Online-Journals.org

iJET International Journal of
Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 13 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i13.38379

PAPER

High-Stakes Online Exams: Faculty Perceptions on Forced Digitization of Assessment During Corona at a Swiss Business School

Douglas MacKevett(✉),
Martin Gutmann

Lucerne School of Business,
Lucerne University of Applied
Sciences and Arts, Lucerne,
Switzerland

Douglas.Mackevett@hslu.ch

ABSTRACT

COVID-19 has affected university assessment procedures on a large scale. This empirical study aims to understand the types of high-stakes exams delivered online at the Lucerne School of Business in Switzerland during the "Corona Semesters" of 2020 and 2021 and the decision-making factors that influenced their implementation. To do so, the authors conducted semi-structured interviews with eight faculty members across a variety of disciplines. Requirements from the exam workflow (preparation, proctoring, grading) were identified and analyzed by course type. Four factors emerged that significantly impacted design and delivery for high-stakes exams online: 1) Digital exam formats significantly impact the nature of exams for procedural subjects such as mathematics; 2) "Group Exams" are not the answer to preventing student collusion on online exams; 3) interrater reliability and low answer variance are considered a central factor for exam quality assurance; 4) second-order effects such as stable wifi and device compatibility will continue to hinder widescale adoption of digital exams. The findings suggest that online exam delivery significantly affects institutional exam practice beyond mere consideration of learning outcomes. The authors conclude by speculating that similar dynamics may have impacted other business schools during their Corona semesters and invite future research on whether the findings from this article can spark discussion and reflection for policy makers in other institutions on post-pandemic legacies.

KEYWORDS

Corona, high-stakes examinations, online exams, academic integrity, interrater reliability

1 INTRODUCTION

While scholarly writing on hybrid teaching and online exams has been around since the 2000s, a comprehensive implementation of digital solutions in higher education remained slow. Digitalization has been heralded as an efficient, cost-saving tool that will help streamline the assessment process; the Corona pandemic, however,

MacKevett, D., Gutmann, M. (2023). High-Stakes Online Exams: Faculty Perceptions on Forced Digitization of Assessment during Corona at a Swiss Business School. *International Journal of Emerging Technologies in Learning (iJET)*, 18(13), pp. 193–208. https://doi.org/10.3991/ijet.v18i13.38379

Article submitted 2023-01-26. Resubmitted 2023-03-25. Final acceptance 2023-04-27. Final version published as submitted by the authors.

has challenged the feasibility of this assumption. In particular, while research prior to Corona focused on exams in general and digital tools in learning experiences, recent literature has begun to address the issue of e-proctoring and open-book exams. Alotaibi [1] examined the factors for adaptation of e-learning assessments from both the faculty and student perspective, identifying both factors for success and the challenges that might prohibit it. Specifically, the topics of exam preparation and academic integrity (i.e. cheating) play a central role. Alghamdi et al [2] propose an intelligent e-learning system for assessment with the goal to reduce the need for proctoring. While such AI-based tools can certainly increase efficiency to a certain extent, the logistics of ensuring exam quality can prove quite challenging [3] and can even prolong the process. Still other have focused on the central question of academic integrity in exams and unethical decision-making [4], the benefits of group exams on the learning experience [5] and the use Bloom's taxonomy to raise the cognitive demand on exams [6], [7], [8]. However, these do not address the issues of impact of the digital medium in exam scenarios nor the question of student collusion during exams.

Therefore, we wish to cast a wider net and identify the determinants of best practice from a faculty perspective. These should focus on individual written exams under non-proctored conditions as we experienced them in 2020–21. This paper investigates the changes made to existing exam practice to accommodate the online format by focusing on exam workflows from preparation to proctoring and finally grading. Exams were a particularly challenging and sensitive dimension of the rapid digitalization of teaching and learning during the Corona period. In the ensuing discussion around these, administration and faculty tended to focus on digitalizing analog content, ensuring failsafe technical delivery and submission of exam papers, while dealing with a myriad of legal issues such as e-proctoring, data privacy and data security [9]. Anecdotal evidence from colleagues around the world and feedback from faculty at our institution suggests that these challenges were by and large managed successfully from a technical and a reliability perspective. However, less clear are the validity and academic integrity of open-book exams with no proctoring. Thus, the specific processes of exam preparation, implementation and evaluation demand further attention.

As we create this framework, our guiding questions will be: what impact did rapid digitalization have on high-stakes exams? How and using what criteria were digitalization decisions taken in the hectic months of the pandemic and what were the implications of these? We find these questions particularly relevant because in university life, as in business, once new procedures have been developed, whether under duress or not, and whether for better or worse, they have a tendency to stick.

For this reason, we decided to conduct an empirical study in our own university environment, the Lucerne School of Business, Lucerne University of Applied Sciences and Arts, Switzerland. We believe that most current studies focus largely on the modalities of exams (online, hybrid, onsite) or offer guidelines on "constructive alignment", suggesting that problems associated with cheating on exams could be solved by creating exams at higher taxonomy levels. This, however, does not mirror actual exam practice in many exams with a high level of reliability and quality. Nor does it reflect challenges brought by online delivery of exams, including system failure and grading at scale [10], [11]. [12]. We offer this empirical study as a first attempt at resolving this.

2 INSTITUTIONAL MANAGEMENT OF HIGH-STAKES EXAMS – AT LUCERNE SCHOOL OF BUSINESS AND BEYOND

The AACSB-accredited Lucerne School of Business is part of the Lucerne University of Applied Sciences and Arts. Adapting the Bologna regulations in 2008, it offers degree programs and executive education to roughly 2000 students. The accreditation system is based on ECTS points, awarded by the successful completion of intramodular or end-of-module exams. With 30 credits (900 hours) constituting a "normal" semester, students at Bachelor or Master level will normally take 5–6 exams, including project work, per semester.

