The right amount of pressure: implementing time pressure in online exams

Matthias Stadler, Nicola Kolb, Michael Sailer

Angaben zur Veröffentlichung / Publication details:

Stadler, Matthias, Nicola Kolb, and Michael Sailer. 2021. "The right amount of pressure: implementing time pressure in online exams." *Distance Education* 42 (2): 219–30. https://doi.org/10.1080/01587919.2021.1911629.

The right amount of pressure: Implementing time pressure in online exams

Matthias Stadler, Nicola Kolb, and Michael Sailer

Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Beginning in December 2019, a new type of coronavirus spread worldwide, becoming a pandemic. To limit the spread almost all universities stopped classroom teaching and moved to online teaching (Telles-Langdon, 2020). However, the virus continues to spread, online teaching is still ongoing, and universities now have to consider how to deal with exams, asking whether they want to postpone, eliminate them, or conduct them online (Crawford et al., 2020).

As online exams are at risk of facilitating student dishonesty by cheating, one possible solution is to proctor students while they are taking exams (Fask et al., 2014). As proctoring is an expensive solution to minimize the risk of student cheating and also questionable for reasons of data protection, Cluskey et al. (2011) published recommendations on how proctoring can function without a machine or a person who watches over the students during the entire exam period and is also cost-effective at the same time. The recommendations consisted of eight control procedures (Cluskey et al., 2011).

The control procedures describe how to present the exams to students digitally and how to display their questions (Cluskey et al., 2011). However, Cluskey et al. did not specify their recommendations on how time pressure should be implemented. To provide further insights into this issue, this paper reports on the results of an experiment comparing different types of time pressure; namely, the same amount of time for each question, different amounts of time for each question, a total amount of time for the entire exam, rather than breaking time periods down per question. This study aimed to ascertain the

CONTACT Matthias Stadler Matthias.Stadler@lmu.de

best type of time pressure so as to discourage student dishonesty and, at the same time, to make examinations as comfortable as possible for them.

Online exams

While multiple types of electronic exams exist (Bohmer et al., 2018), this study focused on online exams that allow students to take the exam at any place using their own devices. Students writing exams in the comfort of their own home have multiple advantages. Generally, students might be more relaxed in their familiar surroundings, which may lead to increased performance (Becker et al., 2019; Schult et al., 2017), and students with disabilities may experience fewer obstacles to full participation (Kotera et al., 2019; Shevlin et al., 2004). However, studies have also demonstrated that online examinations may increase students' dishonest behavior (Okada et al., 2019). Students may use external sources, such as books, notes, or the Internet, or even collude with other students to cheat on the exam (Ullah et al., 2016).

This possibility of dishonest behavior has resulted in research on proctoring or surveillance during online exams (Sarrayrih & Ilyas, 2013). These resources are often challenging to implement (Okada et al., 2019) or may conflict with existing data protection laws (Bohmer et al., 2018). Cluskey et al. (2011) suggested several online exam control procedures that allow some degree of security without the need for additional proctoring. They suggested that online exams should be multiple-choice questions that are open book, in random order without the opportunity to revisit tasks, and to be solved under time pressure. Additionally, exam questions should be presented one at a time and be accessible only within a specific timeframe and just once on a specific date. Apart from these control procedures, Cluskey et al. (2011) also suggested using a specific Internet page called Blackboard's Respondus Lockdown Browser (Cluskey et al., 2011, p. 5); they also suggested changing one-third of the questions for each iteration of the exam. Neither of these suggestions are relevant to this study.

Open-book testing (i.e., examinations that allow students to use all available resources other than colluding with other students) has long been proposed for testing in online learning environments (Rakes, 2008). Studies have shown that open-book exams may enhance student learning and motivation (e.g., Green et al., 2016); yet there are also opposing results (for an overview, see Durning et al., 2016). From the perspective of online exams, open-book testing provides a solution to the problem of cheating using external sources, thus limiting the need for proctoring.

However, collusion among students remains a more critical problem for online exams (Ullah et al., 2016). Students are well connected via social media, and without proper proctoring, scores might reflect the knowledge of the class's most capable students (Best & Shelley, 2018). This threat can be partially addressed by providing the tasks in random order so that students will not work on the same tasks simultaneously. This strategy can be effective only if students cannot revisit tasks they have already seen as they could otherwise view all tasks before answering the first, rendering the randomization effectively moot.

