Psychological Test and Assessment Modeling, 2024.66:63-91

DOI: https://doi.org/10.2440/001-0013

Unraveling Performance on Multiple-Choice and Free-Response University Exams: A Multilevel Analysis of Study Time, Lecture Attendance and Personality Traits

Tuulia M. Ortner, Verena Keneder, Sonja Breuer, Freya M. Gruber, Thomas Scherndl

Department of Psychology, University of Salzburg, Salzburg, Austria

Abstract:
Assessment methods impact student learning and performance. Various recommendations address challenges of assessment in education, emphasizing test validity and reliability, aligning with ongoing efforts in psychological assessment to prevent test bias, a concern also relevant in evaluating student learning outcomes. Examinations in education commonly use either free-response (FR) or multiple-choice (MC) response formats, each with its advantages and disadvantages. Despite frequent reports of high construct equivalence between them, certain group differences based on differing person characteristics still need to be explained. In this study, we aimed to investigate how test takers' characteristics and behavior—particularly test anxiety, risk propensity, conscientiousness, lecture attendance, and study time—impact test scores in exams with FR and MC format. Data was collected from 376 students enrolled in one of two Psychology lectures at a large Austrian University at the beginning of the semester and post-exam in a real-life setting. Multilevel analyses revealed that, overall, students achieved higher scores on FR items compared to MC items. Less test anxiety, higher conscientiousness, and more study time significantly increased student examination performance. Lecture attendance impacted performance differently according to the exam items' response format: Students who attended more lectures scored higher on the MC items compared to the FR items. Risk propensity exhibited no significant effect on exam scores. The results offer deeper insights into the nuanced interplay between academic performance, personality, and other influencing

Correspondence:

Univ. Prof. Dr. Tuulia M. Ortner, Department of Psychology, University of Salzburg. tuulia.ortner@plus.ac.at

factors with the aim of establishing more reliable and valid performance tests in the future. Limitations and implications of the results are discussed.

Keywords: evaluation methods, student evaluation, test performance, response format, personality

Almost every educational institution, from schools to universities, places significant emphasis on the ongoing evaluation of student learning outcomes and performance. These outcomes include the knowledge, skills, and attitudes of students upon successfully completing a course. Therefore, assessments of learning outcomes typically manifest as evaluations conducted at the end of a course (Suskie, 2010). Various assessment methods, such as knowledge or skills tests, essays, oral examinations, assignments, presentations, simulations, or case-based evaluations, are employed for this purpose (Flores et al., 2015; Hartel & Iwaoka, 2016; Struyven et al., 2005). Within the last few years, research on student learning outcome assessments and assessment methods has increased, especially in the field of health care education (Fielding & Regehr, 2017; Hartel & Iwaoka, 2016) and for assessments of e-learning outcomes (Pilli & Admiraal, 2017; Shute & Rahimi, 2017). The overarching goals of research in the domain of student learning outcomes included the improvement of curricula, enhancement of pedagogical practice, and, of course, the development of more valid or economic assessment methods (Fielding & Regehr, 2017; Suskie, 2007). The current study aims at increasing knowledge about assessments of learning outcomes by analyzing whether the effects of student behavior and student characteristics on exam results depend on the response format.

Several guidelines have been formulated to address issues regarding assessments in education. For example, referring to the Quality Assurance Agency in Higher Education (QAA) guidelines, Rust (2002) outlined implications for general assessment strategies in tertiary education. These QAA guidelines are designed to ensure that university students' expected workload is realistic, that the assessment system is non-threatening and non-anxiety-provoking, that the students clearly understand the assessment criteria, and that continuous assessments with formative feedback are favored over summative assessments. According to Kaefer et al. (2021), formative feedback gathers information about students´ current level of understanding, aiming to utilize this information to support them in their upcoming endeavors (e.g., constructive feedback on homework, guiding during in-class tasks). In contrast, summative assessment offers a retrospective evaluation of a student´s performance up to a specific moment (e.g., exams)—an assessment method still widely used in German-speaking countries (Falck, 2023). Moreover, the Assessment Reform Group (ARG) initiated various projects to ensure that assessment policies and practices integrate relevant research and ideas about the appropriate ways to assess school and university student outcomes. Furthermore, they outlined the importance of reliability and validity of assessments as well as research-based principles (Mansell et al., 2009). Besides these guidelines, individual recommendations have been put forth: Gardner (2012) and Brown (2005) outlined validity and reliability as important criteria concerning learning outcome assessments. To establish validity, an assessment needs to fulfil its intended purpose effectively by aligning with the learning objectives and content that students are expected to master. With regard to reliability, different assessors are expected to grade similar work equally (Brown, 2005; Gardner, 2012). Furthermore, variations in students' scores should stem from construct-relevant differences rather than construct-irrelevant factors, such as effects resulting from the person who conducted the scoring, the particular selection of items, or students' constitutions (Gardner, 2012). This aim can be met by the current efforts that are invested in the domain of psychological assessment in order to avoid test bias (see Van de Vijver, 2016). Bias that is particularly related to certain assessment methods therefore also attracts interest in the field of student learning outcome assessments. This study aims to contribute to this field by investigating the impact of different kinds of learning and the test takers' personality on free-response compared to multiple-choice test scores in university exams.

As Cohen-Schotanus (1999) asserted, assessment drives student learning. The influence of particular assessment methods may extend beyond mere evaluation, also affecting learning through the quality and effectiveness of student engagement, exam preparation, and test performance (see also Lindner et al., 2015; Stanger-Hall, 2012). This influence is evident, for example, through students' expectations and preparation for the assessment or formative feedback that learners receive after the assessment (Fielding & Regehr, 2017). As referred to, there are different ways of evaluating student performance and learning outcomes. Hartel and Iwaoka (2016) differentiated between direct and indirect assessment. Direct assessment methods include traditional tests, grades, and multiple-choice exams, whereas indirect assessment methods include journal entries, alumni surveys, and external reviews (Hartel & Iwaoka, 2016). According to Flores et al. (2015), written tests, oral presentations in groups, and team work represent the most commonly employed assessment methods today. The assessment methods of interest in the present study are direct assessment methods such as written tests with varying response formats, namely, free-response (FR) and multiple-choice (MC) examinations. The prevalence of MC tests has increased in universities in German-speaking countries in recent years due to their efficiency in addressing the growing number of tests within the bachelor-master-system (Lindner et al., 2015). The following section will delve into a discussion of the advantages and disadvantages of MC and FR examinations.

MC questions contain predetermined response options, requiring test takers to select varying numbers of correct responses (i.e., sometimes more than one), whereas FR questions require written responses (e.g., Breuer et al., 2023; Ozuru et al., 2007). Due to guessing possibilities in MC exams, items may be scored as correct even when the test taker lacked the relevant knowledge, skill, or ability assessed by the question (Kubinger & Gottschall, 2007). Careful designing of response options, such as employing a (rule-based) construction rationale, and the application of psychometrically based methods of item evaluation, such as distractor analyses (see Mittring & Rost, 2008) may decrease the impact of guessing. Nevertheless, complete elimination of guessing is impossible (see Diedenhofen & Musch, 2017; Lee et al., 2011), indicated by higher scores achieved on MC exams compared to FR tests (e.g., Breuer et al., 2023). In addition, MC scores are often perceived as providing limited information about test takers' abilities as they only require one particular mode of reaction, that is, the recognition of correct answers among a number of distractors (Ackerman & Smith, 1988; Kubinger, 2014). Conversely, exams based on FR questions require students to reproduce content and reponses independently (Ackerman & Smith, 1988; Kubinger, 2014). In general, it has been stated that MC questions require only surface learning strategies, whereas FR exams require deep learning strategies to pass the exam (Struyven et al., 2005). In fact, studies have identified differences when comparing scores from MC versus FR exam results (e.g., Breuer et al., 2020; Breuer et al., 2023; Bulut et al., 2023; Gültekin & Demirtaşlı, 2012; Rodriguez, 2003). In their meta-analyses, Breuer et al. (2023), In´nami and Koizumi (2009), and Rodriguez (2003) analyzed the construct equivalence of MC questions and FR questions. The results identified construct equivalence, for example, as a function of item design, highlighting the importance of the same item stem: When the items in the two response formats were presented with the same item stem (i.e., the question), the results showed much greater similarity than when tasks were presented with different stems. By contrast, Powell (2012) revealed higher scores on MC questions (one out of four correct) compared to FR questions, even when they were constructed with the same item stem and, therefore, assessing exactly the same content. Further research suggested that FR questions show higher reliability (e.g. Breuer et al., 2020; Birenbaum & Tatsuoka, 1987) and might be more suitable for diagnostic conclusions in the extreme ranges of performance (Lee et al., 2011; Rauch & Hartig, 2010; Schult & Sparfeldt, 2016). In their recent meta-analysis, Breuer et al. (2023) revealed that, even though positive relations emerged between scores from assessments with FR and MC response format, negative pooled effect sizes for differences between the scores strengthened the doubts about construct equivalence between the two formats. These results demonstrate the importance of considering the response formats used in student examinations.

