Educational Assessment, Evaluation and Accountability (2023) 35:503–523
https://doi.org/10.1007/s11092-023-09410-9

Large-scale diagnostic assessment in first-year university students: pre- and transpandemic comparison

Melchor Sánchez-Mendiola[1,2] · Abigail P. Manzano-Patiño[1] · Manuel García-Minjares[1] · Enrique Buzo Casanova[1] · Careli J. Herrera Penilla[1] · Katyna Goytia-Rodríguez[1] · Adrián Martínez-González[1,2]

Received: 25 October 2021 / Accepted: 30 May 2023 / Published online: 21 June 2023

Abstract

COVID-19 has disrupted higher education globally, and there is scarce information about the "learning loss" in university students throughout this crisis. The goal of the study was to compare scores in a large-scale knowledge diagnostic exam applied to students admitted to the university, before and during the pandemic. Research design was quasi-experimental with static group comparisons, taking advantage of the pandemic "natural experiment," to assess knowledge in students admitted to the National Autonomous University of Mexico. Four student cohorts were analyzed: 2017 and 2018 (prepandemic, paper-and-pencil exams), 2020 and 2021 (transpandemic, online exams). The same instruments were applied in each pair of cohorts (2017–2021; 2018–2020) to decrease instrumentation threat. Propensity score matching was used to create balanced comparable groups. 35,584 matched students from each of the 2018 and 2020 cohorts were compared and 31,574 matched students from each of the 2017–2021 cohorts. Reliability and point biserial correlation coefficients were higher in the transpandemic online applications. Knowledge scores were 2.3 to 7.1% higher in the transpandemic assessments, Spanish scores in the 2018–2020 comparison were 1.3% lower, and English results in 2021 were 7.1% lower than in 2017. Before the pandemic, there was a 3.1% higher test performance in men; this gap decreased to 0.34% during the pandemic. There was no documented learning loss in this large student population, with an increase in knowledge in the pandemic cohorts. Some influence in scores due to the online testing modality cannot be ruled out. Longitudinal follow-up is required to continue evaluating the impact of the pandemic in learning.

Keywords Diagnostic assessment · Learning loss · Large-scale testing · Assessment of learning · COVID-19

Extended author information available on the last page of the article

1 Introduction

Higher education lived a severe disruption in 2020 with the COVID-19 pandemic, when universities all over the world did their best to continue educational activities using online teaching and assessment (Hodges et al., 2020). The consequences of this prolonged crisis are still unknown; there is a need to obtain rigorous assessment data that can provide a glimpse of its impact on learning (IESALC-UNESCO, 2020; UNESCO, 2020). Analysis of assessment data obtained during the pandemic will provide a clearer picture of the size and direction of the "educational catastrophe" that some scholars and the media have predicted, so appropriate measures can be implemented (Jankowski, 2020; Lake & Olson, 2020; UNESCO, 2021). The term "pandemic learning loss" has appeared frequently in the media and academic literature, and it is crucial to document its size. It is important to explore if the pandemic effects on learning are selective in educational levels, areas of knowledge, or in different countries and socioeconomic realities (Lake & Olson, 2020; Skye et al., 2020). Data are beginning to emerge regarding learning loss in basic education, but so far, these are scant in higher education (Azevedo et al., 2021; Engzell et al., 2021; Maldonado & De Witte, 2020; Donnelly & Patrinos, 2022).

For universities, the change was unexpected and required emergency measures of heterogeneous quality, including online assessment testing (IESALC-UNESCO, 2020; Lake & Olson, 2020; UNESCO, 2020). The validity and use of test results in this context required complex analyses and decisions, with mixed arguments about their usefulness and risk of biases (Jankowski, 2020; Lake & Olson, 2020).

Objective tests are the preferred instrument to assess students' knowledge in a valid and reliable manner (AERA, 2014; Lane et al., 2015; Sánchez-Mendiola, 2020). There is widespread agreement that multiple-choice question (MCQ) large-scale testing can measure knowledge in a precise, timely, and cost-effective manner with large populations of students (Lane et al., 2015). Online testing has added advantages in terms of ease of application, speed of analysis and scoring, and less use of paper and is becoming a preferred option if the technology and infrastructure are available (Butler-Henderson & Crawford, 2020; Jankowski, 2020). Computer-aided testing has several disadvantages: cost of software licenses, equipment, and connectivity infrastructure, need for digital skills in faculty and students, among others (Butler-Henderson & Crawford, 2020; Dennick et al., 2009). An important challenge during the pandemic has been the difficulty of direct proctoring and supervision during testing, which is more complex when testing occurs in the students' home (technological issues, Internet access, electricity, possibility of cheating) (Daffin, 2018; Reisenwitz, 2020; Şenel and Senel, 2021; UNESCO, 2020).

The National Autonomous University of Mexico (UNAM) has a long tradition of applying large-scale standardized testing for diagnostic and high-stakes exams (Valle, 2012). The university has applied yearly, since 1995, a diagnostic exam to students that are admitted to the institution, to assess knowledge in several areas, plus English and Spanish (Martínez-González et al., 2018, 2020; Valle, 2012).

The Educational Evaluation area of the university is a centralized department that develops these exams and has the task of developing, validating, implementing, applying, and scoring the test results, as well as publishing institutional reports. The exams are MCQ tests developed with an evidence-based process to create academically grounded instruments. The diagnostic exams have been applied in face-to-face modality during the first weeks of the academic year, at the university schools' classroom facilities, with faculty proctoring and a massive logistical display of resources, since the student cohorts are in the tens of thousands (Martínez-González et al., 2018, 2020; Valle, 2012). A goal of the diagnostic exams is to obtain information about students' knowledge level at admission, to identify areas of high and low performance, and implement remediation strategies to decrease dropout and academic delay during the first year of their academic trajectory.

The pandemic forced the university to apply the admission diagnostic exams online, opening a window of opportunity to collect data about the process and contrast the information obtained with previous exam administrations. The goal of this study was to compare and contrast knowledge levels in newly admitted university students, in pre- and transpandemic cohorts, and to analyze the results by knowledge area and gender.

2 Methods

2.1 Research design

The research design was quasi-experimental with static group comparisons, taking advantage of the pandemic "natural experiment" (Campbell and Stanley, 1963; Fraenkel et al., 2018).

2.2 Setting

UNAM is the largest public university in Mexico (>369,000 students, >42,000 faculty, 133 careers, http://www.estadistica.unam.mx) and one of the largest in the world. Each year, about 35,000 students are admitted to the university (Sánchez-Mendiola et al., 2020).

2.3 Population and sample

The studied populations were four cohorts admitted to the University in 2017, 2018 (before the pandemic, control groups) and 2020, 2021 (during the pandemic, intervention groups). Each of the two pairs of cohorts was evaluated with the same instrument (2017 and 2021 used the same test, 2018 and 2020 had the same exam). The diagnostic exam is not mandatory and has no summative implications, although the majority of students decide to take the test. The vast majority of applicants are based in Mexico City and the surrounding metropolitan area. In Mexico, all schools

and universities closed due to the pandemic lockdown in the middle of March 2020 and stayed closed for more than 200 days. Students that responded the diagnostic exam in August 2020 finished high school at the end of Spring 2020, so they spent about three months with emergency remote teaching from home before finishing high school. The August 2021 student cohort did the last year or so of their high school education via distance learning.

2.4 Instrumentation

The diagnostic exam is an MCQ test composed of two portions, with 120 items each: one general knowledge (GK) test (mathematics, physics, chemistry, biology, world history, Mexican history, literature, and geography) and one English and Spanish (ES) exam. The proportion of items in the general knowledge test varies depending on the career area where the students are admitted. At UNAM, careers are classified in four areas: area I, physics, mathematics, and engineering sciences (PMES); area II, biological, chemical, and health sciences (BCHS); area III, social sciences (SS); and area IV, humanities and arts (HA). There are four versions of the general knowledge exam, one for each area. Only area IV has philosophy items (Table 1).

