*	*SE*	*β*	*β* 95% CI	*p*
Intercept	0.12	0.13			0.354
Low effort	0.17	0.03	0.14	[0.10, 0.18]	<0.001***
Show ability	0.53	0.02	0.46	[0.42, 0.50]	<0.001***
Test-wiseness strategies	0.01	0.02	0.01	[−0.03, 0.04]	0.760
Objectivity	0.07	0.03	0.08	[0.02, 0.14]	0.006**
Perceived fairness	0.30	0.03	0.29	[0.24, 0.33]	<0.001***
Format: Written CR examinations	−0.82	0.07	−0.37	[−0.42, −0.31]	<0.001***
Format: Oral examinations	−0.50	0.08	−0.24	[−0.31, −0.17]	<0.001***
Format: SP examinations	−0.18	0.06	−0.08	[−0.13, −0.03]	<0.001***
Sample source	0.00	0.04	0.00	[−0.03, 0.04]	0.921

Note. Parallel-slopes regression analysis. Format-specific expectations and dummy variables for the examination format were entered as predictors. Written MC examinations were used as a reference level for the examination formats. The rows including the dummy variables for the examination formats (e.g. 'Format: Written CR examinations') can be interpreted as differences in the intercept relative to the reference level of written MC examinations. *b* Represents unstandardized regression weights. SE is the standard error. *β* Represents standardized regression weights. CI stands for the confidence interval.
***$p < 0.001$; **$p < 0.01$.

Table 6. Multiple regression with the criterion of perceived fairness.

Variable	*b*	*SE*	*β*	*β* 95% CI	*p*
Intercept	0.51	0.12			<0.001***
Low effort	0.13	0.02	0.11	[0.07, 0.15]	<0.001***
Show ability	0.28	0.02	0.25	[0.21, 0.29]	<0.001***
Test-wiseness strategies	−0.03	0.02	−0.03	[−0.07, 0.01]	0.097
Objectivity	0.49	0.02	0.60	[0.55, 0.65]	<0.001***
Format: Written CR examinations	−0.09	0.06	−0.04	[−0.09, 0.02]	0.157
Format: Oral examinations	−0.36	0.07	−0.18	[−0.25, −0.11]	<0.001***
Format: SP examinations	−0.04	0.05	−0.02	[−0.07, 0.03]	0.405
Sample source	0.01	0.03	0.00	[−0.03, 0.04]	0.848

Note. Parallel-slopes regression analysis. Format-specific expectations and dummy variables for the examination formats were entered as predictors. Written MC examinations were used as a reference level for the examination formats. The rows including the dummy variables for the examination formats (e.g. 'Format: Written CR examinations') can be interpreted as differences in the intercept relative to the reference level of written MC examinations. *b* Represents unstandardized regression weights. SE is the standard error. *β* Represents standardized regression weights. CI stands for the confidence interval. ***$p < 0.001$.

The results of additional separate regression analyses for the different examination formats are provided in Supplementary Material 3, Tables 5–8. All these analyses demonstrated at least a moderate share of explained variance ($R^2 \geq 0.25$), and a pattern of findings relatively similar to the overall analysis results (Table 6) emerged. The potential to show ability and objectivity were predictors in all four examination formats. Low effort was only a predictor in written MC examinations and SP examinations. Altogether, the reported results largely corroborate the findings reported for the parallel-slopes regression across all examination formats (Table 6).

Discussion

Differences in preference and perceived fairness across the examination formats (RQ1)

Regarding preferences, we found medium-sized differences ($\eta^2 = 0.23$) across the examination formats, which align with our hypothesis H1.1. MC examinations were the most preferred format per the descriptive statistics. However, their contrast with SP examinations, which were also very popular, was not statistically significant. Oral examinations were significantly less preferred than these examination formats, while written CR examinations had the lowest preference overall, differing significantly from all three other formats. Combining our findings and the results from the studies described in the literature review (Chamorro-Premuzic et al. 2005; Furnham, Batey, and Martin 2011; Holzinger et al. 2020; Neto, Neto, and Furnham 2023; Lazarus and Kent 1983), a clearer picture emerges across different study subjects. MC and SP examinations seem to be somewhat favored in most study subjects. However, written CR examinations and oral examinations tend to be preferred somewhat less. These findings mainly help classify the preference for SP examinations within the different written, spoken, and practical examination formats, as little empirical evidence is available for SP examinations.