Recent research suggests that considering personality traits, such as risk-taking (e.g., Bereby-Meyer et al., 2003; Rubio et al., 2010), conscientiousness (e.g., Lakhal et al., 2017; Lakhal et al., 2015), and test anxiety (e.g., Miesner & Maki, 2007), is crucial in relation to response formats. Feelings of test anxiety are common among students and have long been considered a serious issue (Huntley et al., 2016; McDonald, 2001; Robson et al., 2023). Test anxiety is defined as phenomenological, behavioral, and physiological reactions in situations that are characterized by concerns about potential negative consequences of failure on an exam or similar evaluative situations (Zeidner, 1998). Either before or during examination situations, individuals suffering from test anxiety may experience unease, anxiety, frustration, uncertainty (negative emotions), as well as trembling, sweating, and palpitations (physiological reactions). Cognitive impairment, desperation, self-doubt, and problems with self-confidence may also occur (Kassim et al., 2008; Metzig & Schuster, 1998; Sarason, 1984; Sellers, 2000). Several studies have documented negative relations between students' test anxiety and academic performance (see Seipp, 1991). Later studies have confirmed these results in student and scholastic contexts (Breuer et al., 2023; Cassady, 2004; Cassady & Johnson, 2002; Chapell et al., 2005; Kassim et al., 2008; Silaj et al., 2021). Feelings of worry accompanied by irrelevant thoughts, potentially disrupting task-relevant thinking (Kassim et al., 2008; Sellers, 2000), have been revealed to be negatively related with task performance and positively with cognitive impairment (Sarason, 1984; Seipp, 1991). Notably, test anxiety might impact performance even beyond a given test situation by influencing learning strategies and learning behavior (Cassady, 2004; Ortner & Caspers, 2011). With reference to MC and FR exams, Miesner and Maki (2007) identified poor academic performance in highly test-anxious college students for both item formats: MC questions and FR essay tests. However, they reported significant negative relations between test anxiety and performance only on the FR essay test but not on the MC test, indicating that the MC format might be more beneficial in the presence of test anxiety. Consequently, a pertinent question arises: To what extent does the influence of test anxiety on students' examination performance depend on the response format (MC versus FR)?

Risk propensity has also been reported to impact students' examination performance, especially in relation to MC questions (Rowley, 1974). Leigh (1999) and Zinn (2017) described risky behaviors as actions that carry potential risks or negative consequences, requiring test takers to carefully weigh them against potential positive consequences. Various methods, including self-reports (Johnson et al., 2004; Leigh, 1999) and Objective Tests (see Ortner & Proyer, 2015), are available for the assessment of risk propensity. School and university students benefit from taking risks in the classroom, such as engaging in challenging tasks with uncertain outcomes to promote increased self-esteem, resilience, and empowerment (e.g., Abercrombie et al., 2022). Encouraging educational risk-taking in the classroom enhances learning, boosts academic motivation, and elevates effort (Clifford, 1991). When faced with difficult questions in examinations, students have to decide whether to skip them or to take a risk and guess the answer. Concerning risk-taking on MC tests, examinees with higher levels of risk propensity tend to guess more often than those with lower levels of risk propensity (Alnabhan, 2002; Ben-Shakhar & Sinai, 1991; Rubio et al., 2010), potentially leading to higher scores. Therefore, it is crucial to examine how risk propensity influences student performance on exams with varying response format (MC versus FR).

Regarding further aspects of personality, numerous studies and meta-analyses have revealed positive relations between conscientiousness and academic, scholastic and professional success (e.g., Busato et al., 2000; Chamorro-Premuzic & Furnham, 2003a; Chamorro-Premuzic & Furnham, 2003b; Mammadov, 2021; O'Connor & Paunonen, 2007; Ostendorf & Angleitner, 2004; Roemer et al., 2022). Meta-analyses by Trapmann et al. (2007), Poropat (2009), Richardson et al. (2012), and—most recently—Mammadov (2021), particularly highlighted the positive relations between conscientiousness and academic grades. Conscientiousness includes properties that are needed to complete tasks and responsibilities, including discipline, efficiency, a sense of order, dependability, ambition, and perseverance (Trapmann et al., 2007).

Furthermore, individuals posessing high levels of conscientiousness are often regarded as determined, strong-minded, and single-minded (Ostendorf & Angleitner, 2004). In their review of the empirical literature, O'Connor and Paunonen (2007) identified narrow personality facets—such as achievement striving, self-discipline, and dutifulness as key components of conscientiousness—as better predictors of academic performance than broad personality factors. This conclusion arises from evidence indicating stronger relations between narrow traits and academic performance indicators, as well as their ability to explain additional variance beyond that predicted by the broad Big Five factors (see O´Connor & Paunonen, 2007). The general positive relation between conscientiousness and academic performance was often ascribed to better organizational skills concerning the development and accountable fulfilment of study plans, as well as academic success through hard work, achievement orientation, and ambition (Chamorro-Premuzic & Furnham, 2003a; Cheng & Ickes, 2009; O'Connor & Paunonen, 2007). In addition, a negative relation has been observed between conscientiousness and absenteeism from university lectures, meaning that students with higher levels of conscientiousness tend to attend lectures more regularly, as reported by Chamorro-Premuzic and Furnham (2003a). Taken together, the above studies indicate a general influence of conscientiousness on the academic performance, taking into account different kinds of evaluation methods, such as grades, FR questions, MC questions, essays, and practical tasks. Lakhal and colleagues, for example, evaluated oral and written examinations, MC tests, practical work, case studies, projects, group work, and simulations in different studies. They revealed that conscientiousness was positively related to performance on oral exams, written exams, MC tests, and practical work, but not for case studies, simulations, group work, and projects (Lakhal et al., 2017; Lakhal et al., 2015). However, none of the previous studies have directly compared the effects of students' conscientiousness on performance in items with different response formats (i.e., MC versus FR).

In addition to the influence of personality traits, the importance of study time and lecture attendance on academic performance is well established (e.g., Andrietti & Velasco, 2015; Dey, 2018; Spitzer, 2022; Zorio-Grima & Merello, 2020). Active engagement in the learning process plays a critical role in shaping exam success. Attending lectures provides students with a first-hand opportunity to absorb complex material, which promotes a deeper understanding of course material and consequently leads to higher exam performance (e.g., Andrietti & Velasco, 2015; Dey, 2018; Karnik et al., 2020; Vale et al., 2020; Zorio-Grima & Merello, 2020). At the same time, dedicating time to studying course content improves information retention and comprehension, further contributing to higher academic achievement (e.g., Andrietti & Velasco, 2015; Cole & Butler, 2020; Spitzer, 2022; Zorio-Grima & Merello, 2020). To the best of our knowledge, there is currently a gap in research exploring potential interactions between the response format of exams and the variables of study time and lecture attendance in relation to academic performance.

Objectives

While numerous studies have analyzed various factors that are relevant to academic performance in general (e.g., Busato et al., 2000; Chamorro-Premuzic & Furnham, 2003a; Chapell et al., 2005; Kassim et al., 2008; Rowley, 1974), little attention has been paid to the characteristics of examinations and evaluation methods and their interactions with aspects of students' learning behavior and personality. Therefore, we aimed to investigate the effects of test anxiety, risk propensity, conscientiousness, and its facets (i.e., achievement striving, dutifulness, self-discipline), as well as lecture attendance and study time, on the students' scores on MC versus FR exam questions.