Good assessment practice theoretically occurs independently of its delivery format, yet the pandemic has inadvertently highlighted the shortcomings of this approach, placing enormous stresses on administration, faculty, and students alike [13], [14]. Beyond the technical aspects, however, the purely online exam regimes raised heated discussions around academic integrity [15], [16], [17]. In Switzerland, for example, where the authors work, strict data privacy laws (including, but not limited to, GDPR) meant that there was no regulatory framework for online exams in place. Thus, proprietary e-proctoring tools were not adopted on a widespread level and most exams were administered "dark" with no proctoring of the exam session.

In this article, we focus "high-stakes exams", exams that lead to federally accredited certification across a multiyear degree program. The initial framework for this paper – the institutional management of exams – has already been incorporated into exam practices at the Lucerne School of Business. In this framework, various stakeholders are shown, including the institution itself, exam proctors (typically faculty), students, and communities of vocational practice. Each of these stakeholders are motivated by different factors when constructing exam practice and respond to external factors in relation to a particular stakeholder. So, for example, while students demand equal conditions for exams (fairness), faculty are concerned with the amount of effort and time involved (efficiency) to secure this. These two factors in relation to one another determine the "Exam Workflow".

Table 1. Institutional management of high-stakes exams

Exam Workflow	**Degree Reputation**		Quality Assurance
	Students (Fairness)	Qualification (Validity)	
	Faculty (Efficiency)	Exam (Reliability)	
	Institutional Practice		

This study investigates high-stakes examinations conducted in June 2020, and January and June of 2021. As no existing legislation regulated the conduct of online exams in Switzerland, educational institutions nationwide declined to adopt e-proctoring and instead ran exams online without proctoring. Our school was no exception. Oral exams and defenses were successfully conducted online with videoconferencing tools and are not considered in this paper, nor is a pilot project with e-Proctoring conducted in June 2021.

Much research has already focused on constructive alignment in terms of instructional practice and assessment design [18], [19], [20], [21], [22]. These have largely been based on Bloom's taxonomy [22] and refined over the years (see, for instance, extensive examples in McBeath [21]). In terms of instructional design, researchers agree that teaching practices should align with the outcomes expected to be assessed on exams [8]. In a systematic review of the literature on take-home exams, Steger et al [23] found much agreement across the board on the benefits of non-proctored exams. This, however, was qualified by the problem of unethical student behavior, leading to the recommendation of higher taxonomies and constructive alignment to offset these behaviors.

More recent efforts in Europe have built on this, adding competence-orientation based on the original work of Dreyfus [24]. This has led, in Switzerland, to very elaborate models of assessing competence ([26], [27], [28], [29]). For the purposes of this study, we will refer to the course alignment diagnostic test model proposed by Whetten [19], based on [22]. Essentially, constructive alignment is consistently applied throughout curricular design from teaching and low-stakes assessment practices at the same taxonomy level, culminating in high-stakes exams. It is important to mention that overlap occurs throughout all dimensions [22]. Here, we assume that "Factual" is a given for higher education. We have furthermore adapted the category "Metacognitive" according to Kaiser, who argues for "Situational Knowledge" [25] as an extension of conceptual and procedural knowledge in a strategic and ethical context, where the "right" answer will vary according to situation.

Table 2. The knowledge dimension by subject (adapted from [19])

Knowledge Dimension	Subject Examples
Conceptual	Management, Sustainability
Procedural	Math, Financial Management
Situational	Communication, Marketing, HR

It is equally important to mention that "Conceptual" can also encompass higher taxonomies including understanding, analyzing and evaluating; the emphasis, however, is on interrelations among concepts. Equally, the term "Situational" does not automatically assume higher level taxonomies, as an exam may call for a summary which explains a theoretical model, for example, or include a very common procedure such as writing an email.

While this study attempts to define the primary impacts of non-proctored digital exams, at its center lies the assumption that exam design is largely determined by quality considerations. Assessment practices should be a reliable instrument to measure student competency in a subject area and to offer objective and consistent feedback on their performance [22]. For this reason, we will take a closer look at the central role that grading plays in faculty workflows and how this defines the quality assurance practices of any credentialling institution.

The research on interrater agreement (IRA) and interrater reliability (IRR) dates to the 1960s. For example, researchers in pharmacy wanted to rate the effectiveness of pharmacists' interventions and used a Likert scale to do so [30]. IRA refers to the level at which a panel of coders award the same numerical rating to each criterion, while interrater reliability measures to extent to which those reflect the actual reality of an intervention [30]. For instance, when measuring a pharmacists' communicative effectiveness informing a client of an interaction, raters may agree that they communicated well, but on the wrong intervention, thus achieving a high IRA but low IRR. These variables have been subject to a significant number of ANOVA tests, for example in multiple-choice exams [31], a factor we will not consider in detail here.

IRR has also been researched in qualitative fields as a measure of quality. As Denzin and Lincoln have pointed out, "terms such as credibility, transferability, dependability, and confirmability have replaced the usual positivist criteria of internal and external validity, reliability and objectivity (cited in [32]). The core assumption in the debate is that by providing coding of transcribed interviews, for example, researchers can check the intercoder reliability by asking a second coder to do so independently [30]. This second coding should then be consistent to a large extent with the first. In the Armstrong et al study, six researchers were asked to code the same transcript. They found that, while the codings were similar, the framing of these varied. Possible reasons for this could range from subjective interpretations to social constructivist explanations for the inherent inconsistency of social realities [32].