Finally, both open-book testing and task randomization protect against academic dishonesty only if there is limited time to work on the tasks. Cluskey et al. (2011) recommended allowing a fixed time to answer all tasks. However, this time constraint may lead to some students not seeing all the exam questions; that is, they may not complete all the tasks. Alternatively, it would be possible to allow for a specific time for each exam question so that all

students would see all questions. This would assume that all tasks are of similar length and difficulty and can be answered in approximately the same time. A final operationalization of time pressure in online exams could, therefore, allow for specific times per exam question but have the questions differ based on specific properties.

Time pressure in exams

Introducing any form of time pressure to an exam will impact the exam's measurement quality. On a general level, time pressure shifts the exam's measurement, from measuring students' maximal ability (knowledge) toward measuring their ability to answer quickly (speed). Although these two indicators of ability may be related, they are certainly not identical (Partchev et al., 2013). Adding a measure of speed, thus an additional ability, to an exam will inevitably impact the exam's difficulty as experienced by students (Lindner et al., 2018; Perlini et al., 1998). This increase in difficulty is particularly evident for students with specific learning difficulties (e.g., dyslexia), who tend to suffer substantially from increased time pressure (e.g., Waterfield & West, 2006). However, even for typically developing students, individual performance is consistently worse under time pressure as students need to apply less advanced heuristics to solve tasks (Rieskamp & Hoffrage, 2008). For instance, students differ in how much time they allocate to specific exam tasks (Son & Metcalfe, 2000). However, time pressure limits the possibility of choosing to spend more time on tasks students consider more difficult. The negative effect of increased time pressure on task difficulty seems stronger for women than for men, with women performing substantially worse under time pressure than men (De Paola & Gioia, 2016; Steinmayr & Spinath, 2019; Voyer, 2011).

These objective impacts of time pressure on exam performance are also very likely affecting students' subjective evaluations of the exam, such as the exam's perceived fairness or the perceived possibility to demonstrate ability. This fact is essential to consider as students' subjective evaluations of an exam influence their performance (Lindner et al., 2018). Thiede (1996) reported results suggesting that students' performance is even more affected by the anticipated test format than by anticipated test difficulty. Therefore, any considerations on the effects of time pressure should also include how the implementations of time pressure will influence students' subjective evaluations of the exams.

This study

The literature reported above indicates that while the comfort for students rises with online exams taken in familiar surroundings, such exams also simplify the chance for academic dishonesty, which can then be controlled with time pressure. Introducing time pressure to an exam may, despite several drawbacks, help solve this issue. We conducted an experiment comparing three different implementations of time pressure in an online exam. Students worked on a set of tasks either within a (1) fixed time, choosing freely how much time to spend on each task, within a (2) fixed but equal time for each question, or within a (3) fixed time that varied for each task based on the length of text for the task.

In our first research question (RQ1), we investigated whether students' objective performance (i.e., their scores) would differ across the conditions. Based on previous research (Lindner et al., 2015), we assumed the tasks with time pressure would be more difficult than

without any time pressure but had no clear hypothesis about whether any condition might be most challenging.

In addition, we investigated whether students' subjective evaluation of the exam would differ across the three conditions (RQ2). We did not have any a priori hypotheses as to whether students would prefer any of the conditions over the others.

Finally, we investigated how much time students allocate to working on the tasks when choosing freely compared to the other two conditions (RQ3). Again, we did not have any a priori hypotheses on whether students spend more or less time on a task when choosing freely than in any of the other conditions.

Method

Sample

The sample consisted of 111 students from Ludwig-Maximilians-Universität München (a large German university) taking a class on empirical research methods. Students were either enrolled in educational science or prevention, inclusion, and rehabilitation. The majority of students were in their first year of university. Of the 111 students, 92 (82.9%) reported their gender as female, 16 (14.4%) as male, and three (2.7%) as diverse. Their mean age was 22.3 years (SD = 3.62).

Participation was anonymous, voluntary, and could be ended at any point without the need for justification. There were no incentives to participating. The tasks and the correct solutions were published online after the data collection was complete. All participants provided informed consent to the data collection and its publication before any data was recorded.

