

Conventionally developed versus large language model-generated exam tasks in higher education: A comparative analysis

Niklas Lettow ^{1*} , Kristopher Pantani ¹ , Jörn Littkemann ¹ 

¹ University of Hagen, GERMANY

*Corresponding Author: niklas.lettow@fernuni-hagen.de

Citation: Lettow, N., Pantani, K., & Littkemann, J. (2026). Conventionally developed versus large language model-generated exam tasks in higher education: A comparative analysis. *Journal of Digital Educational Technology*, 6(2), Article ep2622. <https://doi.org/10.29333/jdet/19699>

ARTICLE INFO

Received: 23 May 2026

Accepted: 15 Aug. 2026

ABSTRACT

This exploratory field study investigates how human-screened exam tasks generated by the large language model (LLM) GPT-4o perform in comparison with conventionally developed exam tasks with regard to selected quality criteria. Using two business administration modules at a German distance-learning university, human-screened LLM-generated tasks were integrated into submission assignments (SA) alongside instructor-created tasks. Four selected task-quality indicators were examined: objective difficulty, subjective difficulty, context relevance, and item discrimination. The analysis combines descriptive statistics with paired-sample t-tests and Wilcoxon signed-rank tests, applied at three aggregation levels: SA-specific, module-level, and cross-module. The results revealed mixed and task-specific patterns. While selected indicators were numerically similar for some task pairs, statistically significant and, in one comparison, substantial differences were observed in others, particularly in objective difficulty. Given that the study examined only four task pairs that were not systematically matched and that prior human screening was conducted, the findings establish neither equivalence nor general comparability. However, the study provides initial exploratory insights from two business administration modules that can serve as a foundation for further research.

Keywords: analysis, exam, higher education, large language models, quality criteria, task generation

INTRODUCTION

In addition to private and corporate use cases, artificial intelligence (AI) is playing an increasingly important role in higher education teaching (Elbanna & Armstrong, 2024, p. 17). In particular, generative AI systems are being used that are designed to produce text, images, video, and audio (Kurpicz-Briki, 2023, p. 2). For example, they serve as assistants for individualized learner support, for answering administrative questions, as tools to help formulate texts, and for conducting teaching evaluations (de Witt et al., 2020, p. 12). Large language models (LLMs) are often used in this context; they represent a subset of generative AI and are the focus of this article.

In this context, LLMs can be understood as AI-based systems that are specifically designed to process text in order to generate new textual content (Caelen & Blete, 2025, p. 14-15). This article therefore examines the extent to which LLMs are suitable for generating written exam tasks in higher education – specifically in the field of business administration. This effort is motivated by two challenges. First, developing high-quality exam tasks in the conventional way is typically very time-consuming, which repeatedly ties instructors to

routine tasks (Cheung et al., 2025, p. 3; Meißner et al., 2024, p. 53). Second, students increasingly seek greater flexibility, especially regarding the choice of exam dates and locations (Kochskämper & Lips, 2022, pp. 168–170). Given the sharp increase in LLM capabilities in recent years, this study investigates them as potentially suitable tools for addressing these challenges, with the aim of drawing conclusions about the quality of human-screened LLM-generated exam tasks (Fan et al., 2024, p. 3-4).

The analysis relies on a systematic comparison of human-screened LLM-generated exam tasks and conventionally developed exam tasks across four real examination settings. To assess question quality, the study uses the following quality criteria: objective and subjective difficulty, context relevance, and item discrimination. At its core, the article therefore addresses the research question of whether human-screened LLM-generated exam tasks differ from conventionally developed exam tasks with regard to the scores of the selected quality criteria. However, neither formal equivalence nor the indistinguishability of the tasks is assessed. Accordingly, the findings are interpreted as exploratory evidence regarding similarities and differences among the specific task pairs examined. No predefined threshold for acceptable differences between the two task-development approaches was specified.

Empirical evidence on the use of LLM-generated tasks in the context of higher education exams remains scarce. Although several studies have used LLM-generated tasks in practice-oriented exercise settings, publications that describe the use of LLM-generated tasks in real exam settings are still extremely rare (Kiyak et al., 2024, p. 2; Rivera-Rosas et al., 2024, p. 2). In the field of business administration in particular, no studies have been identified to date. This article therefore aims to take an initial step toward closing this research gap. The remainder of the paper explains how the study was implemented in detail and what results it produced. It is structured as follows:

Sections *Conceptual Background* and *Methodology* first outline the conceptual foundations and the methodological approach. Section *Results* then presents the empirical results. Section *Discussion and Limitations* provides a critical discussion of the methodology and findings and addresses key study limitations. Finally, Section *Conclusions and Future Work* concludes the article with a summary and an outlook on further potential research.

CONCEPTUAL BACKGROUND

The evaluation of LLM-generated exam tasks can be grounded in established foundations of assessment research: assessment validity concerns the appropriate interpretation and use of assessment results (Kane, 2013), while fairness is regarded as a central quality criterion in standardized test development (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In addition, research on automated item generation (Gierl & Lai, 2013) as well as work on item writing and item review provide important points of reference for the systematic development and evaluation of tasks (Haladyna et al., 2002). Furthermore, psychometric research emphasizes criteria such as item difficulty, item discrimination, reliability, and construct validity (Moosbrugger & Kelava, 2012), while recent studies on the quality of AI-generated assessment tasks show that LLM-generated items continue to require expert review and empirical item analysis (Isley et al., 2025).

Against this background, the present study applies these assessment-related quality considerations to a concrete higher education context. For this purpose, two business administration modules were selected as the objects of analysis. They are currently used in several degree programs within the Faculty of Business and Economics and can be studied each semester by enrolled students. Based on a consistently developed case study featuring a fictitious mid-sized company, the modules place particular emphasis on the practical application of selected controlling tools. In addition to explaining discipline-specific knowledge, a substantial share of the content therefore focuses on applying qualitative and quantitative methods. For this reason, both modules appear particularly suitable for generating LLM-based exam tasks, as they do not merely aim to convey definitional knowledge but also promote the development of methodological competencies.

From a scheduling perspective, students complete each module at the end of the semester with a written final exam. In addition, by the middle of the semester they have to submit an intermediate assessment in the form of two so-called submission assignments (SA). Successfully completing these SA serves as the eligibility requirement for the module's final exam. Passing the SA therefore constitutes an at least indirect but mandatory prerequisite for completing the module overall. Unlike the final exam, which is limited to a two-hour working time and is proctored, the SA can be completed without supervision, independent of time and location, up to a specified submission deadline. In terms of content-related difficulty, the SA tasks are broadly comparable to those of the final exam. Beyond functioning as an admission criterion for the final exam, the SA also serve to monitor students' individual learning progress. To date, both the SA (as well as the final exams) have been designed exclusively in a conventional manner. In the winter semester 2025/2026 – the study period of this article – the conventionally developed tasks were supplemented with LLM-generated tasks. As a result, the SA offered in that semester consisted for the first time of a mix of conventionally developed and LLM-generated tasks.

It should be noted at the outset that, in this article, the use of LLM-generated tasks is limited exclusively to the SA. The decision to introduce LLM-generated tasks initially only in the SA – and not (yet) in the module final exams – rests primarily on didactic, organizational, and quality-assurance considerations. The SA are particularly well suited as a testing ground because, compared with the final exams, they carry lower stakes. This makes it possible to gather initial systematic evidence on the quality, clarity, and acceptance of LLM-generated tasks without immediately incurring disproportionate legal and regulatory risks associated with examinations. In addition, the SA allow for more intensive reflection on task design by instructors and students, since feedback, processing difficulties, and content-related ambiguities can be identified and documented at an early stage. This feedback provides the basis for a robust evaluation and, if necessary, targeted adjustments before any later use in summative formats such as the module final exam. Finally, this staged approach also offers greater legal certainty, because the validity and fairness of exam tasks must meet particularly stringent requirements. Accordingly, prior use in lower-threshold assessments is intended to test the reliability and non-discriminatory nature of LLM-generated exam tasks under real-world conditions.

With regard to the LLM-based generation of tasks, GPT-4o – an LLM developed by OpenAI – was selected as the tool for this study (OpenAI, 2024). This choice was made because, at the time of the study, GPT-4o was among the most capable LLMs, was comparatively easy to use, and produced the most promising results in a series of pilot tests. In addition, GPT-4o is extensively pre-trained and comes with a broad knowledge base without further customization. It also allows users to configure customized chatbots that can be enriched with user-provided data. This approach was implemented in the present study and was considered essential, because a set of preliminary tests showed that LLM-generated exam tasks remained too superficial and unspecific when they did not

Table 1. Overview of personalized chatbots for task generation

Module	Chatbot	Thematic Area	Course Unit
Instruments of Controlling	1	Procurement KPIs	1 – Procurement Controlling
	2	Production Program Planning	2 – Manufacturing Controlling
	3	Revenue Variance Analysis	3 – Marketing Controlling
	4	Order Quantity Optimization	4 – Logistics Controlling
	5	Human Capital Valuation	5 – HR Controlling
Group Controlling	6	Coordination of Group Controlling	1 – Organization of Group Controlling
	7	Risk Controlling	3 – Risk Controlling

incorporate module-specific, discipline-related contexts. This targeted customization is intended to ensure that the LLM-generated exam tasks are clearly solvable on the basis of the course materials, content-accurate, and didactically appropriate.

METHODOLOGY

Data Basis and Chatbot Configuration

This section first provides additional information on the selected modules, namely “Instruments of Controlling” (module no. 31601) and “Group Controlling” (module no. 32591). While the former is used in several bachelor’s programs, the latter is part of the master’s programs offered by the Faculty of Business and Economics. In addition to the conventionally developed tasks, customized chatbots were configured for both modules and were used to generate the LLM-based tasks. All teaching materials for the respective module served as the data basis for the chatbots. Accordingly, the dataset comprised course texts/units, final exams from previous semesters including sample solutions, Excel spreadsheets with formulas and calculation schemes for various quantitative tasks, and SA from previous semesters.