These results have been discussed in educational assessment circles in terms of quality assurance for exams with text-based short answers. Interestingly, many of these have been conducted in fields with specific vocational applications, for example in pharmacy, nursing, medicine (see, for example, [33]). In a recent study, researchers in Germany investigated two schools conducting the same exam for physiotherapy students using an evaluation sheet with a 6-point rating scale and 20 evaluation criteria. Applying statistical tests shows the IRA was poor, with a central tendency and intergroup bias [34]. The study concludes that this process needs improvement to ensure objective feedback and professional competence among their graduates [34]. This is perhaps not surprising due to the nature of the profession but does raise the question of whether the same results apply to subject areas with unspecified vocational outcomes – business administration students, for instance, with skills across the board but no specific job description as its goal. It is from this perspective that we will be investigating decisions faculty make when designing non-proctored exams.

In the following sections, the decision-making processes for exam design that faculty undertook when considering a non-proctored exam will be discussed under “Preparation”. Secondly, the provisions made for proctoring without digital surveillance are outlined under “Proctoring”. Finally, the evaluation procedures necessary to mark digital exam scripts will be highlighted under “Grading”.

2.1 Preparation

In a first stage, faculty considered current exam scenarios and how these would align with the purely online, open-book format without proctoring. Two main issues became salient: Academic integrity and technical requirements. These will be discussed in detail below.

Academic integrity. In addition to the significant institutional burdens that purely online exams entail, the brunt of the discussion focused on “academic integrity” or, more to the point, “cheating”. This comprises both student collusion and copy-paste plagiarism. While guidelines for creating online exams already existed prior to Corona, most of these focused on reducing student collusion and copy-paste plagiarism (see, for example, [10], [11], [23], [35]). For online exams focusing on multiple-choice and text entry question types, countermeasures included random order of questions and tighter time restraints. The central assumption of this discussion is that students will cheat during non-proctord online exams if given the opportunity to do so [36]. A meta-study by Steger et al found that under high-stakes conditions, students will be more likely to cheat on lower taxonomy question types [23]. In high-stakes exams, this behavior is triggered by opportunity given under anonymity and little risk of personal consequences [20]. However, both faculty and students have a vested interest in ensuring that exam validity remains high. The implication of widespread cheating due to non-proctoring of online exams would seriously harm the institutional reputation [20].

Technical requirements. Students were suddenly faced with new technical requirements for taking exams. Years of exam practice in public schooling for paper-and-pencil tests had not prepared them to deal with compatibility problems, back-up features and the threat of system interruption or failure, including system spikes and slower upload rates. File size could play a role here too: Few students know the throughput rates of their Wi-Fi connections, so that large files uploaded "at the last minute" might in fact take several minutes to do so. Therefore, an important consideration was to make the online exam environment as familiar and easy-to-use as possible. In this way, the institution can reduce student anxiety and stress [35]. Our Learning Management System provided its own set of challenges and faculty were constantly worried about inadvertently making exam documents visible before the exam period began.

The reality of non-proctored online exams with a bring-your-own-device policy added an additional layer of complexity. This included technical issues such as system spikes, compatibility issues (operating systems, versions, processing speed), administrative issues (exam scheduling with additional retake slots in case of system failure), fairness vis-à-vis students in exam rooms with better connectivity, and specific application issues (formatting in Word v PDF, media size in PPT v PDF, and so on). This led most faculty to choose solutions based on "affordances" to improve exam security: Not always the optimal solution, but rather the safest, easiest or most convenient (see [3] for this and for more detailed information on the use of PDFs in exams).

2.2 Proctoring

Here, we define "Proctoring" as the in-exam practice of candidate surveillance, whether in person or via camera. Proprietary software solutions will be specifically referred to as "e-proctoring", and typically include an array of intrusive measures such as camera and microphone surveillance, browser lockdowns, keyboard- and eye-tracking.

At our institution, exam proctoring practices have grown organically from basic exam-room supervision into a procedure with protocols, procedures and documentation including seating assignments, room layout, minute-taking and ID checks. This serves a two-fold purpose: 1) to standardize exam practice in order to reduce formal discrepancies and thus increase exam "Fairness"; and 2) to reconstruct student proximities in case of unpermitted student collusion and thus increase "validity". The administrative effort to ensure both fair and valid exams twice a year is considerable, ending around the beginning of the following semester and beginning a few weeks later again.

While e-proctoring is seen as the digital equivalent of the procedures listed above, it has its detractors. The main criticism placed against e-proctoring is, first and foremost, that it does not prevent cheating [37], though the research-based evidence for or against still remains scant at the time of publishing (see, for example, [38]). Reasons cited for cheating during proctored exams include an "us v them" mentality that encourages students to game the system. From the faculty perspective, AI-based proctoring (for example, eye tracking) leads to a false flag rate of more than 90%, significantly reducing faculty incentives to check each case. Instead, spot checks are preferred, reducing e-proctoring's effectiveness [38].

In favor of e-proctoring is clearly the cost factor, with full-fledged proctored exams costing around USD 20 per student [38], with lower license costs for untimed proctoring. In a pilot project run at the Lucerne School of Business with a proprietary e-proctoring tool in June 2021, this could be confirmed, with costs for proctoring around USD 25 per student. Nevertheless, at our home institution, e-proctoring was never seriously considered due to a variety of regulatory and technical reasons. Thus, the decision was clear that online exams would not be proctored. For these reasons, we will focus on non-proctored exam delivery and the measures taken to ensure fairness and validity.

2.3 Grading

In contrast to paper-based exams, digital exams entail entirely new aspects that affect exam grading. Faculty effort focused on double-checking for complete data entry, complete and timely submissions, and potential compatibility and multimedia rendering issues, for example [39]. From a faculty perspective, the prospect of creating entirely new exam items for the online format is daunting, especially for grading at scale. In any case, previous studies on the prevalence of cheating in online versus in-class exams found no significant variation, leading to the conclusion that cheating is already prevalent on high-stakes exams [35], [40].

This suggests that, when creating high-stakes online exams, faculty will seek to optimize interrater reliability. While exam grading practice has grown over the years with this in mind, most studies have focused on the number of raters involved and the number of criteria assessed [33], [41]. This leads to safeguards assuring exam reliability by identifying item discriminators while simultaneously increasing faculty cognitive offloading [3]. It is our contention that online exam design will favor items that discriminate, such as multiple choice, over higher taxonomy items with high variance, as the latter would lower interrater reliability.