Procedure

The entire study was conducted online using the SoSci Survey website (https://www.soscisurvey.de). Data collection took place in July 2020. Participation was voluntary and could be stopped by the participant at any point without explanation. Students were randomly assigned to one of three conditions, which were one-time (n = 32), equal-time (n = 37), and different-times (n = 42). At the beginning of the study, students were asked to answer demographic questions and accept the data privacy statement. Next, students were asked to solve six application-oriented multiple-choice questions in open-book format from their current curriculum (empirical research methods) presented in random order. The total time available for these tasks was 480 s; however, the time differed according to the experimental condition. Students in the one-time condition saw a timer counting down the full 480 s and could choose how much time they spent on each task. After completing a task, students in this condition could press a button to access the next task but going back to the question was impossible. Students in the equal-time condition saw a timer counting down 80 s for each task, after which the next task would start automatically. Students in this condition could not move on to the next task on their own. Students in the different-times condition saw a timer that counted down a specific time for each task relative to the number of words in the text (instruction and response options). The time varied between 37 and 172 s. As with the equal-time condition, students did not move on to the next task independently but were directed to the next task automatically after the time allotted to the task was up. After completing all tasks

or time-out in the one-time condition, students were asked to rate their experience with the exam. These questions on subjective experience were identical across all three conditions and had no time limit.

Measures

Score

To estimate objective performance, the students worked on six different single-choice questions within their current curriculum. All items were worded in German but are translated to provide examples. Responses were scored as zero for any incorrect response and 1 for the correct response, with the sum of all tasks representing the final score.

Subjective experience

In addition to the objective score, we asked students to rate their subjective experience of the exam on six dimensions.

Low effort. To measure students' estimates of how much effort was required by exams in the three different formats, we used a scale by Linder et al. (2018). Students rated their agreement to four statements on a Likert scale from 1, *Does not apply* to 4, *Fully applies*. An example is "Exams in that form require low effort from me." The scale showed acceptable reliability of $\alpha = 0.72$.

Potential to show ability. To measure students' estimates of their potential to show their ability in the exams in the three different formats, we used a scale by Lindner et al. (2018). Students rated their agreement to four statements on a Likert scale from 1, *Does not apply* to 4, *Fully applies*. An example is "Exams in that form allow me to show my exact knowledge." The scale showed good reliability of $\alpha = .81$.

Objectivity. To measure students' estimates of how objectively exams in the three different formats measured their performance, students rated their agreement to a statement on a Likert scale from 1, *Does not apply* to 4, *Fully applies*. The statement was "Exams in that form are scored very objectively," which was used as an item by Lindner et al. (2018).

Fairness. To measure students' estimates of how fair the exams in the three different formats were, students rated their agreement to a statement on a Likert scale from 1, *Does not apply* to 4, *Fully applies*. The statement was, "Exams in that form are fair," which was used in the study of Lindner et al. (2018).

Familiarity. To measure students' estimates of how familiar exams were in the three different formats, we used a scale by Lindner et al. (2018). Students rated their agreement to three statements on a Likert scale from 1, *Does not apply* to 4, *Fully applies*. An example is "I am familiar with exams in that form." The scale showed good reliability of $\alpha = 0.84$.

Preference. To measure students' estimates of how their preference was by exams in the three different formats, students rated their agreement to two statements on a Likert

scale from 1, *Does not apply* to 4, *Fully applies*. One statement was "Exams in that form suit me"; the other was "I would like more of my exams to be provided in that form," which were used as a scale by Lindner et al. (2018). The scale showed high reliability of α = 0.94.

Statistical analysis

All analyses were conducted using JAMOVI (https://www.jamovi.org). To compare the three conditions as to objective scores and subjective evaluations, we computed ANOVAs and robust ANOVAs with post hoc tests following all significant main effects (Wilcox, 2012). To find the difference between the time on task and the three different conditions, we used the one-sample *t* test and compared the one-time condition with the equal-time condition and the one-time condition with the different-times condition. Due to our study's exploratory nature, all tests were conducted two-sided with an alpha level of 5%.

Results

Descriptive statistics

Table 1 displays the mean, standard deviation, range, and bivariate correlations for all variables used in the study. Both objective scores and subjective evaluations were generally low, with the notable exception of students' evaluation of the exams' objectivity (M = 3.34; SD = 0.82). Regarding the bivariant correlations, there is no significant correlation between the score and the subjective scales. Notably, students preferred the exams if they considered them to require low effort (r = .57), allowed them to show their ability (r = .67), and seemed fair (r = .53). Shapiro-Wilk tests indicated a non-normal distribution for all scales (Table 1).