Concerning fairness, we identified substantial differences ($\eta^2 = 0.36$) among the examination formats, consistent with our hypothesis H1.2. Written MC examinations received the highest ratings, scoring significantly higher than all other formats. SP examinations ranked second, showing a significant difference from written MC examinations. Written CR examinations ranked third, differing significantly from both written MC and SP examinations. Oral examinations received the lowest ratings for fairness, with significant differences from all three other formats. From our results and those of several studies comparing the perceived fairness of different examination formats (Lazarus and Kent 1983; Lindner, Mayntz, and Schult 2018; Neto, Neto, and Furnham 2023; Zeidner 1987), it is clear that both SP and written MC examinations tend to reach higher perceived fairness values than do written CR examinations and oral examinations. One reason for

this seems to be that when open and closed written examinations were compared, higher perceived fairness scores were reported for closed than for open examinations (Lindner, Mayntz, and Schult 2018; Zeidner 1987). An exception to this pattern is evidenced in the study by Neto, Neto, and Furnham (2023), in which essay examinations, as a type of open written examination, were rated as fairer than written MC examinations.

The described results suggest trends in how most students rate their preferences for and perceived fairness of examination formats. Deriving a clear rank order valid for all contexts, institutions, and students does not seem to be possible. One reason is that preferences and perceptions of fairness may be shaped to some extent by the culture and particular domain in which students learn (Kellaghan and Madaus 2003; Stobart 2005). Another reason is that prior experiences with assessments affect preference and fairness perceptions (Struyven, Dochy, and Janssens 2005). For instance, in a survey of $N = 177$ Austrian secondary school students by Sonnleitner and Kovacs (2020), students with lower grades rated the perceived fairness of examinations as worse.

While this was not the primary focus of our study, we also investigated differences in several other format-specific expectations: objectivity, anticipation of low effort, the potential to show ability, and the use of test-wiseness strategies. Our results can be compared with the findings by Lindner, Mayntz, and Schult (2018), who examined the same variables on written MC and CR examinations. In both studies, significant differences on most of these format-specific expectations were found. However, these differences appear to be larger in our sample. We have one compelling explanation for this pattern of results: Differences in format-specific expectations could generally be larger when comparing spoken, written, and practical examinations than when comparing written examination formats with each other.

Explanation of preference and perceived fairness with format-specific expectations (RQ2 & RQ3)

Explaining preference of examination formats (RQ2)

With this study, we aimed to assess the extent to which students' format-specific expectations explained their preferences for different examination formats. Format-specific expectations explained about 25% of the variance in preferences across the examination formats. Personality variables and learning strategies, extensively studied as predictors of examination format preferences, explained substantially less variance (Chamorro-Premuzic et al. 2005; Furnham, Batey, and Martin 2011). Given these promising results, further research should be conducted on the relationship between format-specific expectations and preferences.

Anticipation of low effort, the potential to show ability, and perceived fairness were significant predictors of preference. These results are in line with our hypotheses (H2.1–H2.3) and with Lindner, Mayntz, and Schult (2018), in which the same variables were examined separately for written MC and CR examinations. They are also consistent with Holzinger et al. (2020), in which the perception of assessment was related to preferences for MC, CR, and oral examinations. Our study goes beyond these studies, as it shows that across written, spoken, and practical examination formats, format-specific expectations are associated with preferences and have meaningful effect sizes. This key finding suggests that students' format-specific expectations affect a wide range of examination format preferences and are not confined to specific formats.

In addition, we found that when objectivity was included in the regression model (Table 5), this variable also explained variance in preferences. This finding is in line with the results by Holzinger et al. (2020) and highlights that future studies should consider objectivity as a predictor of preferences.