The content structure of module 31601 comprises five thematic focal areas. Following the corporate value chain (procurement → production → sales), and including logistics and human resource management as cross-functional areas, the module therefore consists of five course units covering procurement controlling (course unit 1), manufacturing controlling (course unit 2), marketing controlling (course unit 3), logistics controlling (course unit 4), and HR controlling (course unit 5). In terms of content, the module provides a comprehensive overview of selected controlling tools within these functional areas. Module 32591, by contrast, is structured into three thematic blocks: organization of group controlling (course unit 1), instruments of group controlling (course unit 2), and risk controlling (course unit 3). At the beginning of each semester, the course units are sent to students and can be regarded as the distance-learning equivalent of traditional in-person lectures at campus-based universities. They contain all discipline-specific content that students are expected to master through self-study.

As part of the chatbot configuration process, the complete dataset was uploaded to the GPT-4o platform and used to create one customized chatbot per module. In addition to uploading the material, the customized chatbots were provided with written instructions to define their core tasks and behaviors. The chatbots could then be tested immediately. Depending on the output produced, both the instructions and the input prompts can be varied and adjusted to improve the

output step by step. Through repeated feedback and refined instructions, the aim is that, after re-entering a prompt, the previously generated output progressively converges toward the intended result. In principle, this cycle can be repeated as often as needed – ideally until the chatbot delivers the desired results or further improvements no longer appear significant. In the present application, however, the initial outputs largely failed to meet the basic expectations. Almost all generated exam tasks were either far too easy in terms of difficulty and/or were not sufficiently tied to the module content (i.e., they showed low context relevance). The scope of the application context was therefore gradually scaled down. From a practical standpoint, this suggests that a thematically narrowly focused customized chatbot – such as one designed to generate exam tasks on “revenue variance analysis” – tends to produce higher-quality outputs than a chatbot expected to generate questions across the full range of an entire module. At least during the configuration processes of the chatbots developed for this paper, this pattern became clearly visible. Empirical evidence – albeit from other application contexts – points to similar relationships (Kwan et al., 2024, pp. 15574-15576; Shi et al., 2023, p. 15339-15341).

Ultimately, the scaling process resulted in a total of seven different chatbots, each tailored to a specific topic area within a course unit. In summary, the chatbot development process thus moved from an initial two “generalist bots” through several steps toward seven “specialist bots,” each ultimately focused on a narrowly defined topic area (see [Table 1](#), which also provides links to the chatbots). From the perspective of the instructors responsible for the courses, approximately 75% of the LLM-generated tasks proved fully usable. Only tasks that had been rated as usable by both raters were included in the task pool. The classification was therefore based solely on the teaching experience of both instructors rather than on a formalized evaluation rubric. Tasks were then randomly selected from this pool and incorporated into the SAs without any manual post-editing.

Development of the Submission Assignments (SA)

Students complete the SA, and the exam results are evaluated, via a web-based, in-house online examination system. In principle, the system can accommodate any conceivable question format. In the two modules considered here (31601 and 32591), the following formats are currently used: Multiple Choice (MC) (either single selection/matching or multiple selection/matching), Numbers (quantitative questions), and Terms.

Overall, each SA analyzed in this article comprises two tasks, with a maximum of 600 points available per task. Each task consists of multiple subtasks. Accordingly, each SA yields a maximum of 1,200 points and is considered passed if

Table 2. Task structure of the Submission Assignments (SA)

Module	SA	Task	Maximum Possible Score	Developed Conventionally	LLM-generated
31601 – Instruments of Controlling	1	1	600		X
		2	600	X	
	2	1	600		X
		2	600	X	
32591 – Group Controlling	1	1	600		X
		2	600	X	
	2	1	600		X
		2	600	X	

Table 3. Quality criteria

Quality Criterion	Reference Basis	Data Collection
Difficulty (objective)	Score Achieved	Automated (via Online Examination System)
Item Discrimination (objective)	Score Achieved	Automated (via Online Examination System)
Difficulty (subjective)	Likert Scale (0-5)	Manually (by Exam Participants)
Context Relevance (subjective)	Likert Scale (0-5)	Manually (by Exam Participants)

students achieve at least 50% of the maximum score. To ensure comparability, the SA were designed from the outset to balance the point distribution between LLM-generated and conventionally developed tasks. The scope of each task was calibrated so that the workload per task (i.e., per 600 points) amounts to approximately 30 minutes. **Table 2** provides a comprehensive overview of the task structure across all SA. A detailed overview of the respective subtasks can be found in **Appendix 1**.

Quality Criteria

Within the scope of this paper, a total of four quality criteria are used to compare the LLM-generated tasks with the conventionally developed tasks. The selection of these criteria was based on an independent literature review, which indicated that the four selected criteria are among the most frequently used measures for assessing exam tasks (Lettow et al., 2025, p. 41). While the two objective criteria (objective difficulty and item discrimination) are based solely on the scores achieved on the tasks, the two subjective criteria (subjective difficulty and context relevance) rely on assessments provided by the exam participants (see **Table 3**).

With regard to the objective quality criteria, the online examination system automatically generates descriptive statistics when evaluating the exam results (mean, minimum, maximum, standard deviation, correlation coefficient). All of these measures refer to the scores achieved and are therefore free of subjective judgments. A subset of these measures is used in the following to analyze objective difficulty and item discrimination. For objective difficulty, this paper assumes that it is determined by the score achieved on a specific task (Coşkun et al., 2025, p. 4). The easier students find a task, the more closely the mean score will tend to approach the maximum attainable score for that task. The second objective quality criterion, item discrimination, is also captured directly by the online examination system. Because the data are measured on a metric scale, item discrimination is expressed quantitatively using Pearson's correlation coefficient (r), where $-1 \leq r \leq 1$ (Moosbrugger & Kelava, 2012, p. 84-86). The higher the value for a given task, the better that task is at differentiating between high-performing and low-performing students. Specifically, a high discrimination value indicates that exam participants who achieved a high total score on the overall exam or SA also tended to achieve a high score on the

task in question. A low value indicates, conversely, that participants who achieved a low total score on the overall exam or SA were nevertheless able to achieve a comparatively high score on the task in question (i.e., despite weak overall performance) (Fisseni, 2004, p. 36-38). As each SA contains only two tasks, the uncorrected item-total correlation is used. To ensure methodological consistency throughout the analysis, the uncorrected measure is also applied at the higher levels of aggregation.

In addition to these objective criteria, which are entirely based on achieved scores, perceived difficulty and context relevance serve as the selected subjective quality criteria. Perceived difficulty refers to the level of difficulty as experienced by the exam participants, whereas context relevance indicates the extent to which a given task is substantively aligned with the course units (Cheung et al., 2025, p. 5; Drori et al., 2022, p. 6). Both criteria were collected from exam participants for each (sub-)task and were coded on a Likert scale from 0 to 5 as follows:

Subjective Difficulty:

0 = The task is very easy.

...

5 = The task is very difficult.

Context Relevance:

0 = The task is not substantively aligned with the course units at all.

...

5 = The task is fully aligned with the course units.

It should be noted that the subjective assessments were collected immediately after participants completed each (sub-)task – and thus before the exam results were evaluated. Because participants did not (yet) know their individual performance at the time of the self-assessments, their ratings were not subject to any score-driven bias. It should also be noted that participants were neither informed which tasks were conventionally developed and which were LLM-generated, nor were they told that the SA contained any LLM-generated tasks at all. As a result, participants were free from expectations or preconceptions about the origin of the tasks. This procedure was intended to ensure that completion and evaluation were based solely on the substantive and formal quality of the tasks.

Table 4. Determination of the differences

Exam Participant	Context Relevance-Score (0-5) Task 1 (LLM-generated)	Context Relevance-Score (0-5) Task 2 (Developed conventionally)	Difference
1	4.22	4.22	0.00
2	2.51	2.10	0.41
3	3.85	3.92	-0.07
...	...	...	...
150	3.70	3.55	0.15

Evaluation Methods

With regard to the evaluation methods, the study uses not only the calculation and presentation of selected measures of central tendency and dispersion, but also statistical testing procedures. These are intended to provide robust evidence with respect to the core question of whether LLM-generated tasks differ significantly from conventionally developed tasks (Hassler, 2018, p. 191-193). It should be noted, however, that this only permits conclusions about the specific task pairs examined, not about the broader population of LLM-generated and conventionally developed tasks.

Because the samples to be compared are dependent, the analysis is conducted using two-sided paired-sample t-tests (also referred to as paired-difference tests) (Bortz & Schuster, 2010, p. 124-126). The tests are carried out separately for each criterion, meaning that the quality criteria listed in [Table 3](#) are analyzed individually. This approach makes it possible to determine, for each quality criterion, whether a significant difference between LLM-generated and conventionally developed exam tasks should be assumed by comparing the score values of the LLM-generated tasks with those of the conventionally developed tasks. This comparison is performed for each exam participant. Accordingly, the sample sizes (n) are to be understood as the number of exam participants. Using objective difficulty as an example, the difference between the score achieved on the LLM-generated task under consideration and the corresponding conventionally developed task is calculated for each participant. The same procedure is applied analogously to the other quality criteria, except for item discrimination. For a total of n participants, the difference δ_i for the i -th participant is therefore given by:

$$\delta_i = (\text{Score}_i \text{ LLM-generated task}) - (\text{Score}_i \text{ conventionally developed task})$$

For example, with $n = 150$ exam participants, this yields 150 such differences per quality criterion. [Table 4](#) shows an excerpt of the dataset using context relevance as an example.

In the next step, the distribution of the calculated differences must be examined. Because the paired-difference test assumes a t-distributed variable – or, for large samples, a normally distributed variable in the population – it is necessary to ensure that the distribution of the differences follows the t-distribution or normal distribution to at least some extent. The skewness coefficient and kurtosis are generally suitable criteria for this assessment. Skewness indicates whether and to what extent a distribution is asymmetric and leans to the left or right, whereas kurtosis reflects how peaked or flat the distribution is compared with a normal distribution (Hassler, 2018, p. 86-87). Depending on sample size, different decision rules are used to judge whether a given distribution is suitable for a paired-difference test.