RQ1 and RQ2: Objective score and subjective evaluations

To compare objective scores, we compared the average scores across the three different conditions, which are the one-time condition (M = 1.78; SD = 1.24), the equal-time condition (M = 1.78; SD = 1.18) and the different-times condition (M = 1.21; SD = 1.14). We found a small but statistically significant main effect for the robust ANOVA (F(2, 108) = 3.81; p = .030; η^2 = .053). The post hoc tests showed that there was a significant difference between the one-time condition and the different-times condition in the score (p = .021) as well as between the equal-time condition and the different-times condition (p = .016). There was no significant difference between the one-time condition and the equal-time condition (p = .524).

Table 1. Descriptive statistics and bivariate correlations for all scales.

Variable	M	SD	Range	$p_{Shapiro\text{-}Wilk}$	1	2	3	4	5	6
1 Score	1.57	1.20	0-6	< .001	-					
2 Low effort	1.67	0.576	1-3.25	< .001	.09	-				
3 Show ability	2.13	0.770	1-4	.014	-.01	.41*	-			
4 Objectivity	3.34	0.824	1-4	< .001	.17	.05	.17	-		
5 Fairness	2.61	0.997	1-4	< .001	.03	.37*	.62*	.26*	-	
6 Familiarity	1.83	0.966	1-4	< .001	-.10	.31*	.26*	-.07	.25*	-
7 Preference	1.99	0.905	1-4	< .001	-.08	.57*	.67*	.11	.53*	.30*

Note. M = mean; SD = standard deviation; * p < 0.05

Table 2 displays ANOVA results for all scales. Besides the conventional ANOVA, we also conducted robust ANOVA, as there was a violation of the assumption of normality for all scales (Table 1). Regarding students' subjective evaluations of the different exams, we found no significant main effects for any of the scales.

RQ3: Time on task

To analyze time-on-task differences between the three experimental conditions, we conducted two different sets of one-sample t test. When the Shapiro Wilk test indicated a non-normal distribution, we performed Wilcoxon rank tests. We compared the time spent on each task in the one-time condition to the time for the equal-time condition (80 s) and the times allotted for the different-times condition. Students spent approximately 80 s on Task 1 and significantly less time on Tasks 2–6 (Table 3). The comparison between the average time spent on each task in the different-times condition to the one-time condition showed significant differences for Task 1 ($p < .001$), Task 2 ($p = .024$), Task 5 ($p = .021$) and Task 6 ($p < .001$). There were no significant differences between Tasks 3 and 4 in the different-times and the one-time conditions. The results are summarized in Table 3.

Discussion

We have reported the experiment's results comparing three different implementations of time pressure in an online exam regarding students' objective performance, subjective evaluations, and time allocated to the exam's tasks. Students worked on a set of tasks either within a (1) fixed time, choosing freely how much time to spend on each task, within a (2) fixed but equal time for each question, or within a (3) fixed time that varied for each task based on the number of words in the task instruction and response options.

Table 2. ANOVA results comparing the three experimental conditions for all scales.

	ANOVA				Robust ANOVA		
Variable	*F*	df_1	df_2	*p*	F_{robust}	p_{robust}	η^2
Score	3.02	2	108	.053	3.81	.030	.053
Low effort	0.204	2	92	.816	0.281	.757	.004
Show ability	0.477	2	92	.622	0.972	.388	.010
Objectivity	0.606	2	91	.548	0.210	.812	.013
Fairness	0.348	2	91	.707	0.723	.493	.008
Familiarity	0.676	2	93	.511	0.378	.688	.014
Preference	0.243	2	91	.785	0.704	.501	.005

Note. *df* = degrees of freedom

Table 3. Results of one-sample t tests for time-on-task.