Explaining perceived fairness of examination formats (RQ3)

Another main objective of this study was to determine the extent to which students' format-specific expectations explained their perceptions of the fairness of examination formats. We found that medical students' format-specific expectations are associated with perceived fairness across all four examination formats. This result is consistent with the findings of Lindner, Mayntz, and Schult (2018), who showed that the perceived fairness of written MC and CR examinations is related to students' format-specific expectations. Moreover, our results also add to the extant literature in showing that format-specific expectations have, on average, a notable effect on the perceived fairness of different examination formats, including written, spoken, and practical examination formats.

Anticipation of low effort, the potential to show ability, and objectivity were significant predictors in our main regression model. These results support the hypotheses H3.1–H3.3 and align well with the findings of Lindner, Mayntz, and Schult (2018). In their study, the same predictors were significant for written MC and CR examinations in the domain of psychology. Conceptually related variables were also associated with perceived fairness in qualitative studies discussed in the literature review of this paper (Murillo and Hidalgo 2017; Sambell, McDowell, and Brown 1997). These qualitative studies also suggest that transparency, appropriate demands, and difficulty—as well as objective and valid administration and grading of examinations—could be predictors of perceived fairness. Future quantitative studies should, therefore, also investigate associations with these variables—in addition to the predictors we have investigated in this study. In the following section, we explain why our results on format-specific expectations are important for educational practice in higher education settings.

Implications

Students' preferences for and expectations of examination formats are linked to important educational outcomes, as the literature shows. Positive evaluations of examination formats can contribute to a deeper learning style and increase motivation (Chory-Assad, 2002; Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, and Kester 2019). Negative evaluations of examination formats, on the other hand, can support the development of test anxiety (von der Embse et al. 2018). In addition to these findings, our study has shown that format-specific expectations, such as objectivity, are associated with the preferences for and perceived fairness of examination formats. By understanding these relationships, instructors can gain valuable insights into how the examination formats they select are connected with students' examination preparation and subjective views.

Instructors and curriculum designers can also gather feedback on their current assessment practices by administering questionnaires that gather data on preferences and expectations (Blair and Valdez Noel 2014), such as the one used in this study. This is all the more important, as often, little feedback is available on subjective examination expectations and experiences. In many European countries, including Germany, evaluations typically occur toward the end of courses, but before the final examinations take place (Braun, Oberschelp, and Schwabe 2020). The reasoning behind this is that instructors are not evaluated based on students' examination performance. However, gathering insights into students' experiences and expectations regarding assessment can help highlight how the chosen examination formats are perceived and the extent to which instruction and assessment are aligned. Instructors and curriculum planners can utilize this untapped feedback to make informed decisions and enhance future assessments in areas that have been identified as problematic.

Moreover, instructors and curriculum designers can benefit from the results of our study when selecting and implementing examination formats. Our study results show that students rate several qualities of examination formats, such as fairness, differently for different examination formats. Of course, instructors and curriculum designers must consider several factors

when selecting and implementing examination formats: Psychometric properties, learning goals, local resources and conditions, and study regulations play a central role in this context (Epstein 2007; Miller 1990; Schuwirth and van der Vleuten 2004). These sources can now be augmented with the subjective views of students when making decisions. Integrating subjective views would enable instructors and curriculum designers to better anticipate students' reactions and behaviors and to communicate more effectively about the selection and implementation of examination formats.

Limitations

One limitation of our study concerns the timing of data collection. The data were collected in the first half of 2020. As described, the participants' experiences relate primarily to the examination formats that were offered in German higher education before the coronavirus pandemic. Newer examination formats and experiences that emerged in Germany after the pandemic were thus not included in the students' expectations. However, after the pandemic, frequent doubts arose about digital examination formats, where participants could cheat more easily (Newton and Essex 2024). It can, therefore, be assumed that some of the examination formats introduced during the pandemic have been revoked (Broadbent et al. 2023) and that the important face-to-face examination formats discussed in this text will continue to be used in a similar form as before at many universities and in many subjects.