The English and Spanish tests have 60 items for each language. The Spanish portion assesses four areas: reading comprehension (16 items), grammar and composition (23), vocabulary (9), and orthography (12). The English test evaluates the first three levels of language domain, in agreement with the Common European Frame of Reference for Languages (CEFRL, 2001). Levels are beginner (Level A1), high beginner (Level A2), and low intermediate level (Level B1). Each level has 20 items. Applicants need to obtain at least 75% of items correct in the block of items that reach the corresponding level; if they do not reach the first level, they are included in

Table 1 Structure of the general knowledge diagnostic exam, UNAM, Mexico City

Subject matter	Area of knowledge							
	I. Physics, math, and engineering sciences		II. Biological, chemical, and health sciences		III. Social sciences		IV. Humanities and arts	
	No. of items	%	No. of items	%	No. of items	%	No. of items	%
Mathematics	36	30.0	32	26.7	30	25.0	24	20
Physics	24	20.0	16	13.3	10	8.3	10	8.3
Chemistry	10	8.3	16	13.3	12	10	10	8.3
Biology	10	8.3	16	13.3	12	10	10	8.3
World history	10	8.3	10	8.3	16	13.3	10	8.3
Mexican history	10	8.3	10	8.3	16	13.3	10	8.3
Philosophy	–	–	–	–	–	–	20	16.7
Literature	10	8.3	10	8.3	14	11.7	16	13.3
Geography	10	8.3	10	8.3	10	8.3	10	8.3
Total	120	100	120	100	120	100	120	100

a "not classified" category. Traditionally, UNAM first-year students have low scores in English, as is frequent in public universities in Mexico.

Since general knowledge exams have a different structure for each area of knowledge, they were analyzed separately. In the case of the English and Spanish test, they were analyzed for the total population, since they are the same for all areas.

The contents of the diagnostic exam after admission to the university are focused on the fundamental learning outcomes at the end of high school and cover the official high school curriculum in Mexico. The test blueprint and items are validated by teachers with content expertise in each of the test explored areas. The main goal of the diagnostic exam is to assess the level of knowledge in students that are admitted to the university.

These exams are developed following good practices for objective large-scale standardized tests, as has been previously reported (Sánchez-Mendiola et al., 2020). In summary, the test blueprint is developed by content expert groups, led by test development staff from UNAM Department of Educational Evaluation, using the official Mexican high school curriculum learning goals as the construct to be measured. Our university has created a large item-bank over the course of many years, piloting and testing with psychometric analysis. Every year, items are added to the pool to renew and grow the number of available items, and different items are applied on each occasion. In this study, for purposes of pre- and transpandemic comparison, the test applied in 2020 was the same as the exam in 2018, to decrease instrument bias. The 2018 test was done face-to-face paper-and-pencil, and the 2020 exam was done online with remote proctoring. We also included the results of two more years, 2017 (prepandemic, paper-and-pencil) and 2021 (transpandemic, online), to provide a broader perspective of the results; these years were chosen because the same instrument was utilized in these cohorts. Test development processes, application, and scoring are performed with several quality-control steps along the process, following standard practices for large-scale testing.

Tests are piloted prior to application, and test item selection from the item bank has strict criteria (moderate difficulty, high reliability, and high discrimination indices). The psychometric data obtained with classical measurement theory (CMT) and item response theory (IRT) have been previously reported (Martínez-González et al., 2018, 2020; Sánchez-Mendiola et al., 2020).

2.5 Test application

The 2017 and 2018 tests were applied in a paper-and-pencil printed format with Opscan MCQ answer sheets, in their respective schools' classrooms. The central Evaluation Department coordinated the logistics, and each school applied the test with supervision. The test was applied in groups of varying size, depending on the school facilities and student population, with direct proctoring by faculty. The total time for answering both exams was three hours. The answer sheets and exams were collected and analyzed in the Evaluation Department, and the results were collated in a report sent to the university authorities and each school.

Due to the pandemic lockdown, in 2020 and 2021, the test was applied online at home, with specific instructions to minimize cheating (Butler-Henderson, 2020; Dennick et al., 2009). The university developed an in-house digital platform for the design, application, and scoring of online exams in large populations (EXAL), which allows real-time monitoring of the responses from each test-taker, as well as the time used in each item and test-taking total time. It registers monitor and mouse/keyboard activity and flags anomalous events as incidents, issuing alerts to the remote test proctors. It does not use video monitoring in real time, which was not feasible in our setting due to the technological and Internet connectivity limitations in many test-takers' homes.

Instructions were sent to the students for pretest verification of technical compatibility and ethical, security, and technical issues. The testing process was piloted previously, and three hours of testing time were allowed. Test applications were achieved without major problems.

2.6 Psychometric and statistical analysis

Student performance was measured with percentage correct response scores (Morduchowicz, 2006). Average test scores were calculated, as well as differences between 2017–2021 and 2018–2020. Since the study uses observational data, we used propensity score matching (PSM) to build balanced and comparable control (prepandemic, face-to-face testing) and experimental (transpandemic, online testing) groups, to account for the covariates that may predict receiving the intervention (Rosenbaum & Rubin, 1983). For each pair of comparisons (2017–2021, 2018–2020), the conditional probability for each individual to be exposed to the intervention was calculated with a combination of the observed variables of interest, using the software R (r-project.org). The variables of interest utilized to calculate the propensity score were gender and high school of origin, in agreement with our previous studies (Martínez-González et al., 2018). PSM assumes that two students with similar propensity scores have the same distribution of explanatory variables, which implies that samples built with this criterion helps to guarantee comparable groups. The technique used was near neighbor with caliper matching.

For the knowledge exams in test cohorts 2018–2020 and 2017–2021, pairing was performed by area of knowledge; in the Spanish and English exams, it was not necessary to divide by area since the total population completed the same instrument in each application. Results were also analyzed by gender.

Inferential statistics for group differences was done with Student's *t*-test for independent samples. Cohen's *d* with pooled standard deviations was calculated as a measure of effect size between cohorts (Cohen, 1988).

Psychometric analyses were performed with CMT and IRT. Descriptive statistics, alpha's Cronbach coefficient for reliability, standard error of measurement, mean difficulty index, and point biserial correlation coefficient for discrimination were calculated. Analysis was done with ITEMAN 3.5, BILOG MG 3.0 and Winsteps 3.0 software.

2.7 Ethical aspects

The study was in compliance with the Declaration of Helsinki for research involving human subjects' data. Data was managed anonymously in a confidential manner.

3 Results

For the first prepandemic-transpandemic (2018–2020) comparison set, 35,584 matched students from each cohort were considered, using the PSM methodology. For the second comparison set (2017–2021), 31,574 matched students from each cohort were considered. For the Spanish and English exams, the first set of comparisons (2018–2020) used 33,585 matched students and for the second set (2017–2021) 33,481 students.

Psychometric results of all exams were appropriate for a test with diagnostic intentionality. Reliability coefficients in the 2017 and 2018 exams were in the range of 0.72 to 0.94, and for the 2020 and 2021 tests from 0.86 to 0.95, overall reliability was 5 to 10% higher in the transpandemic online exams (Table 2). Mean difficulty indices were moderate in all exams, 2 to 7% higher in the transpandemic cohorts; discrimination indices were appropriate with point biserial correlations of 0.15 to 0.48, with higher indices in the transpandemic cohorts (Table 2).

Figure 1 shows Wright's maps comparing area II (BCHS) exams, mapping the test items' difficulty with the students' ability in the construct evaluated, using the item response theory Rasch model (Andrich & Marais, 2019). These data show that the test difficulty is adequately calibrated for the student population, and the patterns are similar in each pair of cohorts. The results in the other exams showed similar patterns, which adds validity evidence about the appropriateness or fit of the exams' difficulty for the students' range of ability levels.

The mean percent correct scores for the four cohorts by area of knowledge, Spanish and English, are shown in Fig. 2. The planned paired comparisons using the same instrument (2017 vs. 2021 and 2018 vs. 2020) showed increased scores in the four areas of knowledge during the pandemic, ranging from 2.3 to 4.4% in the 2018 vs. 2020 comparison and from 4 to 7.1% in the 2017 vs. 2021 contrast.