	One-time		Comparison with equal-time				Comparison with different-times			
Task	*M*	*SD*	Time	estimate	*p*	*d*	Time	estimate	*p*	*d*
Task 1	83.8	50.2	80	.415	.681	0.08	172	-9.63	< .001	-1.76
Task 2	38.7	23.4	80	9.00*	< .001	-1.77	46	123.00*	.024	-0.31
Task 3	51.7	38.2	80	76.00*	.001	-0.74	61	138.00*	.053	-0.24
Task 4	45.4	32.6	80	40.00*	< .001	-1.06	37	268.00*	.471	0.26
Task 5	55.8	31.3	80	-4.17	< .001	-0.78	70	-2.45	.021	-0.46
Task 6	62.3	39.1	80	111.00*	.007	-0.45	95	67.00*	< .001	-0.84

Note. M = mean; SD = standard deviation; estimate = *t* value or *Wilcoxon; *df* = degrees of freedom; *d* = effect size

In line with previous research, we found that the tasks were generally challenging under all conditions (Lindner et al., 2018; Perlini et al., 1998). Students performed least well if the time was based on the number of words in the task instruction and response options. The other two conditions did not differ regarding students' performance.

Subjectively, students considered all three conditions to be difficult, albeit relatively fair and highly objective. The subjective evaluations did not differ across conditions. Interestingly, students who could allocate their time freely (Condition 1) did not spend more time on any of the tasks than they did in the condition with fixed but equal times (Condition 2). If based on the number of words in the task instruction and response options (Condition 3), times were either too long or too short for most tasks compared to the free allocation in Condition 1.

Implications

Our results suggest several aspects to be considered when moving from traditional, place-based to online exams. First, any form of time pressure will introduce a measure of speed, thus increasing the exam's difficulty (Lindner et al., 2018; Perlini et al., 1998). This needs to be considered in constructing the exam to avoid bottom effects, with most students receiving low scores.

In implementing the time pressure, free allocation of time or fixed but equal amounts of time seem superior to differing amounts of time. This may be different if there is an exact model of the tasks' difficulty allowing the prediction of the required time for most students (Ha et al., 2019), but merely adapting the time to the number of words in the task instruction and response options seems inadequate, given that challenging tasks may be quite short while relatively easy tasks may have an extended instruction. Comparing the remaining two conditions, it seems surprising that students did not perform better when they could allocate their time freely as this condition should have allowed them to choose more advanced heuristics in answering difficult questions (Son & Metcalfe, 2000). However, due to the task randomization and the missing option of going back to previous tasks, students did not have full information on the scope or relationship between questions and had to consider each task individually. Correspondingly, students spent a minimal amount of time necessary to solve each task (Edland & Svenson, 1993; Svenson & Maule, 1993).

Since students did not prefer any of the conditions over the others, having the same amount of time for each task seems to be the best option. In the case of substantial differences in task difficulty or other constraints, this approach could be amended by having multiple blocks of randomly selected tasks with the same amount of time for each block.

Limitations

Several limitations should be considered in interpreting our results. First, we could not differentiate the effect of the different implementations of time pressure for men and women due to the massive overrepresentation of women in our sample. The sample distribution is typical for educational sciences in German universities, but time pressure has been found to have a more substantial negative effect for women than for men (Steinmayr & Spinath, 2019; Voyer, 2011). Future studies should investigate the

moderating effect of gender on the different implementations of time pressure on performance and students' subjective evaluations of the exams. Otherwise, test fairness could be limited with the exam disadvantaging women.

Also, it needs to be noted that the exams in our study were low-stakes occasions with no disadvantage from poor performance. However, ethical concerns prohibited conducting our study in a realistic high-stakes situation. In a high-stakes exam, students might have been more motivated to perform better (Weiner & Hurtz, 2017), but this factor should not have impacted subjective evaluations. A high-stakes exam would also be a lot longer (usually about 80 min), potentially prompting different reactions to the different exam conditions. On the one hand, a longer duration might result in some familiarity with the exam (Boevé et al., 2015). Also, more tasks would increase the observed score's reliability. On the other hand, the different conditions may well vary with regard to student fatigue or the build-up of anxiety over the duration of a full exam (Jensen et al., 2013). Finally, the study was carried out in one subject only, and it is unclear how well the results generalize to other subjects with other exam conditions.

Future studies should, therefore, try to corroborate our findings in a more realistic exam setting and extend the samples to various kinds of learners (at school and university) and various kinds of subjects (Rummer et al., 2019).

Conclusion

In conclusion, Cluskey et al.'s (2011) recommendations seem adequate to minimize academic dishonesty in unproctored online exams. The way time pressure is implemented is essential, though, as it has potential impacts on students' performance. Based on our results, we suggest having equal times for all tasks. The implementation with equal times should be preferred to free time allocation as students tend to use less time than available when allocating freely to deal with the uncertainty of unknown future tasks. Generally, time pressure should be used with caution, as it impacts the tasks' difficulty and may disadvantage students with specific learning difficulties. For students who declare their difficulties, access arrangements such as additional time may remedy this disadvantage (Duncan & Purcell, 2017), but not all students declare their difficulties or are even aware of them (Richardson, 2009).