3 FINDINGS ON FACULTY WORKFLOWS FOR ONLINE EXAMS

For this empirical study, we interviewed eight faculty members currently teaching and testing in degree programs at our school. Interviewees were selected by purposive sampling of those faculty who the researchers identified as having the specific experience required at modular level. Additionally, experts were chosen evenly among a mix of module types and their position as head, or coordinator of, modules. Primary data was collected in this study by means of semi-structured interviews initiated with a small number of broader questions to gain information about the participant's background and experience, before focusing on more specific subject matter.

After initial questions involving the interviewees' professional experience and subject-matter expertise, they were asked to describe their experiences with online exams during the "Corona" semesters of 2020–2021. This was followed by module-specific questions concerning the educational goals that they wished to assess and the specific technical delivery of the exam. These included as email and/or our learning management system. Questions then may have varied depending on the level of technical detail – mathematics, for example, had different technical requirements than management theory. Finally, they were asked to reflect on the entire workflow of high-stakes exam creation from preparation to proctoring and, finally, assessment. This section was followed by a reflection on exam quality compared to pre-Corona semesters and a brief overview of exam practices that they will keep, revise or discard going forward in 2022.

These interviews were conducted in either German or English. After transcription, the interviews were coded and the results added to thematic categories including exam purpose, preparation, proctoring and grading. The initial coding was based on taxonomies derived by Whetten's assessment diagnostic tool [19]. These identify three main types of knowledge to be assessed: Conceptual, Procedural and Situational. The types of knowledge associated with these (declarative, procedural and situational) have a long standing in the literature, most recently with Kaiser in the German-speaking regions in which our school operates [25], [26]. The original template was derived from themes identified in initial discussions, then refined and adapted as additional information became available [42], [43].

Table 3. Matrix of interviewee, knowledge dimension and subject examples

	Knowledge Dimension	Subject Examples
P2, P4	Conceptual	Management, Sustainability
P1, P5, P6	Procedural	Math, Financial Management
P3, P7, P8	Situational	Communication, Marketing, HR

In the section that follows, results are presented by the factors that influenced faculty online exam design along each stage of the workflow.

3.1 Preparation

After accounting for the learning curve to figure out appropriate digital tools for exam delivery, time invested in the workflow was distributed differently compared to paper exams, with much more responsibility being placed on the module coordinator (P1, P2, P4, P5, P6):

Preparation time (compared to exams prior to Corona) was enormous, especially for the module coordinator. The other faculty members had little to do, but he had to collect test items, double-check for accuracy, discard items that didn't work digitally. He was also responsible for creating sub-groups for the student cohort online to reduce student collusion, something we normally didn't have to do (P5).

Additionally, most faculty felt that, given the opportunity and the exam format, students will collude with others and copy-paste answers from the internet (P1, P2, P4, P5, P6, P7). Three interviewees from procedural subjects (Department of Mathematics and Statistics) said that, while the online format changed the nature of the exam, it was nearly cheat-proof: The combination of single-answer sheet, randomized questions, marginally substituted values, and time pressure made collusion extremely inefficient and plagiarism nearly impossible (P1, P5, P6). For exams in situational subjects, faculty agreed that cases needed to be fictionalized and answers text-based in order to increase academic integrity (P3, P7, P8).

If items could not be randomized, then faculty often chose fewer MC-type questions with a higher proportion of open-answer questions. For mathematics and similar subjects, the nature of the exam scenario meant that calculations could not be awarded points for partially correct answers: "*Before (Corona) we could correct exams directly on the exam itself by checking the calculations and awarding partial points; exam validity is very high this way*" (P1). As "showing the work" was not an option in the online format for technical reasons (P1, P5, P6), this changed the form of the exam and thus impacted the learning outcomes.

3.2 Proctoring

Faculty were faced with different tasks as proctors: With paper-based exams "*you get what you get*" (P4) and the exam procedures are well documented and supported by years of practice (P5); with online exams, on the other hand, late or alternative (email) submissions needed to be checked, formatting issues discussed, and data transfers (i.e. PDF to XLS) checked manually for accuracy (P2, P6). For situational subjects (P3, P7, P8), the switch to online delivery was a welcome relief befitting a university in the 21st century: "*Prior to Corona, we had all kinds of discussions about using laptops during exams – will students cheat, will it be too noisy, what happens if a laptop crashes. Now, digital exams are standard practice*" (P3). It was also a major improvement over the logistics of pen-and-paper exams delivered in class: "*We have finally entered the 21st-century*" (P8).

A significant factor in conducting exams online with no proctoring was to highly restrict student collusion. In one conceptual-based module typical of its kind, faculty deployed a variety of measures to reduce opportunities to copy-paste and/or collude with other students:

We reduced the number of multiple-choice items and increase text-based answers above all to prohibit student collusion. In (our subject), we could create randomly generated test items.… We also defined random exam sub-groups in later exam sessions and reduced the number of points that students could earn in multiple-choice items in favor of open test questions (P4).

In general, however, faculty agreed that running exams with no direct proctoring left them with an "*uneasy feeling*" (P3) about the overall integrity of the exams.

3.3 Grading

The faculty members interviewed continued their pre-Corona grading practice of either divvying up correction by item (P1, P2, P4, P5) or by student cohort (P3, P6, P7, P8). The practice of single-item correction is considered an additional quality and plagiarism control (P1, P2, P4, P5). This allows faculty a post-exam check of single items across hundreds of answers to see if faults slipped in, something highly difficult to achieve in paper formats (P2). In one case (P6), a poorly performing answer item was spotted in an excel file that most likely would have gone unnoticed by single raters across 300–400 answers:

Of course, we could have spent even more time on each individual answer. If there are several combinations of possible answers, how can we be sure that the student didn't accidentally type the wrong value? We also had to check the integrity of data imported from one data format to the other – we did this manually and it was very time-consuming. I don't know if the effort was worth it, but we did find a test item that was incorrect that we never could have found with pen-and-paper formats (P6).