3.1 Area of knowledge

Scores in the four areas of knowledge increased in the 2020 and 2021 cohorts; all the differences were statistically significant. The largest increases were observed in area I (PMES) with 4.4% in the 2018–2020 comparison and 7.1% in the 2017–2021 pair (Table 3). The lowest were in area III (SS) with 2.3% in the 2018–2020 comparison and 4% in 2017–2021. Results of the general knowledge exam in the four areas were higher in area IV (HA) in all cohorts, with mean percent correct score values ranging from 50.2 and 56.8.

Table 2 Classical measurement theory parameters for the 2017, 2018, 2020, and 2021 diagnostic assessment examinations in UNAM students (general knowledge exams: 2018 $n=35{,}821$; 2020 $n=41{,}909$; 2017 $n=34{,}003$; 2021 $n=37{,}688$) (English and Spanish exams: 2018 $n=35{,}837$; 2020 $n=35{,}534$; 2017 $n=34{,}078$; 2021 $n=35{,}097$)

	General knowledge exams																							
	Area I (PMES)				Area II (BCHS)				Area III (SS)				Area IV (HA)				Spanish				English			
Parameters	2018	2020	2017	2021	2018	2020	2017	2021	2018	2020	2017	2021	2018	2020	2017	2021	2018	2020	2017	2021	2018	2020	2017	2021
Reliability coefficient (alpha)	0.85	0.90	0.87	0.92	0.82	0.89	0.84	0.90	0.78	0.87	0.84	0.91	0.81	0.89	0.86	0.91	0.76	0.86	0.76	0.86	0.94	0.95	0.94	0.95
Standard error of measurement	4.94	4.91	4.96	4.87	4.95	4.92	4.96	4.94	4.91	4.88	4.98	4.89	4.82	4.83	4.90	4.87	3.51	3.46	3.47	3.40	3.14	3.08	3.12	3.01
Mean difficulty index	0.46	0.51	0.48	0.55	0.45	0.48	0.46	0.51	0.46	0.48	0.49	0.53	0.49	0.52	0.52	0.56	0.55	0.53	0.51	0.53	0.58	0.60	0.61	0.65
Mean point biserial correlation coefficient	0.19	0.25	0.21	0.29	0.17	0.24	0.18	0.25	0.15	0.21	0.18	0.27	0.16	0.23	0.20	0.26	0.20	0.29	0.20	0.28	0.44	0.48	0.45	0.47

Area I (PMES) physics, mathematics, and engineering sciences, *Area II (BCHS)* biological, chemical, and health sciences, *Area III (SS)* social sciences, *Area IV (HA)* humanities and arts

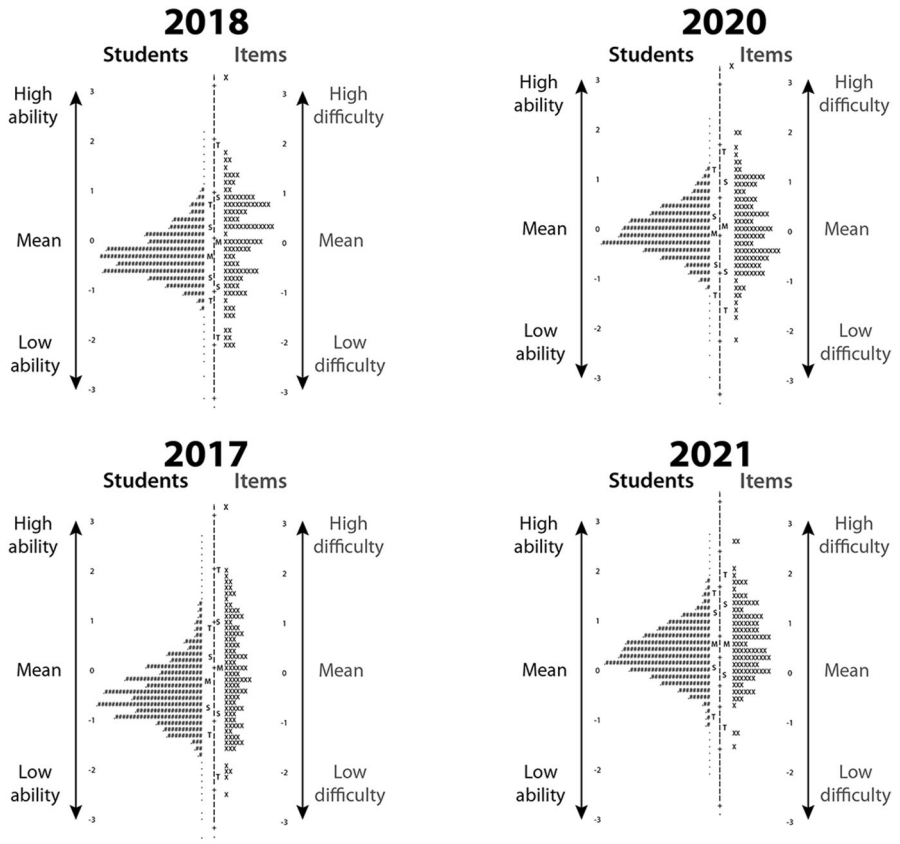

Fig. 1 Wright's maps using Rasch model of test items' difficulty levels and students' ability in area II (biology, chemistry, and health sciences) general knowledge diagnostic exams at UNAM, Mexico (2017, 2018, 2020, and 2021 cohorts)

The scores in the Spanish exam had a small decrease of 1.3% in the 2020 pandemic exam compared to the 2018 prepandemic test, although in contrast there was a 1.6% increase in the 2017 vs. 2021 comparison (Fig. 2 and Table 3). The English test had a 1.7% increase when comparing pandemic vs. prepandemic scores (2020 vs. 2018), but when comparing the 2021 vs. 2017 exam, the English scores were significantly lower during the pandemic, about 7.7% less (Fig. 2 and Table 3).

Figure 3 shows the effect sizes using Cohen's d for all paired comparisons (2017 vs. 2021, 2018 vs. 2020) for total scores in the four areas of knowledge, Spanish and English, and by gender. Effect sizes were generally larger in the 2017–2021 comparisons; they were also larger in area I (PMES) and in women (Fig. 3). Cohen suggested that $d = 0.2$ could be considered a "small" effect size, 0.5 "medium," and 0.8 a "large" effect size (Cohen, 1988).

The analysis by subject topic showed that in most categories there were higher scores in the pandemic cohorts (2020 and 2021) compared to prepandemic

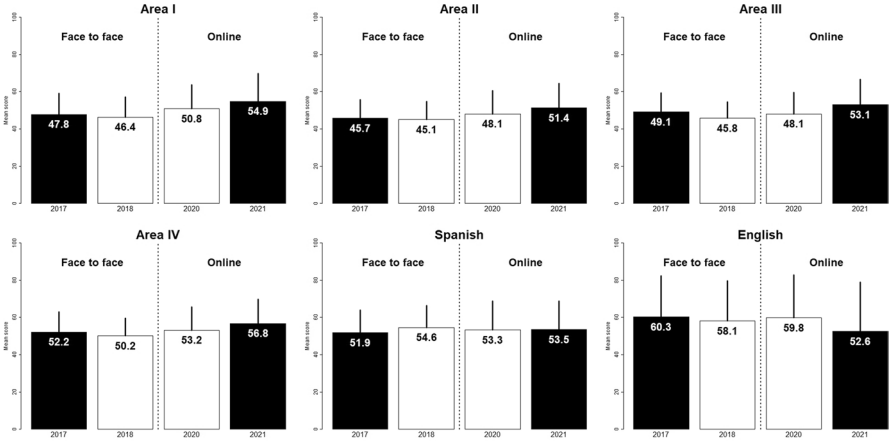

Fig. 2 Mean percent correct scores with standard deviations (SD) for the 2017, 2018, 2020, and 2021 diagnostic assessment exams at UNAM, Mexico. Results by cohort, area of knowledge (area I = physics, mathematics, and engineering sciences; area II = biological, chemical, and health sciences; area III = social sciences; area IV = humanities and arts), and testing modality (face to face, online). Paired comparisons' data are shown in black bars for 2017 vs. 2021 and white bars for 2018 vs. 2020. All differences were statistically significant