Moving from in-person exams to online exams is a great challenge for students. We hope that this paper will provide some guidance to university teachers and help them create pedagogically valid exams while allowing students to demonstrate their knowledge and abilities optimally.

Disclosure statement

No potential conflict of interest was declared by the authors.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (for2385; COSIMA; Teilprojekt M).

Notes on contributors

Matthias Stadler is an assistant professor for educational psychology at Ludwig-Maximilians-Universität München. His research interests are the educational applications of computer-based assessment and the analysis of behavior in complex learning environments.

Nicola Kolb is a student at Ludwig-Maximilians-Universität München. Her major is educational science. At the moment, she is preparing her bachelor's thesis. In addition to her academic studies, she works for a research project.

Michael Sailer is an assistant professor for education and educational psychology at Ludwig-Maximilians-Universität München. He is currently conducting research about gamified learning, simulation-based learning, and the use of digital technology in education.

ORCID

Matthias Stadler http://orcid.org/0000-0001-8241-8723
Nicola Kolb http://orcid.org/0000-0002-9180-8413
Michael Sailer http://orcid.org/0000-0001-6831-5429

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. They are not publicly available due to their containing information that could compromise the privacy of the research participants.

References

Becker, N., Koch, M., Schult, J., & Spinath, F. M. (2019). Setting doesn't matter much, *European Journal of Psychological Assessment*, *35*(3), 309–316. https://doi.org/10.1027/1015-5759/a000402

Best, L. M., & Shelley, D. J. (2018). Academic dishonesty: Does social media allow for increased and more sophisticated levels of student cheating? *International Journal of Information and Communication Technology Education*, *14*(3), 1–14. https://doi.org/10.4018/IJICTE.2018070101

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS One, 10* (12), Article e0143616. https://doi.org/10.1371/journal.pone.0143616

Bohmer, C., Feldmann, N., & Ibsen, M. (2018). E-exams in engineering education — online testing of engineering competencies: Experiences and lessons learned. In *Proceedings of the 2018 IEEE Global Engineering Education Conference* (pp. 571–576). IEEE. https://doi.org/10.1109/EDUCON.2018.8363281

Cluskey, G. R., Ehlen, C. R., & Raiborn, M. H. (2011). Thwarting online exam cheating without proctor supervision, *Journal of Academic & Business Ethics*, *4*, 1–7. https://www.aabri.com/manuscripts/11775.pdf

Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P. A., & Lam, S. (2020). COVID-19: 20 countries' higher education intra-period digital pedagogy responses, *Journal of Applied Learning & Teaching*, *3*(1), 1–20. https://doi.org/10.37074/jalt.2020.3.1.7

Duncan, H., & Purcell, C. (2017). Equity or Advantage? The effect of receiving access arrangements in university exams on humanities students with specific learning difficulties (SpLD), *Widening Participation and Lifelong Learning*, *19*(2), 6–26. https://doi.org/10.5456/WPLL.19.2.6

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Jr., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review, *Academic Medicine*, *91*(4), 583–599. https://doi.org/10.1097/ACM.0000000000000977

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 27–40). Springer, https://doi.org/10.1007/978-1-4757-6846-6_2

Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment, *Journal of Academic Ethics*, *12*(2), 101–112. https://doi.org/10.1007/s10805-014-9207-1

Green, S. G., Ferrante, C. J., & Heppard, K. A. (2016). Using open-book exams to enhance student learning, performance, and motivation, *Journal of Effective Teaching*, *16*(1), 19–35. https://uncw.edu/jet/articles/vol16_1/index.htm

Ha, L. A, Yaneva, V., Baldwin, P., & Mee, J. (2019). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 11–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4402

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PloS One, 8*(8), Article e70270, https://doi.org/10.1371/journal.pone.0070270

Kotera, Y., Cockerill, V., Green, P., Hutchinson, L., Shaw, P., & Bowskill, N. (2019). Towards another kind of borderlessness: Online students with disabilities, *Distance Education*, *40*(2), 170–186. https://doi.org/10.1080/01587919.2019.1600369