These rules should be understood less as strict laws and more as practical guidelines. For sample sizes in the range $n > 300$, Kim, for example, suggests rejecting the assumption of normality if the absolute skewness $|g_3| > 2$ or the absolute kurtosis $|g_4| > 7$. For sample sizes in the range $50 < n < 300$, the assumption of normality should instead be rejected if the so-called Z-values exceed a specified threshold. The Z-values are computed separately for skewness (Z_{Skewness}) and kurtosis (Z_{Kurtosis}). The decision rule in this case is: if $|Z_{\text{Skewness}}| > 3.29$ or $|Z_{\text{Kurtosis}}| > 3.29$, the sample is considered not normally distributed (Kim, 2013, p. 53). The calculation procedures for the relevant parameters are provided in [Appendix 2](#) (Hassler, 2018, p. 86-87; Kim, 2013, p. 53). The specific values obtained for the data analyzed in this article are reported in [Appendix 3](#). They show that, based on the decision rules described above, not all samples are fully suitable for the paired-difference test.

Therefore, in addition to the paired-difference tests, the study also applies a nonparametric test to strengthen the evidential value of the results with an additional foundation. In general, nonparametric tests are statistical procedures that, unlike the paired-difference test, do not assume a specific distributional form for the data under analysis (Bortz & Schuster, 2010, p. 130). For the present study, the Wilcoxon signed-rank test is particularly well suited as the selected nonparametric test. Based on two dependent samples, it examines whether the central tendency of the underlying populations differs significantly (Bortz & Schuster, 2010, p. 133-134). However, the Wilcoxon signed-rank test was used solely as a supplementary procedure to complement and corroborate the results of the t-tests, which served as the primary inferential statistical tests in this study. When conducting both test procedures, the following null hypothesis (H_0) is formulated on the basis of two dependent samples X and Y :

$$H_0: \mu_X - \mu_Y = 0.$$

Expressed in words, H_0 states that, with respect to the quality criterion under investigation, there is no significant difference between the LLM-generated task and the conventionally developed task. If H_0 must be rejected, this accordingly implies that a significant difference should be assumed. This is the case when the p-value is smaller than the significance level (Hassler, 2018, p. 178). Conversely, failure to reject the null hypothesis merely indicates that no statistically significant difference was detected in the present data. However, this demonstrates neither equivalence nor comparability between the two task types, as no equivalence margins were predefined and neither equivalence nor non-inferiority tests were conducted. The test results should therefore be interpreted solely as exploratory findings and not as evidence of statistical equivalence.

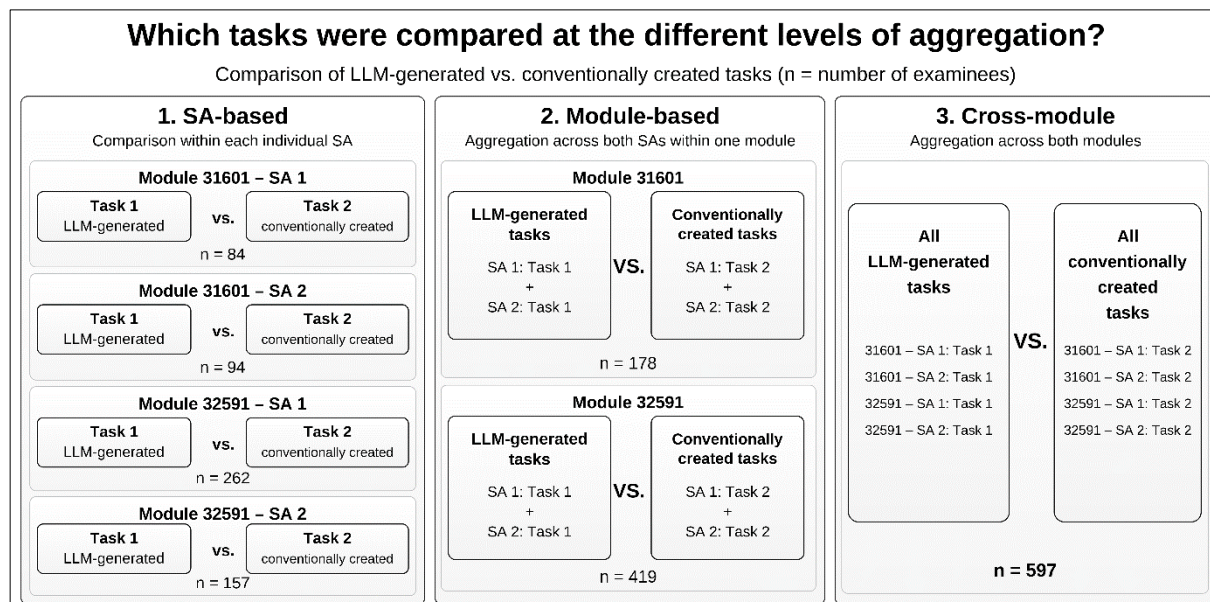


Figure 1. Levels of aggregation in the analysis (Source: Authors' own elaboration, created with the assistance of generative AI (ChatGPT, OpenAI))

For both test procedures, the probability of error was set to $\alpha = 0.01$ – a deliberately low value – to minimize the probability of erroneous conclusions and thereby uphold high standards of exam quality and fairness. At the same time, however, caution is warranted when interpreting individual p-values in isolation, as the large number of comparisons across different quality criteria and aggregation levels increases the risk of alpha-error accumulation. Therefore, in addition to the fixed significance level of $\alpha = 0.01$, Holm-Bonferroni-adjusted significance thresholds were calculated for each family of tests and used in the statistical decision-making process (Holm, 1979, p. 66-68). At this point, it should be noted that the test-family sizes (m) were specified as follows: $m = 12$ for the SA-specific analysis (four SAs with three quality criteria each), $m = 6$ for the module-level analysis (two modules with three quality criteria each), and $m = 3$ for the cross-module analysis (one test for each of the three quality criteria). It should also be stated that the fixed error probability of $\alpha = 0.01$ and the Holm-Bonferroni-adjusted significance thresholds led to different test decisions in three of the total $m = 21$ test constellations. Therefore, the reported p-values should be interpreted with appropriate caution and always in conjunction with the corresponding effect sizes and confidence intervals. For the final classification of the results as statistically significant or non-significant, the test results based on the Holm-Bonferroni correction were ultimately used in Figures 2, 3, 4, and 5.

It should also be noted that, for the quality criterion “item discrimination”, the study does not apply inferential statistical tests but instead relies solely on a qualitative descriptive analysis. The reason is that correlation coefficients are derived from the overall performance of all exam participants and therefore are not available at the individual level, as would be required for difference tests. In addition, procedures such as the paired-difference test require metric raw data from which variance and the mean can be calculated directly – an assumption that is not inherently met when correlation coefficients are available only as aggregated group-level

values. To make correlation coefficients amenable to inferential testing, they would first need to be transformed into a normally distributed series using Fisher's z transformation (Silver & Dunlap, 1987, p. 146-148). Given that only a single correlation coefficient is available per task, comparing only two values per SA would be neither methodologically defensible nor meaningful.

Before presenting the results in the following section, it should be noted that the analysis is conducted at three levels of aggregation: (1) SA-based, (2) module-based, and (3) cross-module. Figure 1 provides a compact overview of these three levels of aggregation. In addition, four methodological limitations should be considered. First, task origin was not isolated from task characteristics such as content, format, and structure, meaning that observed differences cannot be attributed unambiguously to the method of task generation. Second, only GPT-4o-generated tasks selected by instructors were examined. Third, although the statistical tests were based on numerous participant-level observations, they included only four independent task pairs. Fourth, generalizability is limited because the study focused on two business administration modules, a single institution, and lower-stakes assessment formats.

RESULTS

SA-Specific Analysis Results

In the present SA-specific analysis, the findings derived from evaluating all SA in isolation are presented first. Table 5 reports the results of the descriptive statistics for all tasks. It includes both the mean values $\bar{x}$ and the coefficients of variation θ for the quality criteria.

Because each SA contains exactly one conventionally developed task and one LLM-generated task, Table 5 allows a detailed assessment of whether the score values for the individual quality criteria differ significantly between the LLM-generated and conventionally developed tasks. One

Table 5. Descriptive results for the quality criteria by task

Module	SA (n = number of participants)	Task	Objective Difficulty-Score (0-600)	Subjective Difficulty-Score (0-5)	Context Relevance-Score (0-5)	Item Discrimination (-1 ≤ r ≤ 1)
Instruments of Controlling	SA 1 (n = 84)	1 (LLM-generated)	$\bar{x} = 457.14$ $\theta = 0.22$	$\bar{x} = 2.04$ $\theta = 0.53$	$\bar{x} = 3.82$ $\theta = 0.31$	0.85
		2 (Developed conventionally)	$\bar{x} = 519.76$ $\theta = 0.21$	$\bar{x} = 1.82$ $\theta = 0.64$	$\bar{x} = 3.88$ $\theta = 0.36$	0.87
	SA 2 (n = 94)	1 (LLM-generated)	$\bar{x} = 504.47$ $\theta = 0.24$	$\bar{x} = 2.15$ $\theta = 0.44$	$\bar{x} = 3.71$ $\theta = 0.32$	0.83
		2 (Developed conventionally)	$\bar{x} = 473.51$ $\theta = 0.31$	$\bar{x} = 2.11$ $\theta = 0.54$	$\bar{x} = 4.00$ $\theta = 0.31$	0.88
Group Controlling	SA 1 (n = 262)	1 (LLM-generated)	$\bar{x} = 546.98$ $\theta = 0.09$	$\bar{x} = 2.32$ $\theta = 0.40$	$\bar{x} = 3.36$ $\theta = 0.39$	0.64
		2 (Developed conventionally)	$\bar{x} = 529.54$ $\theta = 0.13$	$\bar{x} = 2.07$ $\theta = 0.54$	$\bar{x} = 3.46$ $\theta = 0.41$	0.85
	SA 2 (n = 157)	1 (LLM-generated)	$\bar{x} = 549.81$ $\theta = 0.10$	$\bar{x} = 2.44$ $\theta = 0.38$	$\bar{x} = 3.13$ $\theta = 0.42$	0.51
		2 (Developed conventionally)	$\bar{x} = 267.01$ $\theta = 0.54$	$\bar{x} = 3.18$ $\theta = 0.33$	$\bar{x} = 3.06$ $\theta = 0.43$	0.94