For grading purposes, the online format highly improved legibility of text answers, thus decreasing time spent deciphering handwriting (P3, P7, P8). P8 mentioned that pen-and-paper exams conducted during in January 2022 were seen by students as inferior to the digital one. For nearly all others working with single and multiple-choice items, machine grading significantly reduced time spent (P1, P2, P4, P5, P6). However, the time spent in the preparation and post-proctoring phase with manual spot-checks of answer scripts, it was generally agreed, offset any benefits of time gained.

For exams at scale (300+ students), faculty were by and large satisfied with the reliability of online exams (P1, P2, P3, P5) as they ensured consistency in correction across a large cohort, an advantage that portfolio-style exams cannot offer:

Why don't we do a … mind map and a voiceover? That's simply for the reason if you have 370 students take an exam, ensuring objectivity on the grading of voice recordings on 370 people – that's nearly impossible. Even if just 300 seconds. So it's difficult enough in ensuring consistency in the evaluation of the exams and there again online exams because you can basically sort the answers just by question and randomly go through all the students and, you know, directly compare screen-to-screen each of the answers (P2).

However, the overall result of this format was considered by two faculty members to be a lower quality exam (P4, P7) as it focused mainly on efficiently managing the large number of students, not the quality of the exam: "Grades on our exams during the Corona period were clearly lower than previously" (P7). Faculty with a large number of exams generally preferred item grading by a single rater, as this promoted the highest possible reliability given the exam format (P1, P2, P5). However, at very large scale (n=860), grading was done by student cohort to ensure that workload was spread evenly (P6).

Table 4. Overview of findings – workflow adaptations due to online delivery

Preparation	Proctoring	Grading
[Conceptual] Peer creation of test items; mock exams; mix of SC, short answer and case-based answers	Item versions with random question and answer order	Machine evaluation and e-scripts by hand, manual checks for data integrity; single rater per item
[Procedural] Peer creation of test items; mock exams; mix of SC, short answer and case-based answers	Item versions with random question and answer order; numerical variabilization	Machine evaluation and e-scripts; single rater per item or per learning group; significant correspondence to ensure data integrity
[Situational] Exam items largely unchanged, case studies fictionalized to prevent copy-paste.	Remained largely unchanged; case studies fictionalized to reduce copy-paste plagiarism	Single rater across learning groups

4 IMPACT OF ACADEMIC INTEGRITY AND TECHNICAL REQUIREMENTS ON EXAM DESIGN

Based on the eight interviews and anecdotal evidence from faculty in other departments, we found the following implications for the delivery of digital exams without proctoring:

1. Digital exam formats significantly impact the nature of exams for procedural subjects such as mathematics;
2. "Group Exams" are not the answer to preventing student collusion on online exams;
3. interrater reliability and low answer variance are considered a central factor for exam quality assurance;
4. second-order effects such as stable wifi and device compatibility will continue to hinder widescale adoption of digital exams.

We will now address these points in light of potential policy decisions. Our assumption is that both course content and exams are not primarily online.

1. Technical affordances change the nature of exams in mathematics, statistics, and financial reporting. While many faculty were familiar with STACK assessment questions, they felt that the disadvantages outweighed the advantages in summative assessments as faculty would have want to train students in its use (P1, P5, P6).

Particularly for STEM subjects, these can be used as a simple way to input mathematical notation using a standard keyboard. Its use could be considered as it offers digital means to offer partial grading by building grading trees, which take into account errors carried forward [44]. This can provide more accurate results for high-stakes digital exams and allow for more effective assessment of student work in digital formats, especially in formative assessments (see [45], [46], [47]).

However, any efforts to increase exam integrity would also lead to a greater potential for system failure. Therefore, faculty in procedural subjects moved away from "showing your work" to answer sheets that decreased complexity and increased security in the following ways: 1) the use of widespread applications such as Adobe PDF [3]; 2) multiple-choice exams with up to ten possible answers, nearly eliminating random guessing; 3) staggered exam delivery with question items in binary up to 25 digits (e.g. instead of "question 4", 1011100010110110101110001), thus severely complicating student collusion. Nevertheless,

2. Given the open-book, non-proctored exam delivery, most online exam solutions were based on both reducing student collusion and opportunities to cheat in order to secure exam validity, particularly among the conceptual and procedural subjects that make up the majority of the curriculum. Undoubtedly, group exams and formative assessments are conducive to learning [5] but the challenge here was to provide non-proctored individual exams under fair conditions. Two institutional specifications support this finding: One, our school, based on AACSB best-practice guidelines, requires at least 50% of assessments to be individual and two, nearly all faculty agree that students receive enough group assessments and formative feedback throughout the semester; we focus here on summative, individual assessments.
3. Recommendations on constructive alignment and higher taxonomies miss the point entirely. Faculty are highly concerned with the Institutional Management of High-Stakes Exams as outlined Table 1 as a central factor of quality control. Against popular conception, they are willing to spend more time creating digital exams to ensure fairness in exam delivery: Many faculty further created sub-groups of students online to administer exam versions to smaller cohorts than usual, significantly adding to their workload to ensure that no duplication or inadvertent releases took place. Others created several versions of the same exam by randomizing question order and/or substituting values in data sets. These extra efforts were motivated by the desire to maintain degree reputation. Exam "best practice" recommendations with low interrater reliability and high answer variance will be met with skepticism at best. Higher taxonomies do not guarantee better exams, especially on large-scale exams at Bachelor level where a basic understanding of the subject matter is required. Blanket statements on exam design that do not consider these points will be dismissed.
4. The institutional effort to manage exams is already considerable; adding extra levels of complexity such as safe-exam browsers in a bring-your-own-device environment unnecessarily complicates exam delivery. To reduce potential system failure due to compatibility or overload, faculty chose familiar formats (e.g. PDF) that could be completed offline, with only brief periods of connectivity needed for down- and uploading. This problem will only be exacerbated for off-campus exam sites with increased technical requirements, with no apparent benefits compared to traditional exams. As one faculty member said: "Writing with a pen on paper is pretty easy, isn't it?" (P1). Proctors were additionally tasked with providing emergency numbers, alternative submission platforms via email, and contingency procedures in the event of system crash on the university's side. On the upside for