With reference to the existing literature, our initial hypothesis proposed that (1) higher scores would be attained on MC format questions in comparison to FR format items (e.g., Breuer et al., 2023), which means that—in a multilevel model—we expected a relevant main effect for response format. Subsequently, we hypothesized that (2) lecture attendance (e.g., Dey, 2018; Karnik et al., 2020; Vale et al., 2020; Zorio-Grima & Merello, 2020) and (3) total study time (e.g., Andrietti & Velasco, 2015; Cole & Butler, 2020; Spitzer, 2022; Zorio-Grima & Merello, 2020) would exhibit positive relations with student's exam performance in both response formats. This entails expectations for significant main effects for lecture attendance and study time, but no interactions with response format. Additionally, our fourth hypothesis suggested that (4) test anxiety would negatively impact student performance, particularly on FR questions (e.g., Miesner & Maki, 2007), anticipating a main effect for test anxiety and an interaction between test anxiety and response format. Based on the results of meta-analyses (e.g., Mammadov, 2021; Richardson et al., 2012), our subsequent hypothesis posited (5) positive relations between conscientiousness and student exam performance on both MC and FR questions. Consequently, we expected a main effect for conscientiousness, but no interaction with response format. Furthermore, our final hypothesis suggested that (6) positive relations between levels of risk propensity and exam scores would be particularly evident in the MC response format (e.g., Alnabhan, 2002; Rubio et al., 2010), which implies an anticipated main effect for risk propensity and an interaction between risk propensity and response format[1].

[1] We note the constraints imposed by H0: $\rho = 0$, as its rejection may provide limited information for the research purpose.

Material and Methods

Participants and Procedure

376 German-speaking students (59 men, 229 women, 88 students did not indicate their sex) with a mean age of 21.44 years (*SD* = 4.10) were recruited from two bachelor's program lectures in Psychology (i.e., Psychological Assessment, Introduction to Methodology) at a large Austrian university. All enrolled students in these courses were invited to participate voluntarily during the respective semester. A total of 169 students from the psychological assessment lecture (out of 198 enrolled) and 207 from the introduction to methodology lecture (out of 229 enrolled) agreed to take part in the study. However, the actual sample sizes for the individual analyses were lower as the complexity of the study and the multiple measurements led to substantial amounts of missing data ($n_{\min} = 124$, $n_{\max} = 298$).

We collected data at two occasions: at the beginning of the semester (T1) and post-exam (T2). The psychological assessment lecture offered three exam dates (i.e., in the last week of the semester, three weeks later, and eight weeks after the first exam), while the introduction to methodology lecture provided two (i.e., in the last week of the semester and about eight weeks later). Post-exam data were gathered immediately following the respective final exam. The two data collection sessions took 10 to 20 minutes and included, inter alia, the collection of demographic information and data on the aspects of personality investigated in this study. We further administered a weekly online survey to collect students' self-reported lecture attendance. Moreover, we obtained students' final exam scores at the end of the semester. Both lectures´ final exams incorporated a mixture of MC and FR questions.

Students gave informed consent for their data to be used in the present study and gave permission for an external and independent person of trust to pseudonymize their data (i.e., match their final exam scores with their participant codes). Students were informed about the contents and goals of the study and were advised that they could terminate their participation in the study at any time. At the end of the study, they received "experimental participant hours" serving as essential credits for their curriculum as a form of recognition for their time and contribution.

Material

The following personality characteristics and variables were gathered by administering the German versions of questionnaires.

NEO-PI-R

The conscientiousness scale from the NEO Personality Inventory-Revised (NEO-PI-R; Ostendorf & Angleitner, 2004), a questionnaire designed to assess the Big Five personality factors, was used at T1 to assess the conscientiousness facets self-discipline, achievement striving, and dutifulness, resulting in 24 items. Participants were asked to respond to the items on a five-point Likert scale, ranging from (1) *strongly disagree* to (5) *strongly agree* (e.g., "I am working hard to reach my goals" for achievement striving). Participants' responses ranged from 2 to 5 for achievement striving (M = 3.55, SD = 0.53), self-discipline (M = 3.28, SD = 0.57), and dutifulness (M = 3.84, SD = 0.49) as well as for the total conscientiousness score (M = 3.56, SD = 0.43). In this study, the internal consistencies of the scale were sufficient in the different lectures (α = .85 in both the psychological assessment and introduction to methodology lectures) for conscientiousness in general. The internal consistencies for the facets ranged from .65 to .79 (see Table 1).

Table 1
Cronbach's Alpha Values for Data from Each Lecture and Scale

Lecture	Scale	α	*n*
Psychological assessment	NEO-PI-R (T1)		
	Conscientiousness	.85	88
	Achievement striving	.70	88
	Dutifulness	.66	88
	Self-discipline	.77	88
	TAI-G (T2): Test anxiety		
	Examination date 1	.89	88
	Examination date 2	.93	35
	Examination date 3	.91	10
	DOSPERT-ES (T1): Risk propensity	.47	88
Introduction to methodology	NEO-PI-R (T1)		
	Conscientiousness	.85	148
	Achievement striving	.72	149
	Dutifulness	.65	148
	Self-discipline	.79	148
	TAI-G (T2): Test anxiety		
	Examination date 1	.90	148
	Examination date 2	.90	15
	DOSPERT-ES (T1): Risk propensity	.74	148

Note. T1 = beginning of the semester, T2 = post-exam.

TAI-G

The short form of the test anxiety inventory (TAI-G; Wacker et al., 2008), which includes the subscales interference, worry, agitation, and lack of confidence, was employed to assess test anxiety at T1 and T2. The participants were instructed to rate how they feel in exam situations and what they think during exams (e.g., "I am concerned about my performance" for worry; "I feel uncomfortable" for agitation) on a scale ranging from (1) *almost never* to (4) *almost always*. Mean scores were calculated for all 15 items, and responses ranged from 1 to 4 (T1: *M* = 2.13, *SD* = 0.52; T2: *M* = 2.07, *SD* = 0.55), resulting in medium to high internal consistencies (.89 ≤ α ≤ .93; see Table 1). The two measures of test anxiety were positively correlated (*r* = .73, *p* = .000). However, we used test anxiety assessed at T2 in the main analysis as a measure of state test anxiety. This decision was based on the proximity of T2 to the actual testing situation, enhancing the likelihood that the assessed anxiety levels were directly related to the recent experience of taking the exam. Consequently, we considered the scores assessed at T2 to be a more reliable measure of actual state test anxiety compared to the scores assessed at T1.

DOSPERT-ES

An adjusted version of the ethics scale from the DOSPERT-G (Johnson et al., 2004) was used at T1 to measure risk propensity. Because the contexts of two items from the original form did not fully apply to students' standard experiences ("…to deduct a substantial amount of my salary from a tax declaration" and "…to have an affair with a married man or a married woman"), one item was replaced by "… to use my neighbor's unprotected wireless network without informing him/her" and one more item was added "…to guess an answer on an exam when I don't actually know the answer". Participants were instructed to rate whether they thought they would pursue the mentioned behavior on a scale ranging from (1) *very unlikely* to (5) *very likely*. Responses ranged from 1 to 4 (*M* = 2.21, *SD* = 0.55), and the internal consistencies were low to medium (.47 < α < .74) for the data collected from this sample (see Table 1).

Lecture Attendance and Individual Study Time

Student attendance in the respective lectures was assessed weekly via online questionnaires over the entire 13-week semester. Students were asked to indicate whether they had been present at that week's lecture (*yes* = 1, *no* = 0), and their summed responses ranged from 0 to 11 (*M* = 3.96, *SD* = 3.52) out of 13 lectures per course. As a control, only the students in the introduction to methodology lecture were asked to estimate their overall attendance (out of thirteen classes) on the questionnaire presented at T2. Responses also ranged from 0 to 11 lecture units per course, with higher mean attendance rates compared to the weekly measure (*M* = 8.50, *SD* = 2.82). Therefore, the two measures of attendance exhibited only a correlation of *r* = .33, *p* = .000. We opted for the weekly attendance measure because all students provided data, and relying on individuals′ immediate responses within the respective week seemed more reliable than depending on their recollection at the end of the semester. Furthermore, each participant was asked to specify the number of hours they devoted to studying for the final exam outside of lecture hours at T2. This question was presented in an open-ended format, allowing students to freely input any numerical value. Students reported studying in a range from 0 to 100 hours (*M* = 26.06, *SD* = 16.15).