Table 6. Results of the difference tests by SA

Module	SA (n = number of participants)	Criterion	p-value ($\alpha = 0.01$)	Holm- Bonferroni adjusted significance thresholds (m = 12)	Cohen's d _z	CI (99%)	Reject H ₀ : $\mu_X - \mu_Y = 0$? (Significant difference between the LLM- generated and conventionally developed tasks?)
							Paired-Difference Test Result ¹ Wilcoxon Signed-Rank Test Result ² Paired-Difference Test Result with Holm-Bonferroni-adjusted significance threshold ³
Instruments of Controlling	1 (n = 84)	Objective Difficulty-Score	< 0.0001	0.0010	-0.59	[-93.35, -31.89]	Yes ¹ /Yes ² /Yes ³
		Subjective Difficulty-Score	0.0043	0.0014	0.32	[0.02, 0.42]	Yes ¹ /Yes ² /No ³
		Context Relevance-Score	0.4765	0.0050	-0.08	[-0.26, 0.15]	No ¹ /No ² /No ³
	2 (n = 94)	Objective Difficulty-Score	0.0368	0.0020	0.22	[-7.48, 69.39]	No ¹ /No ² /No ³
		Subjective Difficulty-Score	0.6087	0.0100	0.05	[-0.15, 0.23]	No ¹ /No ² /No ³
		Context Relevance-Score	0.0055	0.0017	-0.29	[-0.55, -0.02]	Yes ¹ /No ² /No ³
Group Controlling	1 (n = 262)	Objective Difficulty-Score	0.0005	0.0013	0.22	[4.53, 30.35]	Yes ¹ /Yes ² /Yes ³
		Subjective Difficulty-Score	< 0.0001	0.0011	0.29	[0.12, 0.40]	Yes ¹ /Yes ² /Yes ³
		Context Relevance-Score	0.0600	0.0025	-0.12	[-0.23, 0.04]	No ¹ /No ² /No ³
	2 (n = 157)	Objective Difficulty-Score	< 0.0001	0.0008	1.97	[252.95, 312.65]	Yes ¹ /Yes ² /Yes ³
		Subjective Difficulty-Score	< 0.0001	0.0009	-0.80	[-0.93, -0.55]	Yes ¹ /Yes ² /Yes ³
		Context Relevance-Score	0.3603	0.0033	0.07	[-0.13, 0.27]	No ¹ /No ² /No ³

particularly striking example is the difference in mean objective difficulty in SA 2 of module 32591 (549.81 points vs. 267.01 points), which points to a highly significant difference. **Table 6** reports the specific results of the paired-difference tests and the Wilcoxon signed-rank tests. For each SA, the LLM-generated task was tested against the conventionally developed task.

With regard to statistical significance, it should first be noted that the results of the paired-difference tests at $\alpha = 0.01$ almost entirely align with those of the Wilcoxon signed-rank tests, as the two procedures led to the same test decision in eleven of twelve cases. Only in one case (see context relevance, SA 2, module 31601, in **Table 6**) did the paired-difference test indicate a significant difference, whereas the Wilcoxon signed-rank test did not. In addition to the p-values, the corresponding effect sizes, measured using Cohen's d_z (Azaad, 2026, p. 9) and the 99% confidence intervals for the mean differences (Bortz & Schuster, 2010, p. 196) are reported. A particularly pronounced difference was observed for the objective difficulty score in SA 2 of module 32591. This difference was already apparent from an initial inspection of the descriptive statistics and is also reflected in the item discrimination values (0.51 vs. 0.94; see **Table 5**). **Figures 2**

(module 31601) and **3** (module 32591) provide a summarized overview of the SA-specific results.

Taken together, the two figures indicate that the objective difficulty score differs significantly in three of the four SA. However, there is no consistent direction of deviation, because the higher score alternates between the LLM-generated and the conventionally developed task. The perceived difficulty score also shows significant differences in two of the four SA. Here, the direction of deviation tends to favor the LLM-generated tasks, as participants rated them as more difficult in three of the four SA. However, the mean differences are partly small, so the statistical significance should not automatically be interpreted as indicating high practical relevance. With regard to context relevance, no statistically significant difference was observed in any of the four SA. The direction of deviation tended to favor the conventionally developed tasks, which achieved the higher score in three of the four SA. From the participants' perspective, the conventionally developed tasks were therefore somewhat more closely aligned with the course units than the LLM-generated tasks. With respect to item discrimination, six of the eight tasks achieved values above 0.80, indicating that most tasks were well suited to distinguishing between high-performing and low-performing participants. Only two tasks yielded comparatively lower

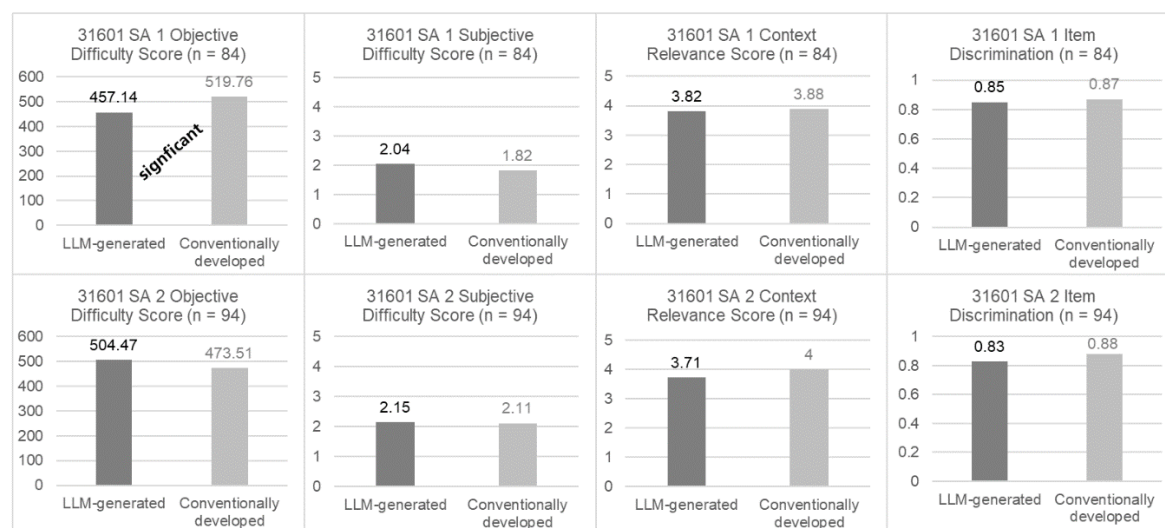


Figure 2. SA-specific results overview (Module 31601) (Source: Authors' own elaboration)

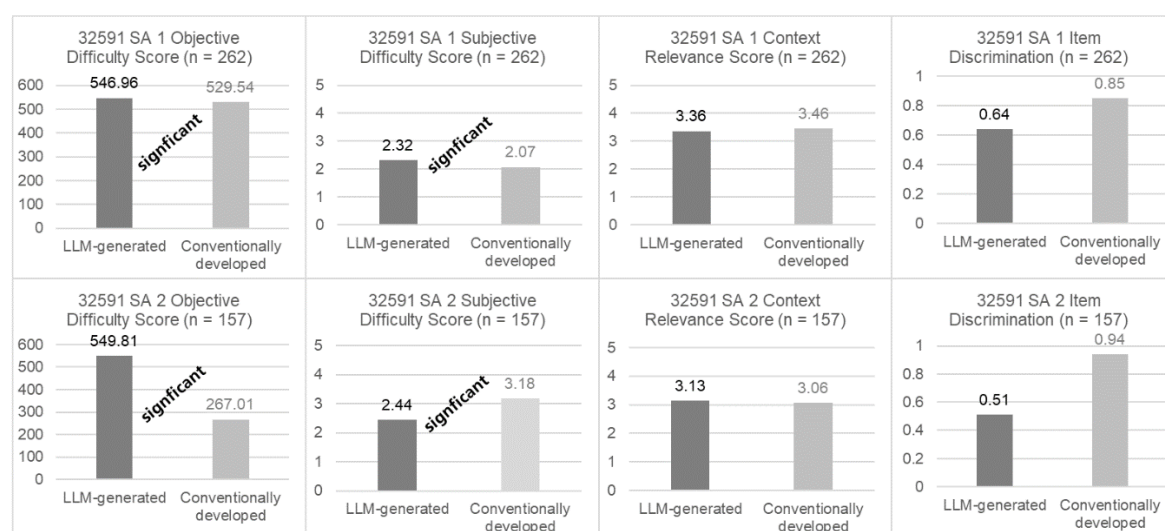


Figure 3. SA-specific results overview (Module 32591) (Source: Authors' own elaboration)

values of 0.51 and 0.64. Because both values stem from LLM-generated tasks, this suggests that the LLM-generated tasks were, on average, somewhat less effective at differentiating between high- and low-performing examinees. A possible explanation could be that, without empirical feedback and without in-depth knowledge of the respective target group, LLMs lack the psychometric fine-tuning required to achieve high item discrimination. However, because task origin and task characteristics are confounded, these differences cannot be clearly attributed to LLM generation.

In summary, the SA-specific analysis thus presents a mixed picture with regard to statistical significance: significant differences were identified in some cases, whereas no significant differences were found in others, with the only clear outlier being the objective difficulty score in SA 2 of module 32591, which is also reflected in the corresponding item discrimination value.