faculty, machine correction greatly sped up the grading process and digital exam types increased legibility dramatically. However, assuring data integrity was challenging, as some faculty had to contact students individually to ascertain whether their exam script was complete (P1, P5). These second-order effects will most likely be a major hindrance to widescale adoption of digital exam formats in the future.

5 THE DIGITAL FUTURE OF HIGH-STAKES EXAMS: LESSONS LEARNED

As we have explored in this paper, the transfer of high-stakes exams to an online format is not merely a consideration of its taxonomy and constructive alignment. For institutional exam management, a range of second-order effects play a decisive role in determining exam design. These entail, in faculty workflows, the reduction of variance and cognitive load in order to increase (or maintain) interrater reliability, especially at scale. While the result of this may be uniform exam types and lower taxonomies, it does ensure fairness for students and validity for exams, and thus the degree program's reputation. In response to the (rhetorical) question whether there is any intrinsic value to traditional closed-book, proctored exams [18], especially in the digital age, the answer must be a resounding yes: They reduce exam variance and thus increase cognitive offloading for faculty, ensuring both a more efficient and reliable exam in the grading process.

This, however, does not imply that this is not possible with digital exams. One of the main criticisms leveled against freer forms of exam assessments (portfolios, blog series, etc) is their variance in answers, increasing faculty cognitive load while decreasing exam reliability, especially at scale. Such assessment is deemed more appropriate for professional development than vocational qualification across a degree program [18]. The additional tasks of checking for plagiarism and comparing widely varying results across hundreds of exams should not be underestimated and has led to practices such as spot checks that often fail to ensure exam integrity [37].

In a similar study conducted on learner perceptions in India and Saudi Arabia in 2021, the author team found that online examinations significantly reduced workload for faculty [3], [11]. One partial reason for this is the selection and grading of multiple-choice items, which our faculty also found time well spent. However, evidence from our school suggests that the exact opposite is the case. Not accounting for the "learning curve" needed, faculty experienced a shift in work responsibilities and spent more time than usual checking data for validity (P1, P5, P6). Non-MC items require close attention to control variance and reliability for quality purposes (P2, P4). This will very likely continue into the future of digital exams, as proctors will need backup plans for power outages, compatibility issues and the like not present in paper-based exams. Interestingly, while the same study clearly advocated for the use of e-exams in formative scenarios only, suggesting the skills, effort and technology require for high-stakes exam is not at the level needed to ensure quality.

In light of institutional reluctance in Switzerland to adopt e-proctoring capability, many faculty are returning to pen-and-paper for proctored in-class exams. This prevents copy-paste plagiarism and significantly restricts student collusion. As safe-exam browser solutions with a bring-your-own-device policy are not technically feasible due to a plethora of compatibility issues, pen-and-paper thus ensures fair conditions for students with no recourse to unallowed assistance from the internet. Whereas some may question the validity of such formats in the 21st-century working world, proponents argue that the exam does a better job of measuring student cognition and ensuring a degree program's reputation. Our study would suggest, however, that the oft-cited faculty reluctance to digital transformation echoed by administrators across the globe (see, for example, [48]) stems less from a fundamental resistance to such change than from a sober calculation of its limitations on the ground.

Going forward, all colleagues interviewed agreed that high-stakes exams should be proctored and digital elements incorporated into exams. In several cases (P1, P2, P5), faculty found several advantages to machine correction of multiple-choice items, including higher reliability and increased data integrity. Those involved in multiple-choice testing also concurred that state-of-the-art item creation must be adhered to for quality reasons. This would include faculty creation and validation of test items against accepted standards; in many cases, this is already common practice.

Another significant element in exam grading at scale was interrater reliability. This was considered highest with multiple-choice items corrected by a single rater after the answers were checked for validity. Accordingly, exams that passively permit or actively allow student collusion amount to group work and should be assessed accordingly (P4). Other faculty found that assessing group discussions of a case study online via Zoom is a superior method to the onsite orchestration of multiple exam rooms and will continue this exam delivery in the future. A minority has moved away from proctored in-class exams entirely and simply requires online submissions at any point up to the due date.

In current exam practice, off-site locations once again inadvertently underscore the importance of technical requirements: External providers will need to guarantee a stable internet connection and offer browser lockdown options. For most non-educational institutions, this capacity is largely non-existent. Furthermore, faculty would need pen-and-paper backups, a precaution seen as necessary but redundant, creating unnecessary paper waste. These factors make offsite exam locations problematic, leading some faculty with procedural exams (math, statistics, accounting) to consider a return to pen-and-paper exams. This, naturally, would eliminate any potential for IT system failure.

A blanket proposal to aim for higher taxonomies in the name of constructive alignment focuses too narrowly on efforts to increase academic validity in high-stakes exams. In doing so, it ignores equally relevant factors in institutional exam management such as reliability, feasibility and variance. It also completely underestimates the effort needed by faculty, staff and IT support to ensure proper examination conditions. In the current technological and legislative environment at the Lucerne University of Applied Sciences and Arts and beyond, e-proctoring does not seem to be the panacea educational institutions had hoped for, often further burdening faculty exam workflows without fully achieving exam integrity.