Final Exam Scores

A total of 367 students provided consent to access their final exam scores from the respective lecturer, while 9 declined. Subsequently, their exam scores were matched with their participant codes. The percentage correct out of the total number of questions was calculated separately for the two response formats: FR (*M* = 74.05, *SD* = 16.84) and MC questions (*M* = 68.83, *SD* = 16.49) and for the combined score (*M* = 69.67, *SD* = 17.33; for further information see also Tables 2 and 3).

Table 2
Maximum Possible Scores for Each Lecture

		Scores		
Lecture	Examination date	Multiple choice	Open-ended	Total
Psychological assessment	1 / 2 / 3	30	26	56
Introduction to methodology	1	40	41	81
	2	39	41	80

Table 3
*Means (*M*) and Standard Deviations (*SD*) for Relevant Scales and Study Variables*

	n	*M*	*SD*
Test anxiety (T2)	298	2.07	0.55
Risk propensity	254	2.21	0.55
Conscientiousness	252	3.56	0.43
Achievement striving	255	3.55	0.53
Dutifulness	254	3.84	0.49
Self-discipline	254	3.28	0.57
Lecture attendance	290	3.96	3.52
Study time	284	26.06	16.15
Score FR (%)	286	74.05	16.84
Score MC (%)	286	68.83	16.49
Score total (%)	292	69.67	17.33

Note. T2 = post-exam; % = percentage of correct items

Statistical Analysis

Statistical analyses were computed with R (version 3.5.3; R Core Team, 2019) using the packages psych (version 1.9.12.31; Revelle, 2019), lme4 (version 1.1-21; Bates et al., 2015), lmerTest (version 3.1-1; Kuznetsova et al., 2019), and sjPlot (version 2.8.3; Lüdecke, 2020).

Our study utilized a design where scores were derived from two different lectures, resulting in nested data. To account for this nested structure, a multilevel analysis was employed to investigate the effects of test anxiety, risk propensity, lecture attendance, and study time on the scores achieved on the MC and FR items. Consequently, all variables with no meaningful zero point were *z*-standardized (i.e., the TAI-G, the NEO-PI-R, the DOSPERT-ES). These standardized variables and lecture attendance, as well as individual study time were entered as predictors of students' actual test scores (%) on the MC and FR questions (grouping variable lecture and ID).

Results

Descriptive Statistics and Sex Differences

Descriptive statistics for all relevant study variables are presented in Table 3. For some variables, significant *t*-test results indicated sex differences. In the DOSPERT-ES, women ($M = 2.17$, $SD = 0.54$) described themselves as less prone to risk-taking than men ($M = 2.40$, $SD = 0.56$), $t(58.95) = -2.43$, $p = .018$, $d = -0.42$. In addition, the analysis of the TAI-G at T2 revealed sex differences, with women ($M_{t2} = 2.15$, $SD = 0.57$) reporting significantly higher test anxiety than men ($M_{t2} = 1.83$, $SD = 0.40$), $t_{t2}(103.11) = 4.40, p = .000, d = 0.57$. The achievement striving facet of conscientiousness revealed sex differences, $t(53.07) = 2.55$, $p = .013$, $d = 0.51$, with women ($M = 3.60$, $SD = 0.50$) indicating higher achievement striving than men ($M = 3.33$, $SD = 0.64$). Similar results revealed for the facet self-discipline, $t(62.57) = 2.57$, $p = .013$, $d = 0.42$ (women: $M = 3.33$, $SD = 0.57$; men: $M = 3.10$, $SD = 0.54$). Over all facets, women ($M = 3.60$, $SD = 0.42$) described themselves as significantly more conscientious than men ($M = 3.42$, $SD = 0.44$), $t(59.15) = -2.45, p =.017, d = -0.42$. In addition, there were sex differences in study time, $t(90.89) = 2.93$, $p = .004$, $d = 0.40$, with women ($M = 26.65$, $SD = 17.11$) reporting significantly more study time than men ($M = 20.16$, $SD = 12.01$). Furthermore, we found sex differences in lecture attendance, $t(71.54) = 2.80$, $p = .007$, $d = 0.424$, with women ($M = 4.40$, $SD = 3.64$) indicating significantly higher lecture attendance than men ($M = 2.89$, $SD = 3.25$).

Correlational Analysis

Correlation coefficients between the relevant study variables are presented in Table 4. For items presented in the FR format, exam scores exhibited small relations with reported study time ($r = 0.20, p = .001$), medium sized relations with test anxiety (T2: $r = -0.27$, $p = .000$), and conscientiousness ($r = 0.27$, $p = .000$)—more specific, with achievement striving ($r = 0.23$, $p = .001$), dutifulness ($r = 0.17$, $p = .021$), and self-discipline ($r = 0.23$, $p = .002$). These results suggest that test takers achieved higher scores on the FR questions the more time they spent learning for the exam, the lower they rated their own test anxiety and the higher they rated their own conscientiousness including all facets. In contrast, for items presented in the MC format, exam scores displayed medium sized coefficients with lecture attendance ($r = 0.27$, $p = .000$), smaller relations with conscientiousness ($r = 0.19$, $p = .009$)—specifically, achievement striving ($r = 0.20, p = .004$) and dutifulness ($r = 0.16, p = .025$). These findings indicate that students achieved higher scores on the MC questions the more lectures they attended and the higher they rated their own conscientiousness, particularly their achievement striving and dutifulness. Notably, the scores on the MC portions of the exam and the scores on the FR portions were also medium sized positively related ($r = 0.45, p = .000$).

Table 4

*Pearson Correlation Coefficients (*r*) for Relevant Scales and Study Variables*

	1.	2.	3.	4.	5.	5a.	5b.	5c.	6.	7.
1. Lecture attendance										
2. Study time	-.12									
3. Test anxiety (T2)	.03	.08								
4. Risk propensity	**-.18**	.04	.03							
5. Conscientiousness	.05	**.21**	**-.32**	**-.22**						
5a. Achievement striving	.00	**.20**	**-.14**	-.10	**.80**					
5b. Dutifulness	.10	.10	**-.25**	**-.24**	**.77**	**.41**				
5c. Self-discipline	.03	**.19**	**-.36**	**-.19**	**.85**	**.52**	**.52**			
6. Score FR (%)	-.03	**.20**	**-.27**	-.00	**.27**	**.23**	**.17**	**.23**		
7. Score MC (%)	**.27**	.06	-.11	.03	**.19**	**.20**	**.16**	.10	**.45**	
8. Score total (%)	.11	.12	**-.23**	.00	**.18**	**.22**	.09	.13	**.84**	**.86**

Note. n available per variable can be found in Table 3; T2 = post-exam; % = percentage of correct items; significant correlations (*p* < .05) are highlighted in bold.

Further results revealed a small, negative relation between lecture attendance and risk propensity (*r* = -0.18, *p* = .004), indicating that students who rated themselves as more prone to take risks attended fewer lectures than their counterparts with lower risk propensity. Study time exhibited a small positive relation with conscientiousness (*r* = 0.21, *p* = .004)—specifically, with achievement striving (*r* = 0.20, *p* = .005) and self-discipline (*r* = 0.19, *p* = .009), indicating that test takers who rated themselves as highly conscientious spent more time on learning for the exam. Furthermore, test anxiety assessed post-exam displayed a medium sized negative relation with conscientiousness (*r* = -0.32, *p* = .000) and its facets: achievement striving (*r* = -0.14, *p* = .043), dutifulness (*r* = -0.25, *p* = .000), and self-discipline (*r* = -0.36, *p* = .000). This means that the higher individuals rated their test anxiety, the lower they rated their own conscientiousness—particularly their self-discipline—and vice versa. Finally, risk propensity revealed a significant negative correlation with conscientiousness (*r* = -0.22, *p* = .000)—specifically, with dutifulness (*r* = -0.24, *p* = .000) and self-discipline (*r* = -0.19, *p* = .002). This means, the higher test takers rated their own tendency to take risks, the lower they rated their conscientiousness.

Multilevel Analysis

As substantial correlations and, therefore, multicollinearity revealed between the facets of conscientiousness, we decided to include only the overall conscientiousness score into the model. We included participant ID and course as random effects, and all other predictors and their interactions with response format as fixed effects. Due to missing data on at least one predictor, the effective *n* for this analysis dropped markedly to 167. The resulting model demonstrated a marginal R^2 of.133 (indicating that 13.3 % of the variance could be explained by using only the fixed effects) and a conditional R^2 of .488 (indicating that 48.8 % of the variance could be explained by the combination of fixed and random effects).