Another limitation concerns the questionnaire used to assess students' format-specific expectations (Table 1). This questionnaire, developed by Lindner, Mayntz, and Schult (2018), consists of six subscales: preference, perceived fairness, anticipation of low effort, the potential to show ability, test-wiseness strategies, and objectivity. Compared with other surveys that assess expectations and experiences (e.g. Holzinger et al. 2020; Neto, Neto, and Furnham 2023), this questionnaire provides a relatively comprehensive evaluation of examination formats. Nevertheless, the questionnaire can be further improved. In its current form, it has only one item on perceived fairness and one item on objectivity. Future researchers could systematically extend these scales to ensure that all test-takers share a common understanding of these concepts and that multifaceted constructs can be assessed more comprehensively. We also suggest extending the perceived fairness items with items measuring equality, equity, and justice. Further, the objectivity scale should be augmented with items including objectivity of measurement, objectivity of evaluation, and objectivity of interpretation. Afterward, a confirmatory factor analysis should be conducted to determine whether the six-factor solution proposed in this study remains valid with the new items.

Generalizability of findings

We examined two types of research questions. RQ1 investigated differences in preferences for and perceived fairness of four frequently used examination formats. Although this research question has been addressed extensively in the literature, the results appear relatively challenging to generalize. Cultural influences, the types of examination formats compared, and domain effects could affect preference and perceived fairness ratings (Kellaghan and Madaus 2003; Stobart 2005). In our field of study, one important point needs to be emphasized: Medical students in general—and especially in Germany with its highly competitive medical school admission process—are people who have done well on examinations in the past. These people may have a different perception of fairness and objectivity, and their use of test-wiseness strategies may also differ from those of other university students. Consequently, we believe that our findings on RQ1 may be especially transferable to medical students in comparable cultural settings with similarly implemented examination formats.

RQ2 and RQ3 concern the extent to which format-specific expectations explain preferences and perceived fairness across the mentioned examination formats. Variations in societal interpretations of the constructs, alongside examination format types and further contextual effects, might impact the size of the effects. Nevertheless, we believe that the described relationships between format-specific expectations and preference and perceived fairness ratings are relatively universal. They should be replicable in different subjects and across various examination formats, including new and alternative assessment methods. There is limited empirical evidence on this topic. In the study by Lindner et al. (2018), which focused on psychology students in Germany and Austria, the same predictors as in our study were associated with preferences and perceived fairness. Likewise, in the study by Holzinger et al. (2020), which examined medical students from one university in Austria, examination qualities—such as the perceived difficulty of examination formats—were found to be associated with preference ratings. These prior studies thus point to a certain level of generalizability of the results for RQ2 and RQ3 to other subjects and contexts.

Avenues for future research

Neither this study nor most of the reviewed literatures have examined *why* students prefer some examination formats more than others. Future qualitative studies could investigate this research question in more detail. The correlations and relationships found in this study provide some clues, but qualitative surveys and analyses could provide deeper insights into students' explicit reasons for their preferences and format-specific expectations. Possible survey methods include web surveys with open-ended questions, structured individual interviews, and focus group discussions. Methods such as inductive and deductive category development (Mayring 2000) or grounded theory (Charmaz and Belgrave 2015) could be selected. These methods could help identify, quantify, and explicate subjective views and perceptions. It would be particularly exciting to use these methods to derive subjective explanations for preferences and format-specific expectations within a common education system, like the German education system. In this way, the context of the examinations could be considered and discussed in more detail.

Another interesting avenue of research would entail quantitative studies that compare format-specific expectations in different subjects and across countries. The questionnaire by Lindner, Mayntz, and Schult (2018, English translation in Table 1) could be used and further improved to this end. For these studies, participants with comparable experience and exposure to relatively similar examination formats should be selected. Regression analysis procedures and structural equation models could be used for analysis. Meta-analyses across different subjects and countries would also be conceivable if a questionnaire that is used by many researchers is established. Such quantitative studies could more accurately determine the extent to which format-specific expectations vary across subjects, countries, and examination formats.