The extent to which our university's experience resembles that of other institutions – in the US or in international locations – is unclear. What we feel comfortable speculating, however, is that no university remained untouched by the Corona pandemic. Looking back on how decisions were made and practices adapted during the hectic days of the "Corona semesters" is a valuable exercise. It is, therefore, our hope that our contribution will inspire others to do the same.

6 ACKNOWLEDGMENT

The Authors would like to thank the management and faculty at Lucerne School of Business for their time and support for this project, including in the form of an IKM publication grant. We declare that there is no financial or other conflict of interest.

7 REFERENCES

[1] Alotaibi, S.R. A novel framework of success using of e-assessment during corona pandemic. International Journal of Emerging Technologies in Learning, 16(12), pp. 215–232, 2021. https://doi.org/10.3991/ijet.v16i12.22063

[2] Alghamdi, A.A., Alanezi, M.A. and Khan, F. Design and implementation of a computer aided intelligent examination system. International Journal of Emerging Technologies in Learning (iJET), 15(1), pp. 30–44, 2020. https://doi.org/10.3991/ijet.v15i01.11102

[3] Agneseus, T., Keimer, I. and Kuechler, C. Interactive PDF forms to conduct online examinations in university education: Practical experience and lessons learned. International Journal of Emerging Technologies in Learning, 17(11), pp. 4–16, 2022. https://doi.org/10.3991/ijet.v17i11.27905

[4] Roberts, F., Thomas, C.H., Novicevic, M.M., Ammeter, A., Garner, B., Johnson, P., and Popoola, I. Integrated moral conviction theory of student cheating: An empirical test. Journal of Management Education, 42(1), pp. 104–134, 2018. https://doi.org/10.1177/1052562917710686

[5] Stark, G. Stop "going over" exams! The multiple benefits of team exams. Journal of Management Education, 30(6), pp. 818–827, 2006. https://doi.org/10.1177/1052562906287965

[6] Bloom, B. *Taxonomy of educational objectives: The classification of educational goal*s. Handbook I: Cognitive domain, New York, Toronto: Longmans, Green, 1956.

[7] Biggs, J. *Enhancing teaching through constructive alignment*, Higher Education, 32(3), pp. 347–364, 1996. https://doi.org/10.1007/BF00138871

[8] Biggs, J. and Tang, C. *Teaching for quality learning at university: what the student does*, Maidenhead: Open University Press, 2011.

[9] Hochschulforum Digitalisierung, *Digitale Prüfungen in der Hochschule*, Whitepaper einer Community Working Group aus Deutschland, Österreich und der Schweiz, September 2021.

[10] Cramp, J., Medlin, J.F., Lake, P. and Sharp, C. Lessons learned from implementing remotely invigilated online exams. Journal of University Teaching & Learning Practice, 16(1), 2019. https://doi.org/10.53761/1.16.1.10

[11] Khan, S.K. Online assessments: Exploring perspectives of university students. Education and Information Technologies, 24, pp. 661–677, 2019. https://doi.org/10.1007/s10639-018-9797-0

[12] Wibowo, S., Grandhi, S., Chugh, R. and Sawir, E. A pilot study of an electronic exam system at an Australian university. Journal of Educational Technology Systems, 45(1), pp. 5–33, 2016. https://doi.org/10.1177/0047239516646746

[13] Abdelrahim, Y. *How COVID-19 quarantine influenced online exam cheating: A case of Bangladesh University students.* Journal of Southwest Jiaotong University, 56(1), 2021. https://doi.org/10.35741/issn.0258-2724.56.1.18

[14] Edwards, M, Leigh, J. *What's the plan? Some ideas about JME's strategy and preparations for post-pandemic teaching & learning.* Journal of Management Education, 45(4), pp. 523–534, 2021. https://doi.org/10.1177/10525629211022418

[15] Holden, O., Norris, M. and Kuhlmeier, V. Academic integrity in online assessment: A research review. Frontiers in Education. Frontiers, 2021. https://doi.org/10.3389/feduc.2021.639814

[16] San Jose, A.E. Academic integrity of students during the COVID-19 pandemic: A mixed method analysis. European Journal of Education and Pedagogy, 3(4), pp. 97–103, 2022. https://doi.org/10.24018/ejedu.2022.3.4.400

[17] Janke, S., Rudert, S., Petersen, A., Fritz, T. and Daumiller, M. Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity. Computers and Education Open, 2, 2021. https://doi.org/10.1016/j.caeo.2021.100055

[18] Braselmann, S., Mathieson, J. and Moisich, O. *Multimodal take-home exams in online teaching and beyond: constructive and professional alignment in teacher education*, Prüfen im Kontext kompetenzorientierter Hochschulbildung, ZFHE.AT, pp. 87–102, https://doi.org/10.3217/zfhe-17-01/06 March 17 / 1 2022.

[19] Whetten, D.A. Republication of "Principles of effective course design: What I wish I had known about learning-centered teaching 30 years ago". Journal of Management Education, pp. 834–854, 2021. https://doi.org/10.1177/10525629211044985

[20] Mackay, A. and Munoz, J. An online testing design choice typology towards cheating threat minimisation, Journal of University Teaching & Learning Practice, 16(3), 2019. https://doi.org/10.53761/1.16.3.5

[21] McBeath, R. J. (Ed). Instructing and evaluating in higher education: A guidebook for planning learning outcomes. Englewood Cliffs, NJ: Educational Technology Publications, 1992.

[22] Anderson, L.W. and Krathwohl, D.R. (Eds.). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. San Francisco: Longman, 2001.