Table 3 Number of students, mean percent correct scores, standard deviation, difference of means, effect size (Cohen's *d*), and 95% confidence interval for the diagnostic assessment exams at UNAM in 2018–2020 and 2017–2021, total results and by gender

			2018			2020				Diff		2017			2021				Diff	
			N	Mean	SD	N	Mean	SD	*d*	M	CI 95%	N	Mean	SD	N	Mean	SD	*d*	M	CI 95%
Total		Area I	7,660	46.4	10.8	7,660	50.8	13.0	0.39	4.4	(4,4.8)	6,812	47.8	11.4	6,812	54.9	14.8	0.57	7.1	(6.7,7.6)
		Area II	11,139	45.1	9.8	11,139	48.1	12.5	0.27	3.1	(2.8,3.4)	10,529	45.7	10.1	10,529	51.4	13.1	0.45	5.7	(5.4,6.0)
		Area III	12,758	45.8	8.8	12,758	48.1	11.5	0.21	2.3	(2.06,2.6)	11,546	49.1	10.2	11,546	53.1	13.5	0.32	4.0	(3.7,4.3)
		Area IV	4,027	50.2	9.4	4,027	53.2	12.5	0.26	3.0	(2.5,3.5)	2,687	52.2	10.9	2,687	56.8	13.0	0.37	4.6	(4.0,5.2)
		Spanish	33,585	54.6	11.9	33,585	53.3	15.5	-0.09	-1.3	(-1.5,-1.1)	33,481	51.9	12.1	33,481	53.5	15.4	0.12	1.6	(1.4,1.8)
		English	33,585	58.1	21.6	33,585	59.8	23.1	0.07	1.7	(1.4,2.03)	33,481	60.3	22.1	33,481	52.6	26.4	-0.31	-7.7	(-8.05,-7.3)
Gender	Male	Area I	4,969	47.3	11.0	4,969	50.8	13.5	0.31	3.5	(3.0,3.9)	4,506	48.6	11.7	4,506	55.0	15.5	0.50	6.4	(5.8,6.9)
		Area II	3,748	47.4	10.6	3,748	48.9	13.5	0.14	1.5	(1.0,2.1)	3,565	48.7	10.8	3,565	51.9	13.8	0.25	3.2	(2.6,3.8)
		Area III	6,077	47.2	9.3	6,077	48.5	11.8	0.11	1.3	(0.9,1.7)	5,570	50.5	10.8	5,570	52.9	14.1	0.19	2.4	(1.9,2.8)
		Area IV	1,473	52.1	10.1	1,473	53.1	13.4	0.09	1.0	(0.02,1.9)	838	54.8	11.7	838	57.0	13.7	0.18	2.2	(1.0,3.4)
		Spanish	15,363	53.2	12.4	15,363	50.8	16.1	-0.17	-2.4	(-2.7,2.0)	15,182	51.0	12.7	14,959	51.0	16.0	0.01	0.1	(-0.2,0.4)
		English	15,363	58.4	22.1	15,363	58.7	23.6	0.01	0.4	(0.1,0.9)	15,182	60.5	22.6	14,959	52.1	26.6	-0.34	-8.4	(-9.0,-7.9)
	Female	Area I	2,691	44.7	10.3	2,691	50.8	12.1	0.54	6.1	(5.5,6,7)	2,306	46.1	10.5	2,306	54.7	13.6	0.68	8.6	(7.9,9.3)
		Area II	7,391	43.9	9.2	7,391	47.7	11.9	0.34	3.9	(3.5,4.2)	6,964	44.1	9.3	6,964	51.2	12.7	0.56	7.0	(6.7,7.4)
		Area III	6,681	44.6	8.2	6,681	47.8	11.2	0.29	3.2	(2.9,3.6)	5,976	47.8	9.5	5,976	53.3	12.9	0.43	5.5	(5.1,5.9)
		Area IV	2,554	49.1	8.9	2,554	53.2	11.9	0.35	4.1	(3.5,4.7)	1,849	51.0	10.3	1,849	56.7	12.6	0.45	5.7	(5.0,6.4)
		Spanish	18,222	55.8	11.3	18,222	55.4	14.7	-0.03	-0.5	(-0.7,-0.2)	18,299	52.7	11.6	18,522	55.5	14.6	0.20	2.8	(2.6,3.1)
		English	18,222	57.9	21.2	18,222	60.7	22.7	0.12	2.8	(2.4,3.3)	18,299	60.2	21.7	18,522	53.0	26.2	-0.28	-7.1	(-7.6,-6.6)

Area I (PMES) = physics, mathematics, and engineering sciences; Area II (BCHS) = biological, chemical and health sciences; Area III (SS) = social sciences; Area IV (HA) = humanities and arts. *d*=Effect size (Cohen's *d*); Diff M= Mean difference, CI 95% = 95% Confidence interval

≤0 0<x≤3 3<x≤6 6<x≤10 x>10

Area I (PMES) physics, mathematics, and engineering sciences, *Area II (BCHS)* biological, chemical, and health sciences, *Area III (SS)* social sciences, *Area IV (HA)* humanities and arts, *d* effect size (Cohen's *d*), *Diff M* mean difference, *CI 95%* 95% confidence interval

$\leq 0\ 0 < x \leq 3\ 3 < x \leq 6\ 6 < x \leq 10\ 10$

values (2018 and 2017), in some cases as high as 15.7% higher (Mexico history) (Table 4). Geography was the only subject that had negative differences in some areas of knowledge during the pandemic, and literature scores in men also had some negative differences.

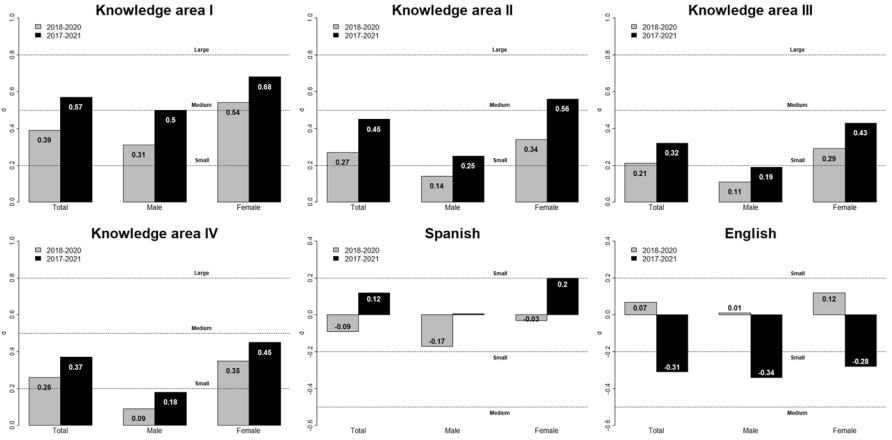

Fig. 3 Effect sizes measured with Cohen's *d* for the 2017, 2018, 2020, and 2021 diagnostic assessment exams at UNAM, Mexico. Grey bars represent effect sizes for the 2018 vs. 2020 comparison and black bars the 2017 vs. 2021 comparison. (area I = physics, mathematics, and engineering sciences; area II = biological, chemical, and health sciences; area III = social sciences; area IV = humanities and arts)

Table 4 Mean scores by topic in UNAM diagnostic exam and differences between 2018–2020 and 2017–2020, by area of knowledge and gender (general knowledge exams: 2018–2020 $n = 35{,}584$ and 2017–2021 $n = 31{,}574$)