Conclusion

We gathered data on students' format-specific expectations toward the four frequently used examination formats of (1) written MC examinations, (2) written CR examinations, (3) oral examinations, and (4) SP examinations. Our study contributes to a growing body of evidence on differences in format-specific expectations. Based on the literature and our findings on preferences, we infer that written MC and SP examinations are somewhat preferred by university students. Written CR examinations and oral examinations, however, tend to be preferred somewhat less. Regarding perceived fairness, the literature and our findings

suggest that SP and written MC examinations tend to reach higher scores than written CR examinations and oral examinations. The reported results should be further empirically tested and generalized with caution, because they may be influenced by domains and culture. Most importantly, this study enhances the current understanding of the average relationships among students' format-specific expectations, preferences, and perceptions of fairness. Parallel-slopes regressions demonstrated that format-specific expectations explained a substantial proportion of variance in preferences and perceived fairness across the four examination formats. This result supports the argument that students' format-specific expectations are associated with their preferences for and perceptions of fairness in written, spoken, and practical examination formats. In addition, we argue that format-specific expectations provide valuable insights into the assessment process. They can help assessors get feedback on their assessment practices, evaluate the alignment between teaching and assessment, and understand their students' subjective perceptions.

Acknowledgments

The authors thank Michael Sailer, Matthias Stadler, and Martin R. Fischer for their support in the early stages of this work. In addition, we thank Joan Little, Balakumar Ravichandran and Lukas Hart for proofreading the manuscript. The OpenAI LLMs O3-mini and GPT4-o were used for assistance in writing the R-code for the confirmatory factor analysis. The code was scrutinized thoroughly. During the preparation of this work the authors used DeepL in order to translate and improve writing of individual paragraphs. Likewise, GPT4-o from OpenAI was used to improve the formulation and grammar of individual paragraphs. Each paragraph was checked carefully, and native proofreaders also edited and proofread the full manuscript. After using the mentioned services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. We would also like to thank the research initiative INDOR at the Universität der Bundeswehr München, which contributed ideas for this article.

Author contributions

Maximilian C. Fink: Conceptualization, methodology, validation, formal analysis, investigation, data curation, writing – original draft, writing – review & editing, visualization, project administration. **Larissa Kaltefleiter**: Methodology, formal analysis, writing – review & editing. **Isabell Reis**: Methodology, investigation, writing – review & editing. **Bernhard Ertl**: Conceptualization, resources, writing – review & editing, supervision.

Ethical approval

In accordance with national regulations and professional association guidelines, no ethics application was required for the study (German Research Association 2022; Rat Rat für Sozial- und Wirtschaftsdaten 2017). The reason for this was that the study only had a survey design and was not expected to have any adverse effects on the participants.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge financial support by Universität der Bundeswehr München.

ORCID

Maximilian C. Fink http://orcid.org/0000-0002-4269-4157
Larissa J. Kaltefleiter http://orcid.org/0000-0002-0921-9047
Bernhard Ertl http://orcid.org/0000-0002-7187-9257

Data availability statement

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

References

Akimov, A., and M. Malin. 2020. "When Old Becomes New: A Case Study of Oral Examination as an Online Assessment Tool." *Assessment & Evaluation in Higher Education* 45 (8): 1205–1221. doi:10.1080/02602938.2020.1730301.

Blair, E., and K. Valdez Noel. 2014. "Improving Higher Education Practice through Student Evaluation Systems: Is the Student Voice Being Heard?" *Assessment & Evaluation in Higher Education* 39 (7): 879–894. doi:10.1080/02602938.2013.875984.

Boulet, J. R., and S. J. Durning. 2019. "What we Measure and What we Should Measure in Medical Education." *Medical Education* 53 (1): 86–94. doi:10.1111/medu.13652.

Braun, E., A. Oberschelp, and U. Schwabe. 2020. "Leistungsbewertung universitärer Lehre: Gegenwärtige Praxis in Deutschland und internationale Beispiele." In *Leistungsbewertung in wissenschaftlichen Institutionen und Universitäten: Eine mehrdimensionale Perspektive*, edited by I. M. Welpe, J. Stumpf-Wollersheim, N. Folger, and M. Prenzel, 71–107. Oldenbourg, Germany: De Gruyter.

Broadbent, J., R. Ajjawi, M. Bearman, D. Boud, and P. Dawson. 2023. "Beyond Emergency Remote Teaching: Did the Pandemic Lead to Lasting Change in University Courses?" *International Journal of Educational Technology in Higher Education* 20 (1): 58. doi:10.1186/s41239-023-00428-z.