Lindner, M. A., Mayntz, S. M., & Schult, J. (2018). Studentische Bewertung und Präferenz von Hochschulprüfungen mit Aufgaben im offenen und geschlossenen Antwortformat [Students' evaluations of and preferences in exams with open-ended and closed-ended questions in higher education], *Zeitschrift Für Pädagogische Psychologie*, *32*(4), 239–248. https://doi.org/10.1024/1010-0652/a000229

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

Okada, A., Whitelock, D., Holmes, W., & Edwards, C. (2019). E-authentication for online assessment: A mixed-method study, *British Journal of Educational Technology*, *50*(2), 861–875. https://doi.org/10.1111/bjet.12608

De Paola, M., & Gioia, F. (2016). Who performs better under time pressure? Results from a field experiment, *Journal of Economic Psychology*, *53*, 37–53. https://doi.org/10.1016/j.joep.2015.12.002

Partchev, I., Boeck, P. de, & Steyer, R. (2013). How much power and speed is measured in this test? *Assessment*, *20*(2), 242–252. https://doi.org/10.1177/1073191111411658

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty, *Canadian Psychology/Psychologie Canadienne*, *39*(4), 299–307. https://doi.org/10.1037/h0086821

Rakes, G. C. (2008). Open book testing in online learning environments, *Journal of Interactive Online Learning*, *7*(1), 1–9. http://www.ncolr.org/issues/jiol/v7/n1/open-book-testing-in-online-learning-environments.html

Richardson, J. T.E. (2009). The attainment and experiences of disabled students in distance education, *Distance Education*, *30*(1), 87–102. https://doi.org/10.1080/01587910902845931

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection, *Acta Psychologica*, *127*(2), 258–276. https://doi.org/10.1016/j.actpsy.2007.05.004

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00463

Sarrayrih, M. A., & Ilyas, M. (2013). Challenges of online exam, performances and problems for online university exam, *International Journal of Computer Science Issues (IJCSI)*, *10*(1), 439–443. https://www.ijcsi.org/articles/Challenges-of-online-exam-performance-and-problems-for-online-university-exam.php

Schult, J., Stadler, M., Becker, N., Greiff, S., & Sparfeldt, J. R. (2017). Home alone: Complex problem solving performance benefits from individual online assessment, *Computers in Human Behavior*, *68*, 513–519. https://doi.org/10.1016/j.chb.2016.11.054

Shevlin, M., Kenny, M., & Mcneela, E. (2004). Participation in higher education for students with disabilities: An Irish perspective, *Disability & Society*, *19*(1), 15–30. https://doi.org/10.1080/0968759032000155604

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204–221. https://doi.org/10.1037/0278-7393.26.1.204

Steinmayr, R., & Spinath, B. (2019). Why time constraints increase the gender gap in measured numerical intelligence in academically high achieving samples, *European Journal of Psychological Assessment*, *35*(3), 392–402. https://doi.org/10.1027/1015-5759/a000400

Svenson, O., & Maule, A. J. (Eds.). (1993). *Time pressure and stress in human judgment and decision making*. Springer.

Telles-Langdon, D. M. (2020). Transitioning university courses online in response to COVID-19, *Journal of Teaching and Learning*, *14*(1), 108–119. https://doi.org/10.22329/jtl.v14i1.6262

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance, *The Quarterly Journal of Experimental Psychology Section A*, *49*(4), 901–918. https://doi.org/10.1080/713755673

Ullah, A., Xiao, H., & Barker, T. (2016, October). A classification of threats to remote online examinations. In I. Staff (Ed.), *Proceedings of the Annual Information Technology, Electronics and Mobile Communication Conference* (pp. 1–7). IEEE, https://doi.org/10.1109/IEMCON.2016.7746085

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis, *Psychonomic Bulletin & Review*, *18*(2), 267–277. https://doi.org/10.3758/s13423-010-0042-0

Waterfield, J., & West, B. (2006). *Inclusive assessment in higher education: A resource for change*, University of Plymouth. https://www.plymouth.ac.uk/uploads/production/document/path/3/3026/Space_toolkit.pdf

Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams, *Journal of Applied Testing Technology*, *18*(1), 13–20. http://www.jattjournal.com/index.php/atp/article/view/113061

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Elsevier/Academic Press.