		Total						Male						Female					
		2018	2020	Diff M	2017	2021	Diff M	2018	2020	Diff M	2017	2021	Diff M	2018	2020	Diff M	2017	2021	Diff M
Mathematics	Knowledge area I	46.0	50.5	4.4	47.5	54.6	7.1	46.9	50.6	3.7	48.3	54.6	6.3	44.3	50.2	5.9	45.9	54.6	8.7
	Knowledge area II	37.7	41.0	3.3	37.2	45.5	8.3	39.6	42.3	2.6	39.0	44.8	5.8	36.7	40.4	3.6	36.2	45.8	9.6
	Knowledge area III	38.0	40.0	2.0	39.9	43.8	3.9	39.3	40.6	1.3	41.2	44.0	2.8	36.9	39.5	2.6	38.8	43.7	4.9
	Knowledge area IV	37.7	41.3	3.6	40.6	46.0	5.4	38.5	40.9	2.3	42.3	46.9	4.6	37.2	41.5	4.3	39.9	45.6	5.7
Physics	Knowledge area I	39.5	41.6	2.1	40.6	45.4	4.8	41.2	42.3	1.1	42.4	46.5	4.1	36.4	40.4	4.0	37.0	43.2	6.2
	Knowledge area II	33.5	34.1	0.6	36.0	42.1	6.1	36.9	36.5	-0.5	40.2	43.5	3.4	31.8	32.9	1.1	33.9	41.4	7.5
	Knowledge area III	38.8	46.5	7.7	40.1	46.4	6.4	40.8	46.6	5.8	41.9	46.2	4.3	37.0	46.3	9.4	38.4	46.7	8.3
	Knowledge area IV	45.3	48.3	3.0	37.2	44.6	7.4	47.5	48.7	1.2	39.7	44.6	4.9	44.0	48.1	4.1	36.1	44.7	8.6
Chemistry	Knowledge area I	54.2	58.4	4.2	44.5	52.1	7.6	54.9	57.3	2.4	45.3	51.7	6.5	52.9	60.5	7.7	43.1	52.9	9.8
	Knowledge area II	49.2	54.8	5.6	48.9	55.0	6.1	51.0	53.9	2.8	52.0	55.1	3.1	48.3	55.4	7.1	47.3	55.0	7.7
	Knowledge area III	42.0	45.0	3.0	40.8	46.9	6.1	42.9	44.3	1.4	41.4	45.5	4.1	41.1	45.6	4.5	40.2	48.2	8.0
	Knowledge area IV	41.0	47.0	6.0	48.4	58.5	10.1	41.7	45.0	3.2	50.7	57.4	6.7	40.5	48.2	7.6	47.4	59.0	11.6
Biology	Knowledge area I	41.9	48.8	6.9	52.3	59.5	7.2	42.4	47.6	5.2	52.1	58.0	5.9	40.8	50.9	10.1	52.7	62.3	9.5
	Knowledge area II	48.3	50.1	1.8	54.1	56.1	2.0	51.2	51.1	-0.1	58.6	57.2	-1.4	46.8	49.6	2.8	51.8	55.5	3.7
	Knowledge area III	37.5	41.7	4.2	53.7	57.8	4.1	37.9	40.8	2.9	53.8	56.1	2.3	37.2	42.5	5.4	53.6	59.4	5.7
	Knowledge area IV	54.8	62.0	7.2	57.9	65.1	7.2	57.0	61.3	4.3	57.6	62.0	4.4	53.6	62.4	8.8	58.1	66.6	8.5
World history	Knowledge area I	55.2	58.9	3.7	57.8	63.7	5.9	56.9	60.1	3.2	59.4	64.7	5.3	52.0	56.7	4.7	54.7	61.7	7.0
	Knowledge area II	51.7	54.7	2.9	44.9	51.7	6.9	54.8	55.8	1.1	49.7	54.1	4.4	50.2	54.1	3.9	42.4	50.5	8.1
	Knowledge area III	50.9	52.6	1.7	54.6	59.8	5.2	53.6	54.4	0.8	58.0	60.8	2.7	48.4	51.0	2.6	51.4	59.0	7.5
	Knowledge area IV	60.0	62.6	2.6	51.8	58.1	6.3	63.9	65.2	1.4	58.0	60.3	2.3	57.8	61.1	3.3	49.0	57.2	8.1
Mexico history	Knowledge area I	40.1	48.8	8.7	40.2	55.8	15.7	40.8	48.8	8.0	41.1	56.1	15.0	38.8	48.7	9.9	38.4	55.4	17.0
	Knowledge area II	41.1	48.5	7.4	49.1	56.4	7.2	43.6	48.7	5.1	51.5	57.4	6.0	39.8	48.3	8.5	47.9	55.8	7.9
	Knowledge area III	47.2	48.6	1.3	51.1	57.1	5.9	48.9	49.5	0.5	54.0	58.4	4.4	45.7	47.7	2.0	48.5	55.8	7.4
	Knowledge area IV	51.0	56.3	5.3	53.5	58.4	4.9	54.8	57.9	3.0	59.3	60.9	1.6	48.8	55.4	6.7	50.9	57.3	6.4
Literature	Knowledge area I	52.9	55.1	2.2	55.4	60.8	5.5	52.0	53.8	1.8	54.3	59.5	5.2	54.5	57.5	3.0	57.5	63.4	5.9
	Knowledge area II	65.2	68.8	3.6	61.6	63.0	1.5	64.2	65.9	1.7	62.2	60.9	-1.3	65.7	70.3	4.6	61.2	64.1	2.9
	Knowledge area III	60.7	61.1	0.3	64.1	64.7	0.6	60.4	59.9	-0.6	63.1	62.2	-0.9	61.0	62.1	1.2	65.0	67.0	2.0
	Knowledge area IV	59.7	60.2	0.5	66.9	67.2	0.3	60.3	58.6	-1.7	68.3	65.7	-2.6	59.4	61.1	1.7	66.3	67.9	1.6
Geography	Knowledge area I	52.3	57.2	4.9	54.9	61.6	6.7	53.2	57.5	4.3	55.8	62.0	6.2	50.6	56.7	6.1	53.3	60.9	7.6
	Knowledge area II	53.3	51.2	-2.1	51.1	55.0	3.8	57.4	54.3	-3.1	54.7	57.3	2.5	51.2	49.6	-1.6	49.3	53.8	4.5
	Knowledge area III	59.0	59.1	0.1	57.0	55.9	-1.1	61.4	60.8	-0.6	58.9	57.2	-1.7	56.9	57.5	0.6	55.2	54.8	-0.4
	Knowledge area IV	61.2	61.0	-0.2	59.5	62.5	3.1	64.7	62.7	-2.0	61.6	63.3	1.7	59.1	60.0	0.9	58.5	62.1	3.7
Philosophy	Knowledge area IV	50.8	52.0	1.2	56.7	58.1	1.5	52.7	51.8	-0.9	59.9	58.8	-1.1	49.7	52.1	2.4	55.2	57.9	2.6

≤0 | 0<x≤3 | 3<x≤6 | 6<x≤10 | x>10

≤ 0 $0 < x \leq 3$ $3 < x \leq 6$ $6 < x \leq 10$ $x > 10$

3.2 Gender

More than 50% of UNAM student population are women (UNAM, 2022). The percentage of women in the four cohorts were as follows: 2017 = 53.8%, 2018 = 54%, 2020 = 53%, and 2021 = 54%. The proportion of female students varies considerably by area of knowledge: in area I (PMES), the majority was male (2018 = 65%, 2020 = 68%), in areas II (BCHS) and IV (HA), the majority were women (area II, 2018 = 66% and 2020 = 68%; area IV, 2018 = 63% and 2020 = 65%). In area III (SS), there is a slightly higher percentage of women (2018 = 53% and 2020 = 51%). In the four exam cohorts, the pattern of gender distribution was similar.

The mean percentage correct scores in general knowledge exams in men were higher in the four areas in all cohorts. However, before the pandemic, there was a 3.1% higher global test performance in men compared to women, and this gap decreased to 0.34% during the pandemic. Women had higher gains than men in both pandemic vs. prepandemic paired comparisons, and these increases were larger in the 2017 vs. 2021 comparison (Tables 3 and 4).

Men had a lower score in Spanish in 2020 compared to 2018 (−2.4%), although the 2017–2021 comparison showed no difference (0.1%) (Table 3). In the English exam, there was a slight increase of 0.4% in 2020 compared to 2018, but there was a large decrease of −8.4% from 2017 to 2021.

Test performance in women showed a better scenario. The general knowledge exam scores were higher in 2020 and 2021 compared to 2018 and 2017. The largest difference was found in area I (PMES) exams with a 6.1% increase in 2020 and an 8.6% increase in 2021; the other areas had increases in scores from 3.2% to 7.0%. All these differences were statistically significant (Table 3).