We observed a main effect of response format (b = -8.24, p = .007), indicating that students scored higher on the FR questions than on the MC questions. Additionally, analyses revealed that students who reported higher conscientiousness (b = 2.68, p = .025, see Figure 1), less test anxiety (b = -3.73, p = .002, see Figure 2), and spending more time studying (b = 0.13, p = .044, see Figure 3) achieved significantly higher test scores. There were no further significant main effects for risk taking (b = -0.40, p = .709) or attendance (b = -0.56, p = .097). A significant interaction between attendance and response format on students' test scores (b = 1.51, p = .000) indicated that students benefitted more from attending lectures when the test questions were in the MC format compared to the FR format (see Figure 4). None of the other predictors (i.e., risk propensity, conscientiousness, test anxiety, or study time) had effects on test scores that differed significantly between the two response formats. For a detailed overview of the multilevel analyses, see also Table 5.

Figure 1
Prediction of Students' Test Scores from Their Conscientiousness Assessed with Scales from the NEO-PI-R. The Grey Area Around the Regression Line Indicates the 95 % Confidence Interval.

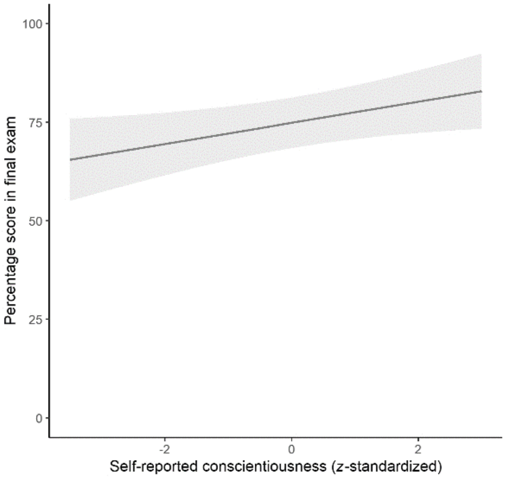

Figure 2
Prediction of Students' Test Scores from Their Test Anxiety Assessed with the TAI-G. The Grey Area Around the Regression Line Indicates the 95 % Confidence Interval.

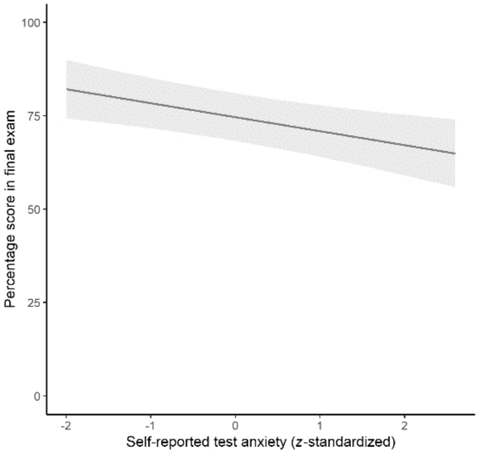

Figure 3

Prediction of Students' Test Scores from the Time They Spent Studying for the Test in Hours. The Grey Area Around the Regression Line Indicates the 95 % Confidence Interval.

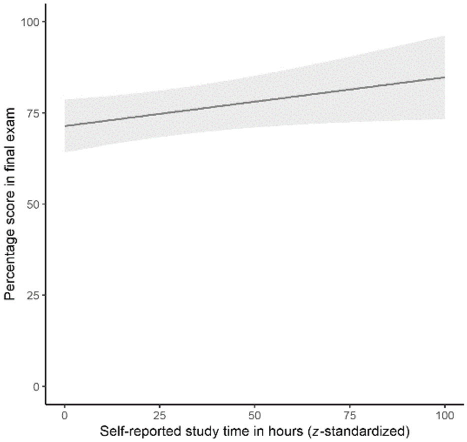

Figure 4

Prediction of Students' Test Scores from Lecture Attendance Presented Separately for Each Response Format: Free-Response (FR) and Multiple-Choice (MC) format. The Colored Areas Around the Regression Lines Indicate 95 % Confidence Intervals.

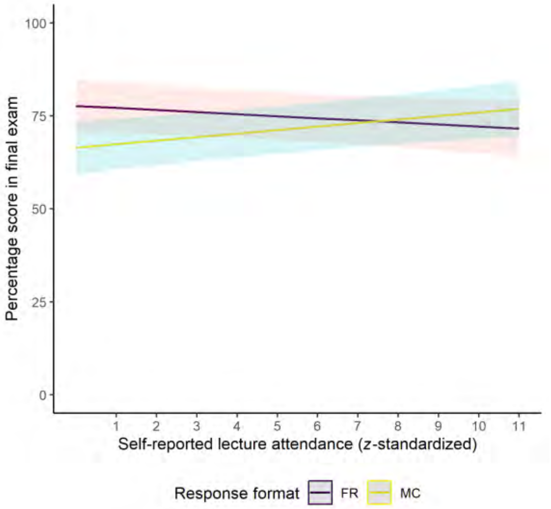

Table 5

Results of the Multilevel Analysis Predicting Test Scores

	Score				
Predictors	*b*	Beta	CI	standardized CI	*p*
(Intercept)	74.01	0.09	66.02 – 82.00	-0.33 – 0.51	<.001
Lecture attendance	-0.56	-0.13	-1.21 – 0.10	-0.27 – 0.02	.097
Response format [MC]	-8.24	-0.23	-14.28 – -2.20	-0.39 – -0.08	.007
Risk propensity	-0.40	-0.03	-2.50 – 1.70	-0.17 – 0.11	.709
Conscientiousness	2.68	0.17	0.34 – 5.03	0.02 – 0.32	.025
Test anxiety (T2)	-3.73	-0.23	-6.12 – -1.35	-0.38 – -0.08	.002
Study time	0.13	0.14	0.00 – 0.26	0.00 – 0.29	.044
Response format [MC] x Lecture attendance	1.51	0.34	0.80 – 2.21	0.18 – 0.50	<.001
Response format [MC] x Risk propensity	1.89	0.13	-0.49 – 4.27	-0.03 – 0.29	.120
Response format [MC] x Conscientiousness	-0.28	-0.02	-2.95 – 2.38	-0.19 – 0.15	.834
Response format [MC] x Test anxiety (T2)	2.02	0.12	-0.64 – 4.69	-0.04 – 0.29	.137
Response format [MC] x Study time	-0.11	-0.12	-0.26 – 0.03	-0.28 – 0.04	.134

Note. n = 354; Random effects: ICC (i.e., intra-class correlation) = .41, n_{ID} = 167, n_{Lecture} = 2; Marginal R^2 = .133, Conditional R^2 = .488; CI = 95 % confidence interval; T2 = post-exam; [MC] indicates the effect of the multiple-choice compared with the free-response format if the other predictors were held constant at their means.

Discussion

With the present research, we aimed to investigate whether the influence of lecture attendance, study time, test anxiety, conscientiousness, and risk propensity on exam performance varied depending on response format (MC vs. FR). In summary, our multilevel analysis with data from real university students revealed that, overall, students achieved higher scores on the FR items of the final exams compared to the MC questions. Test takers with higher conscientiousness, less test anxiety, and more study time were more successful on both response formats. Lecture attendance revealed as the only factor that impacted exam performance differently according to the used response format: Students who attended more lectures scored higher on the MC portions of the exam compared to the FR items. Test takers´ risk propensity exhibited no effect on their exam scores.

It was quite surprising to discover that students performed better on the FR sections of the final exams in contrast to the MC sections (see Table 3). This finding did not confirm our hypothesis and defied the common trend observed in existing literature, which typically indicates that MC items tend to yield higher scores than FR items (see, for example, the meta-analysis of Breuer et al., 2023). Considering the general perception that MC questions are typically seen as easier than FR items (e.g., In´nami &

Koizumi, 2009; McCoubrie, 2004), it is conceivable that in this case, instructors might have taken this assumption into account while designing the final exams. They may have leaned towards incorporating conceptually simpler content within the FR format or scoring those items with less rigor, aiming to counteract the prevailing perception that often deems these questions as more challenging by students.