Brown, G. T. 2010. "The Validity of Examination Essays in Higher Education: Issues and Responses." *Higher Education Quarterly* 64 (3): 276–291. doi:10.1111/j.1468-2273.2010.00460.x.

Chamorro-Premuzic, T., A. Furnham, G. Dissou, and P. Heaven. 2005. "Personality and Preference for Academic Assessment: A Study with Australian University Students." *Learning and Individual Differences* 15 (4): 247–256. doi:10.1016/j.lindif.2005.02.002.

Charmaz, K., and L. L. Belgrave. 2015. "Grounded Theory." In *The Blackwell Encyclopedia of Sociology*, edited by G. Ritzer, 1st ed. Malden, MA: Wiley-Blackwall.

Chory-Assad, R. M. 2002. "Classroom Justice: Perceptions of Fairness as a Predictor of Student Motivation, Learning, and Aggression." *Communication Quarterly* 50 (1): 58–77. doi:10.1080/01463370209385646.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY: Erlbaum.

Cohen, J. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155–159. doi:10.1037/0033-2909.112.1.155.

Doherty, C., M. Kettle, L. May, and E. Caukill. 2011. "Talking the Talk: Oracy Demands in First Year University Assessment Tasks." *Assessment in Education: Principles, Policy & Practice* 18 (1): 27–39. doi:10.1080/0969594X.2010.498775.

Epstein, R. M. 2007. "Assessment in Medical Education." *The New England Journal of Medicine* 356 (4): 387–396. doi:10.1056/NEJMra054784.

Faul, F., E. Erdfelder, A. Buchner, and A. G. Lang. 2009. "Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods* 41 (4): 1149–1160. doi:10.3758/BRM.41.4.1149.

Fey, C. F., T. Hu, and A. Delios. 2023. "The Measurement and Communication of Effect Sizes in Management Research." *Management and Organization Review* 19 (1): 176–197. doi:10.1017/mor.2022.2.

Furnham, A., M. Batey, and N. Martin. 2011. "How Would You like to be Evaluated? The Correlates of Students' Preferences for Assessment Methods." *Personality and Individual Differences* 50 (2): 259–263. doi:10.1016/j.paid.2010.09.040.

German Research Association. 2022. "Guidelines for Safeguarding Good Research Practice." Bonn, Germany: DFG. doi:10.5281/zenodo.3923602.

Gerritsen-van Leeuwenkamp, K. J., D. Joosten-ten Brinke, and L. Kester. 2019. "Students' Perceptions of Assessment Quality Related to Their Learning Approaches and Learning Outcomes." *Studies in Educational Evaluation* 63: 72–82. doi:10.1016/j.stueduc.2019.07.005.

Gierl, M. J., O. Bulut, Q. Guo, and X. Zhang. 2017. "Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review." *Review of Educational Research* 87 (6): 1082–1116. doi:10.3102/0034654317726529.

Harden, R. M., and R. G. Cairncross. 1980. "Assessment of Practical Skills: The Objective Structured Practical Examination (OSPE)." *Studies in Higher Education* 5 (2): 187–196. doi:10.1080/03075078012331377216.

Harter, C., R. G. Chambers, and C. J. Asarta. 2022. "Assessing Learning in College Economics: A Sixth National Quinquennial Survey." *Eastern Economic Journal* 48 (2): 251–266. doi:10.1057/s41302-021-00205-8.

Holzinger, A., S. Lettner, V. Steiner-Hofbauer, and M. Capan Melser. 2020. "How to Assess? Perceptions and Preferences of Undergraduate Medical Students concerning Traditional Assessment Methods." *BMC Medical Education* 20 (1): 312. doi:10.1186/s12909-020-02239-6.