The performance of women in the Spanish exam had a small decrease of 0.5% in 2020 compared to 2018 and a 2.8% increase when comparing 2017 to 2021 (Table 3). The English exam in women showed an increase of 2.8% from 2018 to 2020 but a large decrease from 2017 to 2021 of −7.1%.

4 Discussion

Diagnostic assessment of knowledge in students admitted to the university is important to obtain their baseline academic status at the beginning of the first year, so institutions can identify students that are in a disadvantaged situation and design interventions to improve their likelihood of success. A well-implemented diagnostic strategy at admission can help decrease dropout and academic delay in the first year of higher education (Bombelli & Barberis, 2012; Martínez-González et al., 2018, 2020; Porta, 2018).

UNAM's admission diagnostic exam is designed following good practices for objective, standardized, large-scale tests (AERA, 2014; Lane et al., 2015) and has validity evidence (Martínez-González et al., 2018, 2020). This study showed an overall increase in scores during the pandemic period in the majority of knowledge domains, which was larger in the second year of the pandemic (2021). There was a small decrease in Spanish scores in the first year of the pandemic, which changed to a small increase in the second year. There was a small increase in English scores in

2020 and a substantial decrease in the 2017–2021 comparison, where students had more than one year of their education in confinement. These data do not support the hypothesis that during the pandemic there would be a large learning loss at all educational levels (Pier et al., 2021; UNESCO, 2021).

There are editorials and opinion articles that predict a large learning loss in the trans- and post-pandemic eras, but so far, concrete evidence backed by data is scarce (Azevedo et al., 2021; Pokhrel and Chhetri, 2021). A recent systematic review has identified only eight published papers related to the subject of pandemic learning loss (Donnelly & Patrinos, 2022). Seven studies found learning loss evidence of 0.03 to 0.29 standard deviations (SD) in at least some of the participants, and one found learning gains. Six studies in this review were in primary level students and only two from higher education (Orlov et al., 2021; Gonzalez et al., 2020). The educational level is critically important during the pandemic, since K-12 level students are in a very different level of maturity and cognitive development compared to university students. The majority of studies that analyzed primary level students showed a decrease in learning that was statistically and educationally significant (Azevedo et al., 2021; Engzell et al., 2021; Maldonado & De Witte, 2020; Pier et al., 2021; Schult et al., 2021). Our findings show an increase in almost all areas of knowledge, except Spanish and English, in a large sample of higher education students. These data are compatible with the findings of Gonzalez et al. (2020) at the Universidad Autónoma de Madrid, who analyzed the effects of pandemic confinement on autonomous learning performance of 458 higher education students in courses related to "Applied Computing," "Metabolism," and "Design of Water Treatment Facilities." The experimental group (during the pandemic, 2020) had a better performance than the control group (prepandemic, 2017–2019); furthermore, the experimental group had more continuous learning activities and better assessment outcomes. The authors suggest that confinement had a positive effect in students' learning strategies (González et al., 2020). These findings agree with our study, where we found an increased academic level in knowledge levels, suggesting that higher education students can overcome the difficult pandemic situation and compensate through several strategies the potential negative academic effects of the pandemic (ten Cate, 2001).

The only other study that analyzed higher education students in Donnelly and Patrinos' systematic review was from the United States, examining seven economics courses, where they found a worse performance in Spring 2020 compared to Spring or Fall 2019 students (Orlov et al., 2021). This paper found a statistically significant drop of 0.185 SD ($p=0.015$) during the pandemic semester. The authors do not report the sample size of the study, and unlike our study, it was only in the field of economics.

Due to the large sample sizes in our study, almost all comparisons are statistically significant, although the question remains about its educational significance, as has been previously discussed in the educational research literature (McLean & Ernest, 1998). We used Cohen's *d* as a measure of effect size to provide a clearer picture of the differences among the prepandemic and pandemic cohorts and found that the differences were in the small and moderate range. The use of propensity score matching helped to have statistically matched and balanced control and intervention groups, to decrease potential biases introduced by confounding variables. To argue causality in this type of studies is difficult without an experimental design, but as far as we know, this is one of the few studies that addresses learning loss in higher education with objective tests.

4.1 Performance by modality of test application

An effect that was observed when comparing our test experience was an increase in the number of students that took the test online vs. face to face. It increased 3.4% in the 2018–2020 comparison (78.8 to 82.2% of the total student population) and 2.6% in the 2017–2021 cohorts (75.8 to 78.4%). The diagnostic exam is not mandatory, so we do not have a clear explanation for these differences, although probably students during the pandemic perceived this diagnostic, non-high-stakes exam as important for themselves and the institution and relevant to their academic history and university success (Bombelli and Barberis, 2012; Martínez-González, 2018; Porta, 2018). Online test application in a home environment was found to be a feasible although logistically complex modality. Before the pandemic, the university used a large amount of financial and human resources for test piloting, printing, reviewing printed exams, quality control, test distribution, and collection. This economic and logistical cost decreased substantially with the online modality. The advantages and disadvantages of online testing for high-stakes tests are controversial, although it could be argued that for formative and diagnostic assessments, where the stakes are not high, the pros of applying exams at home outweigh the cons (Baleni, 2015; Butler-Henderson and Crawford, 2020; Dlab et al., 2015).

Psychometric analysis of the tests in our study, obtained via the same instruments with proven validity and reliability, helped confirm that reliability was appropriate and even increased in the online testing modality. The psychometric data about standard error of measurement, difficulty and discrimination indices, showed a pattern consistent with good practices and international standards for objective large-scale testing, and it is interesting that these psychometric data were better in the online modality than the face to face. The patterns of performance in both tests were not affected by the testing modality. A potential explanation for the improvement in psychometric data when the instrument is applied online is that despite the possible technological asymmetries in equipment sophistication and connectivity at home, answering the test online may help standardize the conditions of testing and decrease students' anxiety. Furthermore, students that lived with the pandemic for a long period of time had likely developed online testing skills and planned conditions at home to interact better with digital devices and online testing platforms.

Another aspect to consider are the likely sources of error in controlled face-to-face exams generated by complex logistics, as well as the different physical facilities where exams were applied in the diverse schools of the university, factors which can influence test scores (AERA, 2014). The online application of the exams was performed by the central Evaluation Department, with the same platform, staff, and instructions to discourage cheating, plus solving technical issues in real time. These factors may have provided a more homogeneous setting for the online test application, and the results appear to be valid as measures of the constructs of knowledge, Spanish, and English in the respective student cohorts.

The mean difficulty indices of the 2020 and 2021 exams were higher than the prepandemic values—with the exception of Spanish—which means that the online versions of the tests were easier for the pandemic student cohorts. Discrimination index (point biserial correlation, an indicator of how well the test discriminates the

more knowledgeable individuals from the less knowledgeable) was higher in all the domains of the 2020 and 2021 tests, which means that the students that correctly answered each item had a better overall performance in the test in comparison with students that had a lower performance. These data, together with the higher reliability, are signals of better instrument performance when applied online and provide results similar to those of Dlab and colleagues (2015), in the sense that mean performance can be higher when tests are applied online. A relevant issue is that non-proctored online exams can have better results, due to the fact that students use other sources of information to improve their scores and thus achieve better outcomes (Backes and Cowan, 2019; Brallier, 2015; Daffin and Jones, 2018). A recent systematic review that studied the effects of the pandemic on more than a million students from the health professions found that learners performed better in online assessments compared to prior in-person evaluations, with differences that ranged from 1 to 3% approximately (Dedeilia et al., 2023; Jaap et al., 2021). We cannot rule out completely that our students did not cheat or use other resources to obtain better scores, although we did not find patterns of cheating in the analysis and, ultimately, if a student searches for the answer in a book or in the digital library that is something that helps him/her learn. Online non-proctored exams at home can be less stressful to students if they are not permanently seen and videorecorded by long-distance proctors. If the test is diagnostic or formative, the incentives for cheating are low and they can improve learning through "assessment for learning and as learning" (Earl et al., 2006). On the other hand, the pattern of responses in all topics of the exam was very similar in all cohorts, suggesting strongly that students did not cheat or artificially increased their scores. This is also supported by the finding that scores in Spanish and English in some comparisons not only did not increase but actually decreased; if students had cheated in the home online test, the scores likely would have been increased in all topics.