Based on previous findings (e.g., Dey, 2018; Karnik et al., 2020; Vale et al., 2020; Zorio-Grima & Merello, 2020), we hypothesized that students who attended a greater number of lectures throughout the semester would attain higher scores on their final exams in both response formats. Interestingly, both our multilevel analysis and correlational examinations unveiled that this phenomenon was particularly pronounced in the MC portions of the exams (see Table 4, Table 5, and Figure 4). This might also lead to the hypothesis that, in these exams, instructors may have inclined towards presenting the more conceptually challenging questions in the seemingly simpler MC format. Increased lecture attendance during the semester could then have been particularly beneficial for mastering these questions compared to the FR ones, assuming the latter were less complex in content. To rule out the possibility that instructors followed this approach, the use of stem-equivalent items, as suggested by, for example, Breuer et al. (2023) and Rodriguez (2003), would be necessary in future studies examining the effects of response format. However, the lecture attendance variable in our study, with a mean of 30.5%, presents limitations due to the notably low attendance rate. This raises concerns about the generalizability of the findings and suggests potential sensitivity to variations in attendance patterns. The variable's constraints highlight the need for cautious interpretation, particularly considering that factors influencing attendance were not explored in this study.

As proposed in the existing literature (e.g., Andrietti & Velasco, 2015; Cole & Butler, 2020; Spitzer, 2022; Zorio-Grima & Merello, 2020) and in our hypotheses, overall study time was positively related to student exam performance in both response formats (see Table 5 and Figure 3). Certainly, one might wonder why the time invested in studying for the exam did not lead to the same higher positive impact on answering the MC questions compared to the FR ones, as observed with increased lecture attendance. Conversely, our correlational analysis revealed that study time was positively correlated only with FR questions, not with MC items (see Table 4). This could be explained by the possibility that the MC questions, in this case, indeed focused on inherently challenging topics, particularly emphasized during lectures. If, in fact, the FR questions primarily addressed simpler content easily answered through memorization of course materials, it would explain why this approach was more beneficial for FR questions than for MC items.

Regarding the hypothesized effects of personality variables on exam performance based on the presented response format, our multilevel analysis unveiled a significant debilitating effect of test anxiety on exam scores for both the MC and FR format (see Table 5 and Figure 2), confirming the results of Breuer et al.´s (2023) meta-analysis. The correlational examinations, on the other hand, indicated significant negative relations between test anxiety and the performance on FR format items only, while such correlations were not significant in the MC format and in the total score including the FR questions (see Table 4), as we expected in our hypothesis. These findings align, to some extent, with the outcomes reported by Miesner and Maki (2007), who specifically highlighted the negative effect of test anxiety on FR essay tests. This may be attributed to the presumption that highly test-anxious individuals experience more off-task or interfering thoughts, more task-irrelevant thinking, and a lower performance efficiency (Kassim et al., 2008; Sarason, 1984; Sellers, 2000). Such cognitive challenges arising from test anxiety could be especially detrimental in FR items, where answers must be generated without the aid of potential cues found in the response options.

Consistent with our hypothesis and supporting numerous results from previous studies (e.g., Lakhal et al., 2017; Lakhal et al., 2015; Mammadov, 2021; Richardson et al., 2012; Roemer et al., 2022; Trapmann et al., 2007), our multilevel and correlational analyses revealed a positive relation between (facets of) conscientiousness and students' exam performance on both MC and FR questions (see Table 4, Table 5, and Figure 1). The only difference observed in the impact of conscientiousness on MC versus FR items surfaced in the correlational analysis, particularly concerning the facet of self-discipline. While the students´ self-rated level of self-discipline exhibited a significant positive relation with scores in the FR section of the exam, no such significant correlation was evident in relation to the MC section (see Table 4). It is possible that self-discipline plays a more significant role or is better reflected in the type of skills required for the FR questions. Adhering to the aforementioned assumption that, in this case, FR questions focused on topics conducive to answers through memorization of course material without necessarily requiring attendance in lectures, it follows that self-discipline might have been particularly advantageous for this type of item.

In contrast to our hypothesis and to findings reported in existing literature (e.g., Alnabhan, 2002; Breuer et al., 2023; Rubio et al., 2010), our data did not reveal any significant relations between risk propensity and final exam scores (see Tables 4 and 5). Despite the commonly reported advantage for test takers with higher levels of risk-taking behavior, who are known to guess more often—particularly advantageous in MC questions (e.g., Rubio et al., 2010)—the self-reported risk propensity of the students in our sample did not relate with their exam performance. The timing of data collection may serve as an explanation here: Data were gathered after the exam, potentially influencing students' perceptions of their risk-taking tendencies. Collecting data before a high-stakes exam could have provided better insights into individuals´ baseline risk propensity and its potential impact on their test-taking strategies. Conversely, the data collected after the exam might be influenced by the testing experience, with its associated stress and challenges, leading to potential under- or overestimations of the baseline risk propensity.

Further analyses have revealed interesting results regarding sex differences and relations among the examined personality factors, along with study time and lecture attendance. Women reported significantly higher attendance rates in lectures throughout the semester and indicated spending more time preparing for exams compared to men. In line with existing literature (e.g., Cassady & Johnson, 2002; Chapell et al., 2005; Johnson et al., 2004; Nunez-Pena et al., 2016), women in our study described themselves as less prone to risk-taking, experiencing higher levels of test anxiety, and displaying greater conscientiousness than men. These findings suggest that individuals—specifically men in this case—who leaned towards taking risks and exhibited lower test anxiety and conscientiousness were more likely to skip lectures and study less for exams, contrasting with individuals—specifically women in this case—who tended to avoid risks, to fear exams, and to display high conscientiousness. This assumption was further supported by our correlational analyses, indicating that students who rated themselves as more prone to take risks attended fewer lectures and test takers who reported higher conscientiousness spent more time on preparing for the exam (see Table 4), as also proposed by Chamorro-Premuzic and Furnham (2003a). Previous studies suggested a positive relation between conscientiousness and lecture attendance (e.g., Chamorro-Premuzic & Furnham, 2003a; also see Cheng & Ickes, 2009; O’Connor & Paunonen, 2007), yet this association was not observed in our data. One plausible explanation could be that students might have attended lectures coincidentally when on campus or solely for social interactions rather than focused learning. Future research should consider controlling for such motivational factors examining the influence of conscientiousness on lecture attendance.

Risk propensity and test anxiety both exhibited an inverse relation with conscientiousness and most of its facets. This suggests that as students reported increased tendencies toward risk-taking behavior and greater fear of exams, they tended to perceive themselves as displaying lower levels of conscientiousness (i.e., dutifulness, self-discipline, and—for test anxiety only—achievement striving; see Table 4). While causality cannot be inferred here, it prompts us to wonder whether less conscientious study habits or learning behaviors might have contributed to experiencing higher test anxiety or a greater inclination for risky choices during exams among the students in this study. Given that previous research highlighted accounting modalities and scoring rules as influential factors in test takers´ risk propensity (e.g., Espinosa & Gardeazabal, 2010; Bereby-Meyer et al., 2003), future studies might benefit from including these as additional variables in multilevel analyses.

Limitations and Future Research

There are several limitations arising from the previous discussion that merit consideration. One crucial aspect to address involves the potential impact of socially desirable response tendencies, particularly concerning variables such as study time, lecture attendance, risk propensity, test anxiety, and conscientiousness. In future research, incorporating objective measures for specific variables may provide a means of evaluating them independently of subjective self-reporting. Furthermore, conscientiousness was operationalized with its facets dutifulness, achievement striving, and self-discipline. Although O’Connor and Paunonen (2007) identified exactly these three facets as the best predictors of academic performance, future studies might benefit from supplementing their analyses with the conscientiousness facets competence, order, and deliberation. Another limitation of the study is that it was conducted in real-world study settings rather than under controlled laboratory conditions. While there was no specific preparation for the exams in both courses, it cannot be ruled out that instructors might have provided hints or guidance regarding the nature and content of the exams, albeit in varying degrees. Additionally, due to the different course content, the final exams were not identical, and this discrepancy in content may have influenced the study results. Furthermore, each exam consisted of a different number of MC and FR items, leading to varying proportions of MC and FR components in the overall score. This disparity should also be considered when interpreting the results. Moreover, variations in examination modalities with reference to item difficulty and distractors could have affected the results by influencing student performance (Bereby-Meyer et al., 2003; Espinosa & Gardeazabal, 2010; Hohensinn & Kubinger, 2011; Verbić, 2012). As mentioned by Breuer et al. (2023) and Rodriguez (2003), employing items with the same stem could enhance the comparability between MC and FR items in future studies, potentially reducing construct-irrelevant variability.