[23] Steger, D., Schroeders, U. and Gnambs, T. *A meta-analysis of test scores in proctored and unproctored ability assessments.* European Journal of Psychological Assessment, 36, pp. 1–11, 2018. https://doi.org/10.1027/1015-5759/a000494

[24] Dreyfus, S. *The five-stage model of adult skill acquisition.* Bulletin of Science, Technology & Society, 24(3), pp. 177–181, 2004. https://doi.org/10.1177/0270467604264992

[25] Kaiser, H. *Situationsdidaktik Konkret,* Bern: Hep Verlag, 2019.

[26] Kaiser, H. Kompetenz: *Versuch einer Arbeitsdefinition, V7*: Kanton Solothurn, 2003.

[27] Hof, C. *Von der Wissensvermittlung zur Kompetenzentwicklung,* Literatur- und Forschungsreport Weiterbildung, pp. 80–89, 2002.

[28] Walzik, S., *Kompetenzorientiert prüfen*, Opladen & Toronto: Verlag Barbara Budrich, 2012.

[29] Gerick, J. Sommer, A. and Zimmermann, G. (Eds.). J. Kompetent Prüfungen gestalten, Münster: Waxmann, 2018. https://doi.org/10.36198/9783838548401

[30] Gisev, N., Bell, S. and Chen, T. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy, 9(3), pp. 330–338, 2013. https://doi.org/10.1016/j.sapharm.2012.04.004

[31] Keller, C. and Kros, J. An innovative Excel application to improve exam reliability in marketing courses. Marketing Education Review, 21(1), pp. 21–28, 2011. https://doi.org/10.2753/MER1052-8008210103

[32] Armstrong, David, et al. The place of inter-rater reliability in qualitative research: An empirical study. Sociology, 31(3), pp. 597–606, 1997. https://doi.org/10.1177/0038038597031003015

[33] Wilmot, M.P., Wiernik, B.M. and Kostal, J.W. Increasing interrater reliability using composite performance measures. Industrial and Organizational Psychology, 7(4), pp. 539–542, 2014. https://doi.org/10.1111/iops.12192

[34] Gittinger, F., et al. Interrater reliability in the assessment of physiotherapy students. BMC Medical Education, 22(1), pp. 1–10, 2022. https://doi.org/10.1186/s12909-022-03231-y

[35] Li, et al. Optimized collusion prevention for online exams during social distancing. NPJ Sci. Learn. 6(5), 2021. https://doi.org/10.1038/s41539-020-00083-3

[36] Forray, K., Jeanie, M. *A silver linings playbook, COVID-19 edition.* Journal of Management Education, 44(4), pp. 399–405, 2020. https://doi.org/10.1177/1052562920931901

[37] Karch, B. *Online proctoring services: Insights from North America*, Swissnex in Boston, Commissioned by Hochschule Luzern, 2022.

[38] Reisenwitz, T.H. Examining the necessity of proctoring online exams. Journal of Higher Education Theory and Practice, 20(1), pp. 118–124, 2020. https://doi.org/10.33423/jhetp.v20i1.2782

[39] Hillier, M., Fluck, A. Arguing again for e-exams in high stakes examinations, in Australasian Society for Computers in Learning in Tertiary Education, Sydney, 2013.

[40] Mata, J.R. "How to teach online? Recommendations for the assessment of online exams with University students in the USA in times of pandemic." IJERI: International Journal of Educational Research and Innovation, pp. 188–202, 2020. https://doi.org/10.46661/ijeri.5003

[41] Sattler, D., et al. Grant peer review: Improving inter-rater reliability with training. PLoS ONE, 10(6), p. e0130450, 2015. https://doi.org/10.1371/journal.pone.0130450

[42] King, N. Doing template analysis in qualitative organizational research: Core methods and current challenges, London, Sage, pp. 426–439. 2012, https://doi.org/10.4135/9781526435620.n24

[43] Mayring, P.F.T. Qualitative Inhaltsanalyse, in Handbuch Methoden der empirischen Sozialforschung, Wiesbaden, Springer, 2014. https://doi.org/10.1007/978-3-531-18939-0_38

[44] Orthaber, M., Stütz, D., Antretter, T. and Ebner, M. Concepts for e-assessments in STEM on the example of engineering mechanics. International Journal of Emerging Technologies in Learning, 15(12), pp. 136–15, 2020. https://doi.org/10.3991/ijet.v15i12.13725

[45] Sangwin, C. (2013). Computer Aided Assessment of Mathematics Using Stack. Oxford: Ox-ford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

[46] Weigel, M., Hübl, R., Podgayetskaya, T. and Derr, K. (2018). Potential von STACK-Aufga-ben im formativen eAssessment: Automatisiertes Feedback und Fehleranalyse. In Fach-gruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), Beiträge zum Ma-thematikunterricht (S. 1419–1422). Münster: WTM.

[47] Weigel, M., Derr, K., Hübl, R. and Podgayetskaya, T. (2019). STACK-Aufgaben im formativen eAssessment: Einsatzmöglichkeiten des Feedbacks. In Contributions to the 1st International STACK conference 2018. Friedrich-Alexander-Universität, Nürnberg, Germany.

[48] Akour, M. Alenezi, M. Higher education future in the era of digital transformation. Educ. Sci., 12, p. 784, 2022. https://doi.org/10.3390/educsci12110784

8 AUTHORS

Douglas MacKevett, MA (ED) is Head of Digital Learning Services (faculty development) at the Lucerne School of Business, Switzerland. A native of the USA, MacKevett has studied in both the US and UK at graduate level. He has taught at the Lucerne of School of Business in Switzerland since 2001 in the Bachelor, Master and Executive Education programs and has most recently headed up the faculty development program in blended teaching and learning. He also co-heads the MScBA Online Business & Marketing. His current research interests include digital exams and hybrid teaching.

Martin Gutmann, Ph.D. is Professor of Management at the Lucerne School of Business, Switzerland. A historian by training, Gutmann's research and teaching focus on the historical perspectives on contemporary challenges. His most recent book is *Before the UN Sustainable Development Goals. A Historical Companion* (Oxford University Press, 2022).