Jones, K. 1998. "Simulations as Examinations." *Simulation & Gaming* 29 (3): 331–341. doi:10.1177/1046878198293010.
Jünger, J., and I. Just. 2014. "Recommendations of the German Society for Medical Education and the German Association of Medical Faculties Regarding University-Specific Assessments during the Study of Human, Dental and Veterinary Medicine." *GMS Zeitschrift für medizinische Ausbildung* 31 (3): Doc34.
Kassenärztliche Bundesvereinigung. 2022. "Berufsmonitoring Medizinstudierende 2022 – Ergebnisse einer bundesweiten Befragung." Kassenärztliche Bundesvereinigung. https://www.kbv.de/media/sp/Berufsmonitoring_Medizinstudierende_2022.pdf.
Kaufman, D., and A. Ireland. 2016. "Enhancing Teacher Education with Simulations." *TechTrends* 60 (3): 260–267. doi:10.1007/s11528-016-0049-0.
Kehm, B. M. 2001. "Oral Examinations at German Universities." *Assessment in Education: Principles, Policy & Practice* 8 (1): 25–31. doi:10.1080/09695940120033234.
Kellaghan, T., and G. Madaus. 2003. "External (Public) Examinations." In *International Handbook of Educational Evaluation*, edited by T. Kellaghan and D. L. Stufflebeam, 577–600). Dordrecht, the Netherlands: Springer.
Lazarus, J., and A. P. Kent. 1983. "Student Attitudes towards the Objective Structured Clinical Examination (OSCE) and Conventional Methods of Assessment." *Suid-Afrikaanse tydskrif vir geneeskunde* [South African Medical Journal] 64 (11): 390–394.
Lindner, M. A., S. M. Mayntz, and J. Schult. 2018. "Studentische Bewertung und Präferenz von Hochschulprüfungen mit Aufgaben im offenen und geschlossenen Antwortformat." *Zeitschrift für Pädagogische Psychologie* 32 (4): 239–248. doi:10.1024/1010-0652/a000229.
Lindner, M. A., B. Strobel, and O. Köller. 2015. "Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung." *Zeitschrift für Pädagogische Psychologie* 29 (3–4): 133–149. doi:10.1024/1010-0652/a000156.
Ling, T., and N. M. Nasri. 2019. "A Systematic Review: Issues on Equity in Education." *Creative Education* 10 (12): 3163–3174. doi:10.4236/ce.2019.1012240.
Mayring, P. 2000. "Qualitative Content Analysis." *Forum: Qualitative Social Research* 1 (2): 20.
Miller, G. E. 1990. "The Assessment of Clinical Skills/Competence/Performance." *Academic Medicine* 65 (Suppl. 9): S63–S67. doi:10.1097/00001888-199009000-00045.
Millman, J., C. H. Bishop, and R. Ebel. 1965. "An Analysis of Test-Wiseness." *Educational and Psychological Measurement* 25 (3): 707–726. doi:10.1177/001316446502500304.
Murillo, F. J., and N. Hidalgo. 2017. "Students' Conceptions about a Fair Assessment of Their Learning." *Studies in Educational Evaluation* 53: 10–16. doi:10.1016/j.stueduc.2017.01.001.
Neto, J., F. Neto, and A. Furnham. 2023. "Predictors of Students' Preferences for Assessment Methods." *Assessment & Evaluation in Higher Education* 48 (4): 556–565. doi:10.1080/02602938.2022.2087860.
Newton, P. M., and K. Essex. 2024. "How Common is Cheating in Online Exams and Did It Increase during the COVID-19 Pandemic? A Systematic Review." *Journal of Academic Ethics* 22 (2): 323–343. doi:10.1007/s10805-023-09485-5.
Norcini, J. J. 2003. "Work Based Assessment." *BMJ* 326 (7392): 753–755. doi:10.1136/bmj.326.7392.753.
Pekrun, R., A. C. Frenzel, T. Goetz, and R. P. Perry. 2007. "The Control-Value Theory of Achievement Emotions: An Integrative Approach to Emotions in Education." In *Emotion in Education*, edited by P. A. Schutz and R. Pekrun, 13–36). Burlington, MA: Academic Press.
Pfister, H.-R., H. Jungermann, and K. Fischer. 2017. *Die Psychologie der Entscheidung: Eine Einführung*. 4th ed. Berlin, Germany: Springer.
Preston, R., M. Gratani, K. Owens, P. Roche, M. Zimanyi, and B. Malau-Aduli. 2020. "Exploring the Impact of Assessment on Medical Students' Learning." *Assessment & Evaluation in Higher Education* 45 (1): 109–124. doi:10.1080/02602938.2019.1614145.
R Core Team. 2023. "R: A Language and Environment for Statistical Computing [Computer Software]." https://www.r-project.org/.
Rat für Sozial- und Wirtschaftsdaten. 2017. "Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften." https://www.konsortswd.de/wp-content/uploads/RatSWD_Output9_Forschungsethik.pdf
Reeves, T. C. 2000. "Alternative Assessment Approaches for Online Learning Environments in Higher Education." *Journal of Educational Computing Research* 23 (1): 101–111. doi:10.2190/GYMQ-78FA-WMTX-J06C.
Sambell, K., L. McDowell, and S. Brown. 1997. "But is It Fair?": An Exploratory Study of Student Perceptions of the Consequential Validity of Assessment." *Studies in Educational Evaluation* 23 (4): 349–371. doi:10.1016/S0191-491X(97)86215-3.
Schleicher, A., and P. Zoido. 2016. "The Policies That Shaped PISA, and the Policies That PISA Shaped." In *The Handbook of Global Education Policy*, edited by K. Mundy, A. Green, B. Lingard, and A. Verger, 374–384). Malden, MA: Wiley.
Schuwirth, L. W., and C. P. van der Vleuten. 2004. "Different Written Assessment Methods: What Can be Said about Their Strengths and Weaknesses?" *Medical Education* 38 (9): 974–979. doi:10.1111/j.1365-2929.2004.01916.x.