Differently from the general knowledge exam, the Spanish test had a slightly higher difficulty for the population in the online modality in 2020. A possible explanation could be that skills that have not been acquired throughout time in reading and grammar will not be reflected in a test of this type irrespective of the application modality. A study performed by Backhoff et al. (2011) highlights the importance that knowledge and skills in Spanish provide students from basic to higher education. On the other hand, the large decrease in English scores in 2021 suggests that the long confinement period during the pandemic lockdown affected learning of English as a second language, as has been reported by several authors (Muftah, 2022; Ying et al., 2021).

4.2 Performance by gender

The analysis by gender showed that men had higher scores in knowledge exams, independently of the area of knowledge, and women had better performance in the Spanish exam and the literature component of the general knowledge test. The English test had more homogeneous results in both sexes. These patterns have been consistent for more than 10 years at our university, although it should be

pointed out that in the pandemic cohort the gap between men and women was substantially decreased in the online test modality during the pandemic years. This was apparent in all domains of knowledge.

Performance in exams by gender has been widely discussed in the literature, since gender roles influence several aspects of cognitive abilities and academic performance (Halpern, 2012). There are unanswered questions about why women and girls have better academic performance and terminal efficiency in education, but consistently have lower scores in MCQ standardized tests. In our institution, several studies of academic trajectories have shown that women have better grades than men throughout the academic career, lower percentage of dropout, and higher percentage of graduation, but lower scores in the high-stakes admission exam to the university and the diagnostic exam applied after admission (Martínez González, et al., 2018, 2020; Campillo et al., 2017). Regarding the Spanish test, the results are consistent with previous findings that women have better performance than men in this area. Other studies have shown that women usually have better performance in exams that assess verbal skills (Brizzio et al., 2008).

4.3 Limitations

The study is a result of the pandemic "natural experiment," so there are other variables that could influence the results, not the least of which is the confounding of the testing modality with the onset of the pandemic, so it is difficult to state with certainty how much of the increase in scores is due to the online modality per se, with its attendant implications (mainly the relatively unexplored aspects of online testing at home), or to a real increase in learning in the 2020 and 2021 cohorts due to intrinsic differences, like more hours dedicated to study because of home confinement. The results of the pandemic cohorts are limited to two generations of students, so it is too early to say that there will not be a learning loss phenomenon in the following years, after the complete effects of the ongoing pandemic settle in. It is necessary to continue monitoring the knowledge levels of students to have a clearer and longitudinal picture of the effects of the pandemic in learning.

5 Conclusions

The performance of two transpandemic cohorts of higher-education students in a large-scale objective standardized diagnostic exam was higher than in two matched prepandemic cohorts. These differences could be explained by a number of factors, including testing modality, although it did not show evidence of a large learning loss in the pandemic groups.

Men had a higher score in the general knowledge exam and women in the Spanish exam, consistent with previous data, although the performance gap between men and women decreased substantially during the pandemic.

Online test application under pandemic conditions at home showed better psychometric data and reliability than the face-to-face modality. Large-scale online testing at home seems to be a valid and cost-effective way of applying diagnostic and formative tests in higher education.

Acknowledgements We acknowledge the participation of the staff at the Directorate of Educational Assessment, Assessment of Data Analysis and Informatics area, UNAM, and the university faculty that participated in the logistics of the test, for their enthusiastic participation in the process during both test applications.

Author contribution MSM and AMG participated in the conceptualization of the study and its design. AMP, MGM, CHP, KGR participated in data curation, validation, and statistical analysis. All authors participated in draft and final manuscript writing and approved the final version.

Data availability The datasets used in the study are available from the corresponding author on request. The complete reports of the diagnostic exam are available in our institution website: https://cuaieed.unam.mx/evaluacion_educativa.

Declarations

Ethics approval The study was in compliance with the Declaration of Helsinki for research involving human subjects' data. Data was managed anonymously in a confidential manner.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Clinical trial registration Not applicable.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer Nature Singapore Pte Ltd.

Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K., & Iqbal, S. A. (2021). Simulating the potential impacts of COVID-19 school closures on schooling and learning outcomes: A set of global estimates. *The World Bank Research Observer*, lkab003. https://doi.org/10.1093/wbro/lkab003

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68*, 89–103. https://doi.org/10.1016/j.econedurev.2018.12.007

Backhoff, E., Larrazolo, N., & Tirado, F. (2011). Habilidades verbales y conocimientos del español de estudiantes egresados del bachillerato en México. *Revista de la Educación Superior, 40*(160), 9–28. https://www.scielo.org.mx/pdf/resu/v40n160/v40n160a1.pdf

Baleni, Z. (2015). Online formative assessment in higher education: Its pros and cons. *The Electronic Journal of e-Learning, 13*(4), 228–236.

Bombelli, E. C., & Barberis, J. G. (2012). Importancia de la Evaluación Diagnóstica en Asignaturas de Nivel Superior con Conocimiento Preuniversitario. *Revista Electrónica Gestión de las Personas y Tecnología, 5*(13), 34–41. https://www.revistas.usach.cl/ojs/index.php/revistagpt/article/view/588

Brallier, A., & Sara, A. (2015). Online testing: comparison of online and classroom exams in an upper-level psychology course. *American Journal of Educational Research, 3*(2), 255–258. https://doi.org/10.12691/education-3-2-20

Brizzio, A., Carreras, M. A., & Fernández, M. (2008). La evaluación de las habilidades de razonamiento verbal y abstracto en estudiantes universitarios. In *Su relación con el rendimiento académico. XV Jornadas de Investigación y Cuarto Encuentro de Investigadores en Psicología del Mercosur*. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires. https://www.aacademica.org/000-032/666

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers & Education, 159*, 104024–104024. https://doi.org/10.1016/j.compedu.2020.104024

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Campillo, M., García-Minjares, M., Martínez-González, A., & Sánchez-Mendiola, M. (2017). Ser hombre, factor para no terminar los estudios de licenciatura: la experiencia mexicana en los últimos 20 años. In *Séptima Conferencia Latinoamericana sobre el Abandono Escolar (VII CLABES)*. https://revistas.utp.ac.pa/index.php/clabes/article/view/1685/2421

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates

Council of Europe (CEFR). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. New York: Cambridge University Press. https://rm.coe.int/1680459f97

Daffin, L. W., Jr., & Jones, A. A. (2018). Comparing student performance on proctored and non-proctored exams in online psychology courses. *Online Learning, 22*, 131–145. https://doi.org/10.24059/olj.v22i1.1079

Dedeilia, A., Papapanou, M., Papadopoulos, A. N., et al. (2023). Health worker education during the COVID-19 pandemic: Global disruption, responses and lessons for the future—a systematic review and meta-analysis. *Human Resources for Health, 21*, 13. https://doi.org/10.1186/s12960-023-00799-4

Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eAssessment: AMEE guide no 39. *Medical Teacher, 31*(3), 192–206. https://doi.org/10.1080/01421590902792406

Dlab, M. H., Katic, M. A., & Čandrlić, S. (2015). Ensuring formative assessment in e-course with online tests. In *2015 10th International Conference on Computer Science & Education (ICCSE)* (pp. 322–327). https://ieeexplore.ieee.org/document/7250264

Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *Prospects, 51*, 601–609. https://doi.org/10.1007/s11125-021-09582-6

Earl, L., Katz, S., & The Western and Northern Canadian Protocol for Collaboration in Education (WNCP) assessment team. (2006). *Rethinking classroom assessment with purpose in mind: Assessment for learning, assessment as learning, assessment of learning*. Manitoba Education, Citizenship, and Youth, School Programs Division. https://www.edu.gov.mb.ca/k12/assess/wncp/full_doc.pdf

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences of the United States of America, 118*(17), e2022376118. https://doi.org/10.1073/pnas.2022376118

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2018). *How to design and evaluate research in education* (10th ed.). Mc Graw-Hill.