Conclusion

This study aimed to investigate whether test anxiety, risk propensity, conscientiousness, lecture attendance, and study time had varying effects on academic performance based on response format (MC vs. FR). A significant strength of this research was its utilization of real-life examination settings, enabling a direct comparison of student performance between MC and FR questions. As expected, the findings indicated a negative impact of test anxiety and positive effects of conscientiousness, study time, and lecture attendance on student exam scores. Interestingly, while study time related with scores on FR items, this association was not mirrored in MC items. Test anxiety detrimentally affected both MC and FR scores but exhibited a stronger negative relation with performance on FR items, indicating a particular disadvantage for test-anxious students in exams utilizing this response format. Surprisingly, no significant relations emerged between risk propensity and exam scores, challenging prior research findings. Overall, these results offer deeper insights into the nuanced interplay between academic performance, response formats, and various influencing factors, with the hope of establishing more reliable and valid test situations for all students to demonstrate their acquired skills and knowledge without bias.

Acknowledgments

We would like to thank Jasmin Buchberger for her help collecting and entering the data. We further thank Walter Gruber and Paul Lengenfelder for supporting our study in their courses. We thank Sabine Stummvoll for serving as an independent person of trust, the students from the Department of Psychology at the University of Salzburg for their participation, and Carina Gargitter and Jane Zagorski for helpful comments on a previous version of this manuscript.

References

Abercrombie, S., Bang, H., & Vaughan, A. (2022). Motivational and disciplinary differences in academic risk taking in higher education. *Educational Psychology, 42*(7), 895–912. https://doi.org/10.1080/01443410.2022.2076810

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple choice, and free-response writing tests. *Applied Psychological Measurement, 12(2)*, 117-128. https://doi.org/10.1177/014662168801200202

Alnabhan, M. (2002). An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality, 30(7)*, 645-652. https://doi.org/10.2224/sbp.2002.30.7.645

Andrietti, V., & Velasco, C. (2015). Lecture Attendance, Study Time, and Academic Performance: A Panel Data Study. *The Journal of Economic Education, 46(3)*, 239-259. https://doi.org/10.1080/00220485.2015.1040182

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1)*, 1-48. https://doi.org/10.18637/jss.v067.i01

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28(1)*, 23-35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychologica, 112(2)*, 207-220. https://doi.org/10.1016/s0001-6918(02)00085-9

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11(4)*, 385-395. https://doi.org/10.1177/014662168701100404

Breuer, S., Scherndl, T., & Ortner, T. M. (2020). Effects of response format on psychometric properties and fairness of a matrices test: multiple choice vs. free response. *Frontiers in Education*, 5: Article 15. https://doi.org/10.3389/feduc.2020.00015

Breuer, S., Scherndl, T., & Ortner, T. M. (2023). Effects of response format on achievement and aptitude assessment results: multi-level random effects meta-analyses. *Royal Society Open Science, 10*: 220456. https.//doi.org/10.1098/rsos.220456

Brown, S. (2005). Assessment for learning. *Learning and Teaching in Higher Education, 1*, 81-89.

Bulut, H. C., Bulut, O., & Arikan, S. (2023). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing, 23(1)*, 10-33. https://doi.org/10.1080/15305058.2022.2044821

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual differences, 29*(6), 1057-1068. https://doi.org/10.1016/S0191-8869(99)00253-6

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*(6), 569-592. https://doi.org/10.1016/j.learnin-struc.2004.09.002

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270-295. https://doi.org/10.1006/ceps.2001.1094

Chamorro-Premuzic, T., & Furnham, A. (2003a). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*(4), 319-338. https://doi.org/10.1016/S0092-6566(02)00578-0

Chamorro-Premuzic, T., & Furnham, A. (2003b). Personality traits and academic examination performance. European *Journal of Personality, 17*(3), 237-250. https://doi.org/10.1002/per.473

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274. https://doi.org/10.1037/0022-0663.97.2.268

Cheng, W., & Ickes, W. (2009). Conscientiousness and self-motivation as mutually compensatory predictors of university-level GPA. *Personality and Individual Differences, 47*(8), 817-822. https://doi.org/10.1016/j.paid.2009.06.029

Clifford, M. M. (1991). Risk taking: Theoretical, empirical, and educational considerations. *Educational Psychologist, 26*(3–4), 263–297. https://doi.org/10.1207/s15326985ep2603&4_4

Cohen-Schotanus, J. (1999). Student assessment and examination rules. *Medical Teacher, 21*(3), 318-321. https://doi.org/10.1080/01421599979626

Cole, Z. J., & Butler, D. L. (2020). Disentangling the effects of study time and study strategy on undergraduate test performance. *Psi Chi Journal of Psychological Research, 25*(2), 110-120. https://doi.org/10.24839/2325-7342.JN25.2.110

Dey, I. (2018). Class attendance and academic performance: A subgroup analysis. International *Review of Economics Education, 28*, 29-40, https://doi.org/10.1016/j.iree.2018.03.003

Diedenhofen, B., & Musch, J. (2017). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment, 33*(5), 336-344. https://doi.org/10.1027/1015-5759/a000295

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*(5), 415-425. https://doi.org/10.1016/j.jmp.2010.06.001

Falck, J. (2023). *Lernförderliches Feedback im Unterricht [Learner-supportive feedback in the classroom]*. Persen.

Fielding, D. W., & Regehr, G. (2017). A call for an integrated program of assessment. American *Journal of Pharmaceutical Education, 81*(4), 1-11. https://doi.org/10.5688/ajpe81477

Flores, M. A., Veiga Simão, A. M., Barros, A., & Pereira, D. (2015). Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. *Studies in Higher Education, 40*(9), 1523-1534. https://doi.org/10.1080/03075079.2014.881348

Gardner, J. (2012). *Assessment and Learning.* SAGE Publications.

Gültekin, S., & Demirtaşlı, N. C. (2012). Comparing the test information obtained through multiple-choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online, 11*(1), 251-263.

Hartel, R. W., & Iwaoka, W. T. (2016). A report from the Higher Education Review Board (HERB): Assessment of undergraduate student learning outcomes in food science. *Journal of Food Science Education, 15*(2), 56-62. https://doi.org/10.1111/1541-4329.12084

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746. https://doi.org/10.1177/0013164410390032

Huntley, C. D., Young, B., Jha, V., & Fisher, P. L. (2016). The efficacy of interventions for test anxiety in university students: A protocol for a systematic review and meta-analysis. International *Journal of Educational Research, 77,* 92-98. https://doi.org/10.1016/j.janxdis.2019.01.007

In´nami, Y. & Koizumi, R. (2009). A meta analysis of test format effects on reading and listening test performance: focus on multiple choice and open ended formats. *Language Testing, 26*, 219-244. https://doi.org/10.1177/0265532208101006

Johnson, J. G., Wilke, A., & Weber, E. U. (2004). Beyond a trait view of risk taking: A domain-specific scale measuring risk perceptions, expected benefits, and perceived-risk attitudes in German-speaking populations. *Polish Psychological Bulletin, 35*, 153-172.

Kaefer, J., Herbein, E., & Fauth, B., (2021). *Formatives Feedback im Unterricht [Formative feedback in teaching]*. Institut für Bildungsanalysen Baden-Württemberg.

Karnik, A., Kishore, P., Meraj, M. (2020). Examining the linkage between class attendance at university and academic performance in an International Branch Campus setting. *Research in Comparative and International Education, 15:* Issue 4. https://doi.org/10.1177/1745499920958855

Kassim, M. A., Hanafi, S. R., & Hancock, D. R. (2008). Test anxiety and its consequences on academic performance among university students. In A. M. Columbus (Ed.), *Advances in Psychology Research* (pp. 75-95). Nova Science Publishers.

Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester“ Multiple-Choice-Prüfungsaufgaben. [Report on the construction of multiple choice exam questions that are valid for legal purposes]. *Psychologische Rundschau, 65*(3), 169-178. https://doi.org/10.1026/0033-3042/a000218

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats - An experiment in fundamental research on psychological assessment. *Psychology Science, 49*(4), 361-374.

Kuznetsova, A., Brockhoff, B., & Christensen, R. H. B. (2019). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lakhal, S., Frenette, E., & Sévigny, S. (2017). The effect of personality on the university students performance in business administration on written exams, multiple-choice exams and practical exercises. *Canadian Journal for the Scholarship of Teaching and Learning, 8*(1), 1-20.