Module-Level Analysis Results

In addition to the SA-specific evaluation presented in the previous section, the following analysis adopts a module-level

perspective. Here, the score values of all LLM-generated tasks within a module are compared with the score values of all conventionally developed tasks within the same module. This aggregation is intended to determine whether, across both SA of a given module, significant differences exist between LLM-generated and conventionally developed tasks. It also makes it possible to examine whether there are significant differences in performance and ratings between the modules (i.e., between bachelor's and master's students). Table 7 presents the corresponding descriptive statistics. It should be noted that, at this level of aggregation, item discrimination refers to the overall result across both SA of the respective module. Accordingly, item discrimination captures only those exam participants who completed both SA of a module in full. For module 31601, this yields a sample size of $n = 45$, whereas for module 32591 the sample size is $n = 104$.

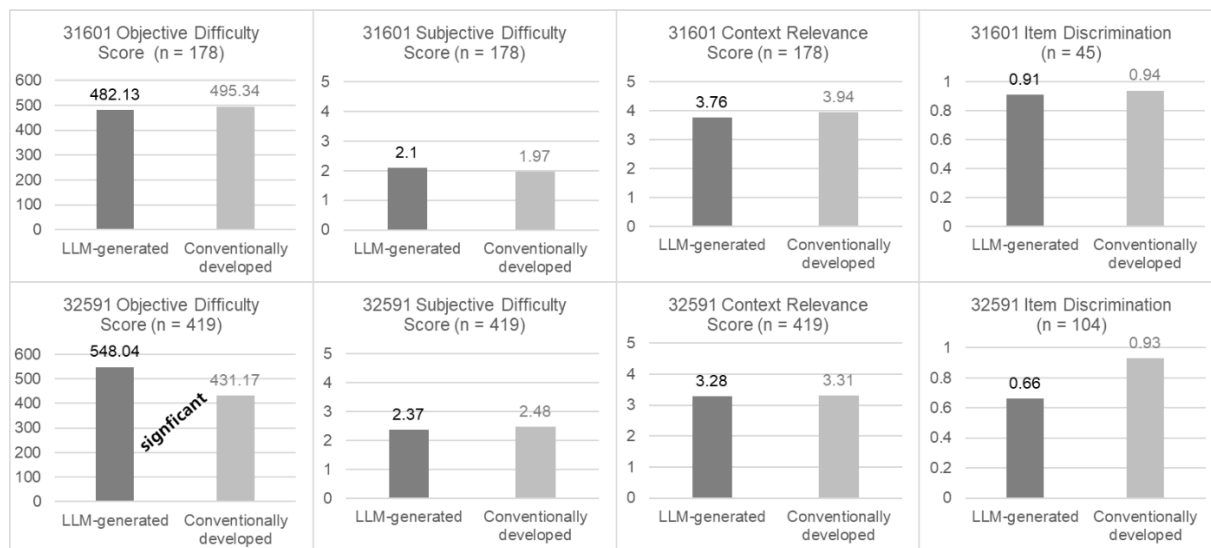
Table 8 reports the results of the paired-difference tests and the Wilcoxon signed-rank tests. Specifically, for each module, the score values of the LLM-generated tasks across both SA were tested against the score values of the conventionally developed tasks across both SA.

Table 7. Descriptive results for the quality criteria by module

Module (n = number of participants)	Task	Objective Difficulty-Score (0-600)	Subjective Difficulty-Score (0-5)	Context Relevance-Score (0-5)	Item Discrimination (-1 ≤ r ≤ 1)
Instruments of Controlling (n = 178)	LLM-generated	$\bar{x} = 482.13$ $\vartheta = 0.24$	$\bar{x} = 2.10$ $\vartheta = 0.48$	$\bar{x} = 3.76$ $\vartheta = 0.31$	0.91
	Developed conventionally	$\bar{x} = 495.34$ $\vartheta = 0.27$	$\bar{x} = 1.97$ $\vartheta = 0.59$	$\bar{x} = 3.94$ $\vartheta = 0.33$	0.94
Group Controlling (n = 419)	LLM-generated	$\bar{x} = 548.04$ $\vartheta = 0.09$	$\bar{x} = 2.37$ $\vartheta = 0.39$	$\bar{x} = 3.28$ $\vartheta = 0.40$	0.66
	Developed conventionally	$\bar{x} = 431.17$ $\vartheta = 0.38$	$\bar{x} = 2.48$ $\vartheta = 0.49$	$\bar{x} = 3.31$ $\vartheta = 0.42$	0.93

Table 8. Results of the difference tests by module

Module (n = number of participants)	Criterion	p-value ($\alpha = 0.01$)	Holm-Bonferroni adjusted significance thresholds (m = 6)	Cohen's d, CI (99%)	Reject $H_0: \mu_x - \mu_y = 0$? (Significant difference between the LLM-generated and conventionally developed tasks?)
Instruments of Controlling (n = 178)	Objective Difficulty-Score	0.1921	0.0050	-0.10 [-39.46, 13.05]	No ¹ /No ² /No ³
	Subjective Difficulty-Score	0.0193	0.0025	0.18 [-0.01, 0.26]	No ¹ /No ² /No ³
	Context Relevance-Score	0.0071	0.0020	-0.20 [-0.35, -0.01]	Yes ¹ /Yes ² /No ³
Group Controlling (n = 419)	Objective Difficulty-Score	< 0.0001	0.0017	0.70 [95.62, 138.13]	Yes ¹ /Yes ² /Yes ³
	Subjective Difficulty-Score	0.0198	0.0033	-0.11 [-0.24, 0.00]	No ¹ /No ² /No ³
	Context Relevance-Score	0.4326	0.0100	-0.04 [-0.15, 0.08]	No ¹ /No ² /No ³

**Figure 4.** Module-level results overview (Source: Authors' own elaboration)

As the test results show, based on the Holm-Bonferroni-adjusted significance thresholds, no statistically significant differences were found in five of the six test constellations at this level of aggregation. Only the objective difficulty score in module 32591 yielded a significant result. **Figure 4** provides an overall summary of the results.

With the exception of the objective difficulty score in module 32591 (and, ceteris paribus, item discrimination), whose significant difference is largely driven by the outlier task 2 in SA 2, the figures show that, even at the module level, the quality-criterion scores are comparatively close to one another. In terms of the direction of deviation, two of the four criteria – context relevance and item discrimination – display

clear (albeit slight) tendencies. As can be seen from the charts in **Figure 4**, in both modules the conventionally developed tasks were somewhat more closely aligned with the course units and were also better suited to differentiating between high-performing and low-performing examinees than the LLM-generated tasks. By contrast, no clear direction of deviation emerged for either objective or perceived difficulty. Overall, the findings at the module level thus show that the observed differences in scores between the LLM-generated and conventionally developed tasks present a mixed picture.

Cross-Module Analysis Results

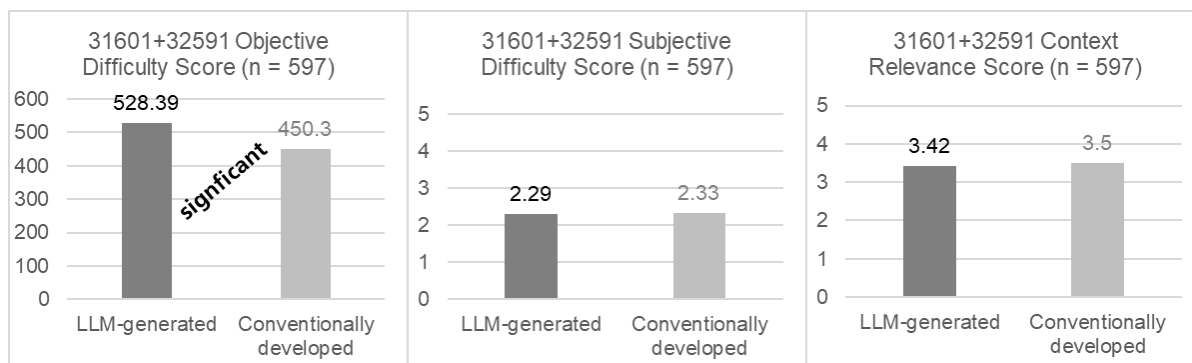
This final section presents a cross-module analysis. Examining the data at this last level of aggregation is intended

Table 9. Descriptive cross-module results for the quality criteria

Tasks (Cross-module; n = number of participants)	Objective Difficulty-Score (0-600)	Subjective Difficulty-Score (0-5)	Context Relevance-Score (0-5)
LLM-generated (n = 597)	$\bar{x} = 528.39$ $\vartheta = 0.15$	$\bar{x} = 2.29$ $\vartheta = 0.42$	$\bar{x} = 3.42$ $\vartheta = 0.38$
Developed conventionally (n = 597)	$\bar{x} = 450.30$ $\vartheta = 0.35$	$\bar{x} = 2.33$ $\vartheta = 0.52$	$\bar{x} = 3.50$ $\vartheta = 0.40$

Table 10. Cross-module results of the difference tests

Criterion (n = 597)	p-value ($\alpha = 0.01$)	Holm-Bonferroni adjusted significance thresholds (m = 3)	Cohen's d, CI (99%)	Reject $H_0: \mu_X - \mu_Y = 0$? (Significant difference between the LLM-generated and conventionally developed tasks?)
				Paired-Difference Test Result ¹ Wilcoxon Signed-Rank Test Result ² Paired-Difference Test Result with Holm-Bonferroni-adjusted significance threshold ³
Objective Difficulty-Score	< 0.0001	0.003	0.46 [60.16, 96.02]	Yes ¹ /Yes ² /Yes ³
Subjective Difficulty-Score	0.2474	0.010	-0.05 [-0.14, 0.05]	No ¹ /No ² /No ³
Context Relevance-Score	0.0343	0.005	-0.09 [-0.17, 0.02]	No ¹ /No ² /No ³

**Figure 5.** Cross-module results overview (Source: Authors' own elaboration)**Table 11.** Ranges of the mean values

Criterion	Range $\bar{x}$ (SA-specific)	Range $\bar{x}$ (module-level)	Range $\bar{x}$ (cross-module)
Objective Difficulty-Score Unit: Score points [0-600]	282.80	116.87	78.09
Subjective Difficulty-Score Unit: Likert scale [0-5]	0.74	0.13	0.04
Context Relevance-Score Unit: Likert scale [0-5]	0.29	0.18	0.08

to determine whether significant differences exist overall between LLM-generated and conventionally developed tasks when both modules are considered jointly. Accordingly, the score values of all LLM-generated tasks are compared with the score values of all conventionally developed tasks. **Table 9** reports the corresponding descriptive statistics.