Gonzalez, T., de la Rubia, M. A., Hincz, K. P., Comas-Lopez, M., Subirats, L., et al. (2020). Influence of COVID-19 confinement on students' performance in higher education. *PLOS ONE, 15*(10), e0239490. https://doi.org/10.1371/journal.pone.0239490

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press. https://doi.org/10.4324/9780203816530

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

IESALC-UNESCO. (2020). COVID-19 y educación superior: De los efectos inmediatos al día después. In *Análisis de impactos, respuestas políticas y recomendaciones*. https://www.iesalc.unesco.org/wp-content/uploads/2020/05/COVID-19-ES-130520.pdf

Jaap, A., Dewar, A., Duncan, C., et al. (2021). Effect of remote online exam delivery on student experience and performance in applied knowledge tests. *BMC Medical Education, 21*, 86. https://doi.org/10.1186/s12909-021-02521-1

Jankowski, N. A. (2020). *Assessment during a crisis: Responding to a global pandemic*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/08/2020-COVID-Survey.pdf

Lake, R., & Olson, L. (2020). Learning as we go: Principles for effective assessment during the COVID-19 pandemic. *Center on Reinventing Public Education*. https://www.crpe.org/sites/default/files/final_diagnostics_brief_2020.pdf

Lane, S., Raymond, M. R., & Haladyna, T. M. (2015). *Handbook of test development (Second)*. Routledge Taylor & Francis Group. https://doi.org/10.4324/9780203102961

Maldonado, J., & De Witte, K. (2020). The effect of school closures on standardised student test. *KU Leuven – Faculty of Economics and Business.* https://doi.org/10.1002/berj.3754

Martínez-González, A., Sánchez-Mendiola, M., Manzano-Patiño, A., García-Minjares, M., Herrera-Penilla, C., & Buzo-Casanova, E. (2018). Grado de conocimientos de los estudiantes al ingreso a la licenciatura y su asociación con el desempeño escolar y la eficiencia terminal. Modelo multivariado. *Revista De La Educación Superior, 47*(188), 57–85. https://doi.org/10.36857/resu.2018.188.508

Martínez-González, A., Manzano-Patiño, A., García-Minjares, M., Herrera-Penilla, C., Buzo-Casanova, E., & Sánchez-Mendiola, M. (2020). Perfil del estudiante con éxito académico en las licenciaturas del área de las Ciencias Biológicas, Químicas y de la Salud. *Revista De La Educación Superior, 49*(193), 129–152. Recuperado a partir de. http://resu.anuies.mx/ojs/index.php/resu/article/view/1029

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15–22.

Morduchowicz, A. (2006). *Los indicadores educativos y las dimensiones que los integran*. Buenos Aires: IIPEUNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000371341

Muftah, M. (2022). Impact of social media on learning English language during the COVID-19 pandemic. *PSU Research Review, Ahead-of-Print.* https://doi.org/10.1108/PRR-10-2021-0060

Orlov, G., McKee, D., Berry, J., Boyle, A., DiCiccio, T., Ransom, T., Rees-Jones, A., & Stoye, J. (2021). Learning during the COVID-19 pandemic: It is not who you teach, but how you teach. *Economics Letters, 202*, 109812. https://doi.org/10.1016/j.econlet.2021.109812

Pier, L., Hough, H. J., Christian, M., Bookman, N., Wilkenfeld, B., & Miller, R. (2021). *COVID-19 and the Educational Equity Crisis: Evidence on Learning Loss from the CORE Data Collaborative*. Policy Analysis for California Education. https://edpolicyinca.org/newsroom/covid-19-and-educational-equity-crisis

Pokhrel, S., & Chhetri, R. (2021). A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher Education for the Future, 8*(1), 133–141. https://doi.org/10.1177/2347631120983481

Porta, M. (2018). La importancia de la evaluación diagnóstica en el proceso de enseñanza-aprendizaje, tanto para docentes como para estudiantes. La problemática de la nivelación en grupos heterogéneos. *Reflexión Académica en Diseño y Comunicación, 19*(35), 179–181. https://revistas.uasb.edu.ec/index.php/ree/article/view/3150/3273

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Sánchez Mendiola, M., García Minjares, M., Martínez González, A., & Buzo Casanova, E. (2020). El examen de ingreso a la Universidad Nacional Autónoma de México: evidencias de validez de una prueba de alto impacto y gran escala. *Revista Iberoamericana De Evaluación Educativa, 13*(2), 107–128. https://doi.org/10.15366/riee2020.13.2.006

Sánchez Mendiola, M., & Martínez González, A. (2022). *Evaluación y aprendizaje en educación universitaria:estrategias e instrumentos* (1a ed.). Ciudad de México: UNAM. https://cuaieed.unam.mx/publicaciones/libroevaluacion/

Sánchez-Mendiola, M., & Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. *Inv Ed Med, 6*(21), 52–62. https://doi.org/10.1016/j.riem.2016.12.001

Schult, J., Mahler, N., Fauth, B., & Lindner, M. A. (2021). Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/pqtgf

Şenel, S., & Şenel, H. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education, 8*(2), 181–199. https://doi.org/10.21449/ijate.820140

Skye, W. F. S., Feinberg, D. K., Zhang, Y., Chan, M., & Wagle, R. (2020). Assessment during the COVID-19 pandemic: ethical, legal, and safety considerations moving forward. *School Psychology Review, 49*(4), 438–452. https://doi.org/10.1080/2372966X.2020.1844549

ten Cate, O. (2001). What happens to the student? The neglected variable in educational outcome research. *Advances in Health Sciences Education: Theory and Practice, 6*, 81–88. https://doi.org/10.1023/A:1009874100973

UNAM-DGPL. (2022). *2022 agenda estadística UNAM*. Cuadernos de planeación universitaria. Ciudad de México, UNAM. https://www.planeacion.unam.mx/Agenda/2022/disco/index.html. Accessed 8 June 2023.

UNESCO. (2020). *Exámenes y evaluaciones durante la crisis del Covid-19: Prioridad a la equidad*. https://es.unesco.org/news/examenes-y-evaluaciones-durante-crisis-del-covid-19-prioridad-equidad

UNESCO. (2021). *One year into COVID: Prioritizing education recovery to avoid a generational catastrophe*. https://en.unesco.org/news/one-year-covid-prioritizing-education-recovery-avoid-generational-catastrophe

Valle, R. (2012). El Sistema Exámenes de diagnóstico y Autoevaluación y estudio de asignaturas del bachillerato de la UNAM. *Revista Mexicana De Bachillerato a Distancia, 4*(8). https://doi.org/10.22201/cuaed.20074751e.2012.8.44271

Ying, Y., Siang, W., & Mohamad, M. (2021). The challenges of learning English skills and the integration of social media and video conferencing tools to help ESL learners coping with the challenges during COVID-19 pandemic: A literature review. *Creative Education, 12*, 1503–1516.

Authors and Affiliations

Melchor Sánchez-Mendiola[1,2] · Abigail P. Manzano-Patiño[1] · Manuel García-Minjares[1] · Enrique Buzo Casanova[1] · Careli J. Herrera Penilla[1] · Katyna Goytia-Rodríguez[1] · Adrián Martínez-González[1,2]

✉ Melchor Sánchez-Mendiola
melchorsm@gmail.com

Abigail P. Manzano-Patiño
abigail_manzano@cuaieed.unam.mx

Manuel García-Minjares
manuel_garcia@cuaieed.unam.mx

Enrique Buzo Casanova
enrique_buzo@cuaieed.unam.mx

Careli J. Herrera Penilla
careli_herrera@cuaieed.unam.mx

Katyna Goytia-Rodríguez
katyna_goytia@cuaieed.unam.mx

Adrián Martínez-González
adrian_martinez@cuaieed.unam.mx

[1] Department of Educational Evaluation, Coordination of Open University, Educational Innovation and Distance Education, National Autonomous University of Mexico (UNAM), Mexico City, Mexico

[2] Faculty of Medicine, UNAM, Mexico City, Mexico