Sonnleitner, P., and C. Kovacs. 2020. "Differences between Students' and Teachers' Fairness Perceptions: Exploring the Potential of a Self-Administered Questionnaire to Improve Teachers' Assessment Practices." *Frontiers in Education* 5: 17. doi:10.3389/feduc.2020.00017.

Stadler, M., N. Kolb, and M. Sailer. 2021. "The Right Amount of Pressure: Implementing Time Pressure in Online Exams." *Distance Education* 42 (2): 219–230. doi:10.1080/01587919.2021.1911629.

Stobart, G. 2005. "Fairness in Multicultural Assessment Systems." *Assessment in Education: Principles, Policy & Practice* 12 (3): 275–287. doi:10.1080/09695940500337249.

Struyven, K., F. Dochy, and S. Janssens. 2005. "Students' Perceptions about Evaluation and Assessment in Higher Education: A Review." *Assessment & Evaluation in Higher Education* 30 (4): 325–341. doi:10.1080/02602930500099102.

Taha, M. H. 2023. "Constructed Response Items." In *Written Assessment in Medical Education*, edited by H. E. E. Gasmalla, A. A. M. Ibrahim, M. M. Wadi, and M. H. Taha, 39–48. Cham, Switzerland: Springer.

Tierney, R. D. 2016. "Fairness in Educational Assessment." In *Encyclopedia of Educational Philosophy and Theory*, edited by M. Peters. Singapore: Springer.

von der Embse, N., D. Jester, D. Roy, and J. Post. 2018. "Test Anxiety Effects, Predictors, and Correlates: A 30-Year Meta-Analytic Review." *Journal of Affective Disorders* 227: 483–493. doi:10.1016/j.jad.2017.11.048.

Walzik, S. 2012. *Kompetenzorientiert prüfen: Leistungsbewertung an der Hochschule in Theorie und Praxis*. 1st ed. Opladen, Germany: Verlag Barbara Budrich.

Wiliam, D., and P. Black. 1996. "Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment?" *British Educational Research Journal* 22 (5): 537–548. doi:10.1080/0141192960220502.

Wormald, B. W., S. Schoeman, A. Somasunderam, and M. Penn. 2009. "Assessment Drives Learning: An Unavoidable Truth?" *Anatomical Sciences Education* 2 (5): 199–204. doi:10.1002/ase.102.

Zeidner, M. 1987. "Essay versus Multiple-Choice Type Classroom Exams: The Student's Perspective." *The Journal of Educational Research* 80 (6): 352–358. doi:10.1080/00220671.1987.10885782.