It should be noted that item discrimination can no longer be computed meaningfully at this level of aggregation, because it would have to refer to the overall result across all SA from both modules. Since the sets of exam participants in the two modules are mutually exclusive, item discrimination is therefore omitted as a quality criterion in the cross-module analysis.

The results of the paired-difference tests and the Wilcoxon signed-rank tests are reported in **Table 10**. First, it can be stated that, in all test constellations, the paired-difference tests led to the same test decisions as the Wilcoxon signed-rank tests. The p-values clearly indicate that there are no significant differences for perceived difficulty and context relevance.

For objective difficulty, however, a significant difference still emerged, as shown in **Figure 5**. As in the module-level analysis, it is largely driven by the outlier. This is confirmed by an additional cross-module paired-difference test excluding the outlier task, which yielded a non-significant result (p-value = 0.3217; see **Appendix 4** for supplementary descriptive statistics). It can be inferred from this figure that, overall, the conventionally developed tasks were more difficult to solve. Subjectively, participants also perceived the conventionally developed tasks as marginally more difficult and as somewhat more closely aligned with the course units than the LLM-generated tasks. Apart from the significant result for achieved scores, the findings therefore suggest that differences between LLM-generated and conventionally developed tasks tend to decrease as sample sizes increase. To further substantiate this assumption, **Table 11** reports the ranges of the mean values. Specifically, the values indicate by how many units the largest observed mean differs from the smallest observed mean. The larger the range, the more pronounced the differences (i.e., greater dispersion) between the LLM-generated and conventionally developed tasks tend to be with respect to the quality criterion considered.

As the values show, the dispersion of the mean values consistently decreases as the number of tasks increases (SA-level → module-level → cross-module). Nevertheless, it should be noted that aggregation across modules may obscure task-specific issues, as opposing differences can offset one another when mean values are calculated. Therefore, the smaller differences observed at higher aggregation levels do not necessarily imply that individual LLM-generated tasks are sufficiently reliable for use in assessments.

DISCUSSION AND LIMITATIONS

The findings of the study present a differentiated picture: Although several comparisons revealed statistically significant differences between the LLM-generated and conventionally developed tasks, other comparisons showed only small numerical differences. These findings suggest that, under the specific conditions examined, LLM-generated tasks may achieve values similar to those of conventionally developed tasks on selected quality indicators. Accordingly, LLM-generated tasks may meet several key requirements of assessment instruments, including the ability to differentiate between higher- and lower-performing examinees, an appropriate level of difficulty, and a close alignment with subject-specific course materials (Bortz & Döring, 2006, p. 220). At this point, however, it should once again be explicitly emphasized that the non-significant findings must be interpreted neither as evidence of equivalence nor as evidence of general comparability between LLM-generated and conventionally developed tasks. They merely indicate that no statistically significant differences were detected under the specific conditions of the present study. Demonstrating equivalence would have required predefined equivalence margins and the use of formal equivalence tests.

Furthermore, the detailed analyses reveal additional critical issues. In the SA-level evaluation, the inferential tests identified significant differences in several test constellations. In particular, objective difficulty showed substantial deviations in some cases. With regard to educational significance, the effect sizes reported using Cohen's d_z should also be interpreted in the context of their potential implications for student assessment. Although the SAs examined in this study are not graded, if the faculty's grading scheme were applied, a score difference of approximately 78 points (see Figure 5; objective difficulty score) would already exceed one full grade interval, while a difference of nearly 117 points (see Figure 4; objective difficulty score, module 32591) would correspond to almost two grade intervals. Differences of this magnitude may therefore be considered educationally meaningful, whereas smaller effect sizes or score differences may merely reflect task-specific variability. This is critical because an unintended increase or decrease in difficulty may not match the intended learning level and could therefore distort performance assessment. Given this remaining instability in balancing difficulty, close oversight by instructors remains advisable. However, this issue is somewhat mitigated by the fact that the effect was less pronounced in both the module-level and cross-module analyses, as the difficulty of individual tasks may vary and may therefore be

locally higher or lower. When a larger number of tasks is considered, these fluctuations increasingly level out, so that average difficulty converges toward a more balanced level as the number of tasks and exam participants increases. The same relationship was also observed for the other quality criteria. This supports the view that the differences observed are less likely to reflect systematic effects and more likely to be driven by task-specific characteristics. However, whether this actually reflects a systematic convergence in task quality would need to be examined in further studies with a larger number of tasks and more rigorously controlled designs. The present inferential results should therefore not be interpreted as population-level evidence regarding LLM-generated and conventionally developed tasks, as only four task pairs were examined despite the relatively large numbers of participants in some cases. Furthermore, the findings may also have been facilitated by the specific characteristics of the two business administration modules examined, as these involve clearly structured quantitative problems and answers that can be assessed as relatively clearly correct or incorrect. More open-ended disciplines, such as humanities subjects with essay-based assessment formats, may, by contrast, pose greater challenges for LLM-supported task generation. Therefore, it should be explicitly emphasized once again that the findings are to be understood as exploratory indications only. They refer exclusively to the specific task pairs examined and do not allow for robust general conclusions about the broader population of LLM-generated and conventionally developed exam tasks.

From a methodological perspective, the study design has several strengths. Particularly noteworthy is the combination of objective and subjective quality criteria, which enables a nuanced assessment of question quality. The decision to pilot LLM-generated and conventionally developed tasks with equal weighting in real exam scenarios also strengthens the study's ecological validity. In addition to descriptive statistics, the study employed paired-difference tests and Wilcoxon signed-rank tests, thereby using two procedures to examine differences between the two task types with statistical rigor.

Despite these strengths, the study also has several methodological constraints and limitations. First, it is important to consider the extent to which the results may be affected by bias inherent in the assessment setting. External confounding factors—including students' motivation, the use of aids, prior knowledge acquired in other modules, potential order effects such as fatigue or practice, and the exchange of proposed solutions among participants—could not be controlled. As a result, the present setting differs substantially from a module final exam with respect to proctoring, administration, and exam organization. Accordingly, the results should be interpreted with caution, because the lack of control over key confounding variables limits internal validity and restricts the extent to which the findings can be transferred to regular, proctored module final exams. Second, it is debatable whether the four selected quality criteria are sufficient to assess reliably and comprehensively whether LLM-generated tasks truly meet the standards of high-quality assessments. It may be useful to include additional criteria, such as the METEOR score as a metric for evaluating machine- or LLM-generated translations (Banerjee & Lavie, 2005), as

well as measures capturing ambiguity, construct alignment, scoring reliability, cognitive demand, fairness, and solution accuracy. Furthermore, the LLM-generated and conventionally developed tasks were not systematically matched in terms of topic, format, or subtask structure; consequently, the observed differences cannot be attributed exclusively to task origin. Another potential concern is the preselection of the LLM-generated tasks used in the SA and the point allocation across the individual subtasks, which relied exclusively on the subjective judgments of the instructors responsible. While pragmatic, this approach risks distortions due to prior assumptions about LLMs or content-related preferences. The same applies to the subjective quality criteria, which were assessed exclusively from the students' perspective. A more systematic validation—for example, involving multiple independent subject-matter reviewers—would therefore be methodologically more robust. The iterative chatbot development process, which followed a trial-and-error approach, is also practical but not fully transparent or reproducible, because there is ultimately no fixed structure linking input (prompt) and output (generated task) (Wang et al., 2022, p. 154). Another point that warrants discussion is the chosen probability of error of $\alpha = 0.01$ in the applied tests. Whether this value can be considered sufficient or should be reduced further remains open. In addition, the dataset is limited to two specific modules in the field of business administration. It therefore remains unclear whether the findings transfer to other disciplines, student populations, or assessment formats. Finally, the study is also limited by the fact that it relied solely on GPT-4o as the model used for LLM-based task generation. It may therefore be valuable to test additional potentially suitable models in this application context alongside GPT-4o.

Although parts of the methodology are exploratory and strongly context-specific, the overall design is nevertheless sound. It provides valuable initial insights into the potential and limitations of LLMs in exam design. However, further studies should complement these findings by using larger samples, more realistic exam settings, and a more systematic chatbot design in order to draw robust and generalizable conclusions.

CONCLUSIONS AND FUTURE WORK

In summary, this exploratory study found that human-screened GPT-4o-generated tasks produced mixed results across four quality indicators in two business administration modules. Some indicators were similar to those of conventionally developed tasks, whereas other comparisons revealed statistically significant differences. The findings therefore do not demonstrate general comparability between LLM-generated and conventionally developed assessment tasks. Rather, they suggest that such tasks may support assessment design under the specific conditions examined. Given the rapid development of increasingly capable LLMs, these findings may provide an initial basis for further research into their potential to support more individualized assessment processes.

However, several open problems remain that have not yet been addressed satisfactorily. First, the chatbots had to be tailored to a very narrow – and thus highly constrained – application context to achieve sufficiently high question quality. Second, and closely related, LLM reliability remains insufficient, as the share of LLM-generated tasks that were fully usable proved not to be high enough. However, a high level of reliability is indispensable in the context of higher education assessments, because otherwise the additional effort – such as dealing with objections triggered by flawed questions – would more than offset the potential benefits of LLM-based examinations. One initial step toward systematically improving reliability could be the use of different prompting techniques. This would make it possible to compare alternative methodological approaches with respect to the results they yield. In addition, so-called retrieval-augmented generation (RAG) approaches have increasingly moved into focus in recent years and are receiving growing attention due to promising results (Hamidi et al., 2025; Li et al., 2025; Nikolovski et al., 2025). These approaches also allow the integration of external data sources and functions.