Lakhal, S., Sévigny, S., & Frenette, É. (2015). Personality and student performance on evaluation methods used in business administration courses. *Educational Assessment, Evaluation and Accountability, 27*(2), 171-199. https://doi.org/10.1007/s11092-014-9200-7

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115-136. https://doi.org/10.1080/08957347.2011.554604

Leigh, B. C. (1999). Peril, chance, adventure: *Concepts of risk, alcohol use and risky behavior in young adults. Addiction, 94*(3), 371-383. https://doi.org/10.1046/j.1360-0443.1999.9433717.x

Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung. [Multiple-choice exams at universities? *A literature review and plea for more practice-oriented research]. Zeitschrift für Pädagogische Psychologie, 29*(3-4), 133-149. https://doi.org/10.1024/1010-0652/a000156

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science* (R package version 2.8.3). https://cran.r-project.org/package=sjPlot

Mammadov, S. (2021). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality, 90*(2), 222-255. https://doi.org/10.1111/jopy.12663

Mansell, W., James, M., Baird, J., Black, P., Daugherty, R., Ecclestone, K., Gardner, J., Harlen, W., Hayward, L., Newton, P., & Stobart, G. (2009). *Assessment in schools. Fit for purpose? A commentary by the Teaching and Learning Research Programme. Economic and Social Research Council.*

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26,* 709-712. https://doi.org/10.1080/0421590400013495

McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology, 21*(1), 89-101. https://doi.org/10.1080/01443410020019867

Metzig, W., & Schuster, M. (1998). *Prüfungsangst und Lampenfieber: Bewertungssituationen vorbereiten und meistern. [Test anxiety and stagefright: Preparing and mastering for evaluation situations].* Springer.

Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology, 19*(4-5), 650-670. https://doi.org/10.1080/09541440701326196

Mittring, G., & Rost, D. H. (2008). Die verflixten Distraktoren. Ueber den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests fuer besser Begabte und Hochbegabte. [The nasty distractors. The utility of a notional distractor analysis of items of matrices tests for the highly gifted). *Diagnostica, 54*(4), 193-201. https://doi.org/10.1026/0012-1924.54.4.193

Nunez-Pena, M. I., Suarez-Pellicioni, M., & Bono, R. (2016). Gender differences in test anxiety and their impact on higher education students' academic achievement. *Social and Behavioral Sciences,* 228, 154-160. https://doi.org/10.1016/j.sbspro.2016.07.023

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*(5), 971-990. https://doi.org/10.1016/j.paid.2007.03.017

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*(3), 157-163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., & Proyer, R. R. (2015). Objectice personality tests. In T. M. Ortner & F. J. R. Van de Vijver (Eds.), *Behavior based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains* (pp. 133-149). Hogrefe.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R). [NEO Personality Inventory by Costa and McCrae, revised version]*. Hogrefe.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, 25(4), 399-438. https://doi.org/10.1080/07370000701632371

Pilli, O., & Admiraal, W. (2017). Students' learning outcomes in Massive Open Online Courses (MOOCs): Some suggestions for course design. *Journal of Higher Education, 7*(1), 46-71. https://doi.org/10.2399/yod.17.001

Poropat, A. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. https://doi.org/10.1037/a0014996

Powell, S. R. (2012). High-stakes testing for students with mathematics difficulty: Response format effects in mathematics problem solving. *Learning Disability Quarterly, 35*(1), 3-9. https://doi.org/10.1177/0731948711428773

R Core Team. (2019). *R: A language and environment for statistical computing*. https://www.R-project.org/

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354-379.

Revelle, W. (2019). *psych: Procedures for personality and psychological research* (R package version 1.9.12.31). Northwestern University. https://CRAN.R-project.org/package=psych

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

Robson, D. A., Johnstone, S. J., Putwain, D. W., & Howard, S. (2023). Test anxiety in primary school children: A 20-year systematic review and meta-analysis. *Journal of School Psychology, 98*, 39-60. https://doi.org/10.1016/j.jsp.2023.02.003

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184. https://doi.org/10.1111/j.1745-3984.2003.tb01102.x

Roemer, L., Lechner, C. M., & Rammstedt, B. (2022). Beyond Competencies: Associations between Personality and School Grades Are Largely Independent of Subject-Specific and General Cognitive Competencies. *Journal of Intelligence, 10*(2): 26. https://doi.org/10.3390/jintelligence10020026

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? *Journal of Educational Measurement, 11*(1), 15-23.

Rubio, V. J., Hernández, J. M., Zaldívar, F., Márquez, O., & Santacreu, J. (2010). Can we predict risk-taking behavior? Two behavioral rests for predicting guessing tendencies in a multiple-choice test. *European Journal of Psychological Assessment, 26*(2), 87-94. https://doi.org/10.1027/1015-5759/a000013

Rust, C. (2002). The impact of assessment on student learning: How can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices? *Active Learning in Higher Education, 3*(2), 145-158. https://doi.org/10.1177/1469787402003002004

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*(4), 929-938. https://doi.org/10.1037/0022-3514.46.4.929

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168-182. https://doi.org/10.1037/0022-3514.94.1.168

Schult, J., & Sparfeldt, J. R. (2016). Reliability and validity of PIRLS and TIMSS: Does the response format matter? *European Journal of Psychological Assessment, 34*(4), 258-269. https://doi.org/10.1027/1015-5759/a000338

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27-41. https://doi.org/10.1080/08917779108248762

Sellers, V. D. (2000). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals, 33*(5), 512-520. https://doi.org/10.1111/j.1944-9720.2000.tb01995.x

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*(1), 1-19. https://doi.org/10.1111/jcal.12172

Silaj, K. M., Schwartz, S. T., Siegel, A. L. M., & Castel, A. D. (2021). Test anxiety and metacognitive performance in the classroom. *Educational Psychology Review, 33*, 1809-1834. https://doi.org/10.1007/s10648-021-09598-6

Spitzer, M. W. H. (2022). Just do it! Study time increases mathematical achievement scores for grade 4-10 students in a large longitudinal cross-country study. *European Journal of Psychology of Education, 37*(1), 39-53. https://doi.org/10.1007/s10212-021-00546-0

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE-Life Sciences Education, 11*(3), 294-306. https://doi.org/10.1187/cbe.11-11-0100

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*, 331-347. https://doi.org/10.1080/02602930500099102

Suskie, L. (2007). Some thoughts and suggestions on assessing student learning. *PS: Political Science & Politics, 40*(1), 102. https://doi.org/10.1017/S1049096507250164

Suskie, L. (2010). How can student learning be assessed? In L. Suskie (Ed.), Assessing Student *Learning: A common sense guide* (pp. 19-35). Jossey-Bass.

Trapmann, S., Hell, B., Hirn, J.-O. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie, 215*(2), 132-151. https://doi.org/10.1027/0044-3409.215.2.132

Vale, J., Oliver, M., & Clemmer, R. M. C. (2020). The influence of attendance, communication, and distractions on the student learning experience using blended synchronous learning. *The Canadian Journal for the Scholarship of Teaching and Learning, 11*(2): Article 11. https://doi.org/10.5206/cjsotl-rcacea.2020.2.11105

Van de Vijver, F. (2016). Test adaptations. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 364 - 376). Oxford University Press.

Verbić, S. (2012). Information value of multiple response questions. *Psihologija, 45*(4), 467-485. https://doi.org/10.2298/PSI1204467V

Wacker, A., Jaunzeme, J., & Jaksztat, S. (2008). Eine Kurzform des Prüfungsängstlichkeitsinventars TAI-G. [A short form of the test anxiety inventory TAI-G]. *Zeitschrift für Pädagogische Psychologie, 22*(1), 73-81. https://doi.org/10.1024/1010-0652.22.1.73

Zeidner, M. (1998). *Test anxiety: The state of the art*. Plenum Press.

Zinn, J. O. (2017). The meaning of risk-taking – Key concepts and dimensions. *Journal of Risk Research, 22*, 1-15. https://doi.org/10.1080/13669877.2017.1351465

Zorio-Grima, A., & Merello, P. (2020). Class-attendance and Online-tests Results: Reflections for Continuous Assessment. *Journal of Teaching in International Business, 31*(1), 75-97. https://doi.org/10.1080/08975930.2019.1698394