Nevertheless, the present findings indicate that human review remains necessary. Any future reduction in human oversight would require robust evidence that LLM-generated tasks consistently meet predefined standards of accuracy, validity, fairness, and reliability. Only once such reliability can be ensured could LLMs support more individualized and flexible assessment formats, potentially offering students greater flexibility regarding when and where they complete assessments. However, such formats would also require appropriate procedures for grading and proctoring. With respect to grading, manual correction is not a viable option – at least from an efficiency perspective. With respect to proctoring, one-to-one human supervision would likewise be of limited practical value. The use of individualized, LLM-generated exams therefore appears sensible only if appropriate options for automated, AI-based grading and proctoring are also available. The literature already includes a number of contributions on this topic (Byun et al., 2025; Han et al., 2024; Heinrich, 2025; Jürgensmeier & Skiera, 2024; Pecuchova et al., 2025). As with LLM-based task generation, reliability and consistently low error rates will be decisive criteria in the contexts of grading and proctoring as well, determining whether – and to what extent – the use of LLMs will become a viable alternative to conventional assessment processes.

Author contributions: NL: conceptualisation, data curation, formal analysis, methodology, visualization, writing - original draft; KP: writing - review & editing; JL: writing - review & editing, supervision. All authors approved the final version of the article.

Funding: No external funding is received for this article.

Ethics declaration: The authors stated that the study was conducted in the context of regular teaching and assessment activities at the University of Hagen. The analysis was based on data generated during regular submission assignments and on students' task-related evaluations. Only anonymized data were used for the analysis, and the results are reported exclusively in aggregated form. According to the applicable institutional requirements, formal approval by an ethics committee was not required for this study.

AI statement: The authors stated that generative AI (GPT-4o, OpenAI) was used as part of the study methodology to generate the examination tasks analyzed in this study, as described in the Methodology section. Generative AI (ChatGPT, OpenAI) was also used to support the creation of Figure 1, which was subsequently reviewed and approved by the authors.

Declaration of interest: The authors declared no competing interest.

Availability of data and materials: All data generated or analyzed during this study are available for sharing when appropriate request is directed to corresponding author.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Azaad, S. (2026). *Computing Cohen's dz from commonly reported statistics: A practical guide for the meta-analysis of paired-samples mean differences* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/7gnj6_v2
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72). Association for Computational Linguistics. <https://aclanthology.org/W05-0909/>
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* [Research methods and evaluation for human and social scientists]. Springer Medizin. <https://doi.org/10.1007/978-3-540-33306-7>
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* [Statistics for human and social scientists]. Springer. <https://doi.org/10.1007/978-3-642-12770-0>
- Byun, G., Rajwal, S., & Choi, J. D. (2025). *LLM-as-a-grader: Practical insights from large language models for short-answer and report evaluation* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2511.10819>
- Caelen, O., & Blete, M.-A. (2025). *Anwendungen mit GPT-4 und ChatGPT entwickeln: Intelligente Chatbots, Content-Generatoren und mehr erstellen* [Developing applications with GPT-4 and ChatGPT: Create intelligent chatbots, content generators, and more] (2. ed., T. Demmig, trans.).
- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T.-H. (2025). Correction: ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE*, 20(6), Article e0327290. <https://doi.org/10.1371/journal.pone.0327290>
- Coşkun, Ö., Kiyak, Y. S., & Budakoğlu, I. İ. (2025). ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher*, 47(2), 268-274. <https://doi.org/10.1080/0142159X.2024.2327477>
- de Witt, C., Rampelt, F., & Pinkwart, N. (2020). *Künstliche Intelligenz in der Hochschulbildung* [Artificial intelligence in higher education] [White Paper]. <https://doi.org/10.5281/zenodo.4063721>
- Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, K., Chen, L., Tran, S., Cheng, N., Wang, R., Singh, N., Patti, T. L., Lynch, J., Shporer, A., Verma, N., Wu, E., & Strang, G. (2022). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences of the United States of America*, 119(32), Article e2123433119. <https://doi.org/10.1073/pnas.2123433119>
- Elbanna, S., & Armstrong, L. (2024). Exploring the integration of ChatGPT in education: Adapting for the future. *Management and Sustainability*, 3(1), 16-29. <https://doi.org/10.1108/MSAR-03-2023-0016>
- Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2024). A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5), 1-25. <https://doi.org/10.1145/3664930>
- Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik: Mit Hinweisen zur Intervention* [Textbook of psychological assessment: With guidelines for intervention] (3. revised and expanded ed.).
- Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32(3), 36-50. <https://doi.org/10.1111/emip.12018>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5
- Hamidi, C., Badiy, M., Gaou, S., Amounas, F., Azrour, M., Tribak, H., Alnajim, A. M., & Alabdulatif, A. (2025). Enhancing automated exam creation with retrieval-augmented generation for scalable educational assessment. *Journal of Advances in Information Technology*, 16(10), 1430-1441. <https://doi.org/10.12720/jait.16.10.1430-1441>
- Han, S., Nikou, S., & Ayele, W. Y. (2024). Digital proctoring in higher education: A systematic literature review. *International Journal of Educational Management*, 38(1), 265-285. <https://doi.org/10.1108/IJEM-12-2022-0522>
- Hassler, U. (2018). *Statistik im Bachelor-Studium: Eine Einführung für Wirtschaftswissenschaftler* [Statistics in undergraduate studies: An introduction for business and economics students]. Springer Gabler. <https://doi.org/10.1007/978-3-658-20965-0>

- Heinrich, E. (2025). A systematic-narrative review of online proctoring systems and a case for open standards. *Open Praxis*, 17(3), 485-499. <https://doi.org/10.55982/openpraxis.17.3.836>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.
- Isley, C., Gilbert, J., Kassos, E., Kocher, M., Nie, A., Brunskill, E., Domingue, B., Hofman, J., Legewie, J., Svoronos, T., Tuminelli, C., & Goel, S. (2025). *Assessing the quality of AI-generated exams: A large-scale field study* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2508.08314>
- Jürgensmeier, L., & Skiera, B. (2024). Generative AI for scalable feedback to multimodal exercises. *International Journal of Research in Marketing*, 41(3), 468-488. <https://doi.org/10.1016/j.ijresmar.2024.05.005>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedem.12000>
- Kiyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, 80(5), 729-735. <https://doi.org/10.1007/s00228-024-03649-x>
- Kim, H.-Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1), 52-54. <https://doi.org/10.5395/rde.2013.38.1.52>
- Kochskämper, D., & Lips, A. (2022). So flexibel wie möglich?! Visionen für ein Studium (Sozial-)Pädagogik 2035 [As flexible as possible?! Visions for (social) pedagogy studies in 2035]. *Sozial Extra*, 46, 167-171. <https://doi.org/10.1007/s12054-022-00488-x>
- Kurpicz-Briki, M. (2023). Generative AI: Was ist das und was kann sie bereits? [Generative AI: What is it and what can it already do?] *SocietyByte – Wissenschaftsmagazin der Berner Fachhochschule*, 1-6. <https://doi.org/10.24451/arbor.19044>
- Kwan, W.-C., Zeng, X., Wang, Y., Sun, Y., Li, L., Jiang, Y., Shang, L., Liu, Q., & Wong, K.-F. (2024). M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15568-15592). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.832>
- Lettow, N., Pantani, K., & Littkemann, J. (2025). LLM-based generation of examination and practice tasks—A systematic literature analysis. *Journal of Convergence in Technology and Management: Global Nexus*, 1, 1-46. <https://doi.org/10.70965/jctmgn.vol.1.issue2.6>
- Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, Article 100417. <https://doi.org/10.1016/j.caeai.2025.100417>
- Meißner, N., Speth, S., Kieslinger, J., & Becker, S. (2024). EvalQuiz—LLM-based automated generation of self-assessment quizzes in software engineering education. In *Software Engineering im Unterricht der Hochschulen (SEUH 2024)* (pp. 53-64). https://doi.org/10.18420/seuh2024_04
- Moosbrugger, H., & Kelava, A. (Eds.). (2012). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire construction]. (2nd, updated and revised ed.). Springer. <https://doi.org/10.1007/978-3-642-20072-4>
- Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation. *Algorithms*, 18(3), Article 144. <https://doi.org/10.3390/a18030144>
- OpenAI. (2024, May 13). *Hello GPT-4o*. <https://openai.com/index/hello-gpt-4o/>
- Pecuchova, J., Benko, L., & Drlik, M. (2025). Automated grading of open-ended questions in higher education using GenAI models. *International Journal of Artificial Intelligence in Education*, 35(6), 3813-3846. <https://doi.org/10.1007/s40593-025-00517-2>
- Rivera-Rosas, C. N., Calleja-López, J., Ruibal-Tavares, E., Villanueva-Neri, A., Flores-Felix, C. M., & Trujillo-López, S. (2024). Exploring the potential of ChatGPT to create multiple-choice question exams. *Educación Médica*, 25(4), Article 100930. <https://doi.org/10.1016/j.edumed.2024.100930>
- Shi, C., Su, Y., Yang, C., Yang, Y., & Cai, D. (2023). Specialist or generalist? Instruction tuning for specific NLP tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 15336-15348). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.947>
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72(1), 146-148. <https://doi.org/10.1037/0021-9010.72.1.146>
- Wang, Z., Valdez, J., Mallick, D. B., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In *Artificial intelligence in education: 23rd International Conference, AIED 2022* (Lecture Notes in Computer Science, Vol. 13355, pp. 153-166). Springer. https://doi.org/10.1007/978-3-031-11644-5_13

APPENDIX 1

SA Overview

Bachelor's Module: 31601 – Instruments of Controlling				
Submission Assignment 1				
Task 1 (600 points; LLM-generated)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	Numbers	Production Program Planning (CU 2)	240
	b)	Numbers	Production Program Planning (CU 2)	80
	c)	MC (1 of 5)	Production Program Planning (CU 2)	40
	d)	MC (1 of 5)	Production Program Planning (CU 2)	40
	e)	Numbers	Order Quantity Optimization (CU 4)	60
	f)	Numbers	Order Quantity Optimization (CU 4)	60
	g)	MC (1 of 5)	Order Quantity Optimization (CU 4)	40
	h)	MC (1 of 5)	Order Quantity Optimization (CU 4)	40
Task 2 (600 points; conventionally developed)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	Numbers	Coefficient of Variation (CU 1)	120
	b)	Numbers	Coefficient of Variation (CU 1)	120
	c)	MC (1 of 5)	ABC/XYZ-Analysis (CU 1)	60
	d)	MC (x of 5)	ABC/XYZ-Analysis (CU 1)	100
	e)	MC (x of 5)	ABC/XYZ-Analysis (CU 1)	100
	f)	MC (x of 5)	ABC/XYZ-Analysis (CU 1)	100
Submission Assignment 2				
Task 1 (600 points; LLM-generated)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	Numbers	Human Capital Valuation (CU 5)	70
	b)	Numbers	Human Capital Valuation (CU 5)	70
	c)	Numbers	Human Capital Valuation (CU 5)	60
	d)	MC (x of 5)	Human Capital Valuation (CU 5)	100
	e)	MC (x of 5)	Human Capital Valuation (CU 5)	100
	f)	MC (x of 5)	Human Capital Valuation (CU 5)	100
	g)	MC (x of 5)	Human Capital Valuation (CU 5)	100
Task 2 (600 points; conventionally developed)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	Numbers	Revenue Variance Analysis (CU 3)	60
	b)	Numbers	Revenue Variance Analysis (CU 3)	60
	c)	Numbers	Revenue Variance Analysis (CU 3)	60
	d)	Numbers	Revenue Variance Analysis (CU 3)	60
	e)	Numbers	Revenue Variance Analysis (CU 3)	60
	f)	Numbers	Revenue Variance Analysis (CU 3)	60
	g)	Numbers	Revenue Variance Analysis (CU 3)	60
	h)	Numbers	Revenue Variance Analysis (CU 3)	60
	i)	MC (1 of 5)	Contribution Margin Analysis (CU 3)	40
	j)	MC (1 of 5)	Fixed-Cost Coverage Analysis (CU 3)	40
	k)	MC (1 of 5)	SWOT Analysis (CU 3)	40

Master's Module: 32591 – Group Controlling**Submission Assignment 1**

Task 1 (600 points; LLM-generated)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	MC (1 of 4)	Management Control in Group Controlling (CU 1)	40
	b)	MC (x of 5)	KPIs, Management Control Instruments and Reporting (CU 1)	100
	c)	MC (matching)	Management Control Instruments (CU 1)	100
	d)	Terms	Coordination of Group Controlling (CU 1)	160
	e)	MC (matching)	Coordination of Group Controlling (CU 1)	100
	f)	Numbers	DuPont Analysis System (CU 1)	100
Task 2 (600 points; conventionally developed)	a)	Terms	Definition of Effectiveness (CU 2)	100
	b)	Terms	Definition of Efficiency (CU 2)	100
	c)	MC (matching)	On the Relationship between Effectiveness and Efficiency (CU 2)	160
	d1)	MC (x of 3)	Distinguishing Organizational Success from Corporate Success (CU 2)	60
	d2)	MC (x of 3)	Distinguishing Organizational Success from Corporate Success (CU 2)	60
	e)	MC (matching)	Relationship Between Organizational and Corporate Success (CU 2)	120

Submission Assignment 2

Task 1 (600 points; LLM-generated)	Subtask	Task Type	Thematic Area (CU = Course Unit)	Maximum Possible Score
	a)	MC (1 of 4)	Operational Value at Risk (CU 3)	40
	b)	MC (x of 5)	Risk Aggregation (CU 3)	100
	c)	MC (matching)	Types of Risk (CU 3)	100
	d)	Terms	Risk Controlling (CU 3)	160
	e)	MC (matching)	Risk Controlling (CU 3)	100
	f)	Numbers	Operational Value at Risk – Historical Simulation (CU 3)	100
Task 2 (600 points; conventionally developed)	a)	Numbers	ROCE (CU 2)	120
	b1)	Numbers	EVA – NOPAT (CU 2)	60
	b2)	Numbers	EVA (CU 2)	60
	b3)	Numbers	Stewart's r (CU 2)	40
	c)	Terms	Beginning-of-Year Adjusting Entry (CU 2)	80
	d)	Numbers	Year-End Adjusting Entry (CU 2)	160
	e)	Numbers	Calculation of EVA After Adjustment (CU 2)	80

APPENDIX 2

Key Parameters for Describing the Shape of a Distribution

$$\text{Skewness Coefficient: } g_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

$$\text{Standard Error of Skewness: } SE_{\text{Skewness}} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

$$Z_{\text{Skewness}} = \frac{g_3}{SE_{\text{Skewness}}}$$

$$\text{Kurtosis: } g_4 = \frac{m_4}{s^4} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

$$\text{Standard Error of Kurtosis: } SE_{\text{Kurtosis}} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

$$Z_{\text{Kurtosis}} = \frac{g_4}{SE_{\text{Kurtosis}}}$$

APPENDIX 3

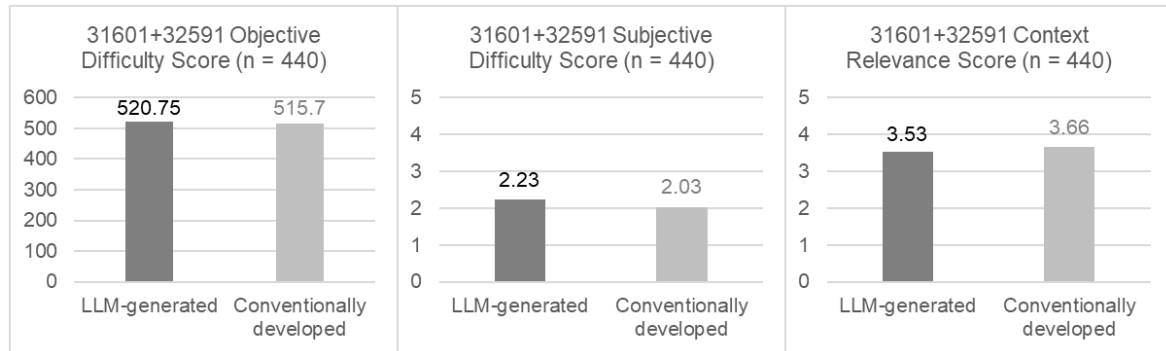
Assumption Checks for Conducting the Paired-Sample t-Tests

Evaluation	Module	SA (n = number of participants)	Criterion	Skewness of the Differences (g ₃)	Kurtosis of the Differences (g ₄)	Z _{Skewness}	Z _{Kurtosis}	Suitable for a paired-difference test?
SA-specific	31601	1 (n = 84)	Objective Difficulty-Score	0.05	0.25	0.19	0.48	Yes
			Subjective Difficulty-Score	-0.49	2.03	-1.87	3.91	No
			Context Relevance-Score	0.18	2.97	0.69	5.73	No
		2 (n = 94)	Objective Difficulty-Score	-0.06	0.36	-0.23	0.74	Yes
			Subjective Difficulty-Score	0.21	0.59	0.84	1.20	Yes
			Context Relevance-Score	0.71	6.84	2.87	13.88	No
	32591	1 (n = 262)	Objective Difficulty-Score	0.41	0.67	2.75	2.24	Yes
			Subjective Difficulty-Score	0.22	1.31	1.48	4.37	No
			Context Relevance-Score	0.38	1.73	2.53	5.78	No
		2 (n = 157)	Objective Difficulty-Score	0.72	0.58	3.73	1.50	No
			Subjective Difficulty-Score	-0.02	0.62	-0.13	1.60	Yes
			Context Relevance-Score	-0.71	1.65	-3.66	4.29	No
Module-level	31601	1+2 (n = 178)	Objective Difficulty-Score	0.22	0.31	1.22	0.86	Yes
			Subjective Difficulty-Score	-0.12	0.96	-0.63	2.64	Yes
			Context Relevance-Score	0.46	6.08	2.51	16.80	No
	32591	1+2 (n = 419)	Objective Difficulty-Score	0.65	-0.57	n/a	n/a	Yes
			Subjective Difficulty-Score	-0.02	0.66	n/a	n/a	Yes
			Context Relevance-Score	0.18	1.94	n/a	n/a	Yes
Cross-module	31601 + 32591	1+2 (31601)	Objective Difficulty-Score	0.61	-0.06	n/a	n/a	Yes
		1+2 (32591)	Subjective Difficulty-Score	-0.13	0.92	n/a	n/a	Yes
		n = 597	Context Relevance-Score	0.26	2.98	n/a	n/a	Yes

APPENDIX 4

Cross-Module Results Excluding the Outlier

Tasks (Cross-module; n = number of participants)	Objective Difficulty-Score (0-600)	Subjective Difficulty-Score (0-5)	Context Relevance-Score (0-5)
LLM-generated (n = 440)	$\bar{x} = 520.75$ $\vartheta = 0.17$	$\bar{x} = 2.23$ $\vartheta = 0.44$	$\bar{x} = 3.53$ $\vartheta = 0.36$
Developed conventionally (n = 440)	$\bar{x} = 515.70$ $\vartheta = 0.20$	$\bar{x} = 2.03$ $\vartheta = 0.56$	$\bar{x} = 3.66$ $\vartheta = 0.38$



(Source: Authors' own elaboration)