So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education. Australia: Asia-Pacific Society for Computers in Education

Can Computer-Based Learning Environments Mitigate Large Class Size?

Ryan S. BAKER[ab*], Aishah AL YAMMAHI[c], Joe EL SEBAALY[c], Ali NADAF [c], Tina KAPP [c], & Seth ADJEI[d]

[a]*Teachers College Columbia University, United States of America*
[b]*University of Pennsylvania, United States of America*
[c]*Alef Education Consultancy LLC, United Arab Emirates*
[d]*Northern Kentucky University, United States of America*
*ryanshaunbaker@gmail.com

Abstract: In this paper, we study the impact of class size in several schools in the United Arab Emirates using a computer-based learning environment as part of their regular classroom practice. Within this study, we compare larger and smaller classes in terms of both learning and time on task. Classes in the same school are compared directly, and a matching quasi-experimental study is also conducted, comparing students across schools. We find no evidence that the larger classes perform more poorly than smaller classes.

Keywords: Class Size, Computer-Based Learning Environments, Implementation

1. Introduction

In many countries, policy-makers and the general public tend to believe that smaller classes are better for children. In Western Europe and North America, many studies seem to suggest that smaller class sizes are associated with better learning outcomes (Glass, 1982; Shin & Chung, 2009). Others have argued that different teaching practices reduce the importance of class size in East Asia (Stigler et al., 1982; Blatchford et al., 2016). Although there has been less research on this topic in the rest of the world, there is some evidence that smaller class size is associated with better learning outcomes in the Middle East, North Africa, and sub-Saharan Africa (Altinok & Kingdon, 2012). However, the desire to decrease class size clashes with worldwide challenges in recruiting and retaining highly-skilled teachers (Madalinska, 2018). This problem has led many to call for improving teaching practices and increasing teacher ability, rather than reducing class sizes (Stigler et al., 1982; Blatchford et al., 2016).

There is now evidence that computer-based learning environments (CBLEs) can promote enhanced pedagogies that may scale better than traditional approaches to teaching (Miller et al., 2015). The majority of students using these systems work at their own pace, receiving help and feedback from the learning system; teachers focus their time and attention on the specific students that are struggling at any given moment. This strategy has the potential to enable master teachers to take on a larger number of students. However, to the best of our knowledge, there has not yet been research on whether the class size remains associated with student outcomes when students learn from CBLEs in class.

We study this question in the context of an initiative that began in the United Arab Emirates (UAE) in 2017. A public girls' school piloted using computer-based learning to mitigate the impact of teacher absenteeism, bringing two grade 6 classes together into one larger room as needed. As a result of this experience, the school designed a larger pilot in the fall term of 2018. Two Grade 7 classes for English, Science, and Math were combined into a single class of 60 students. At the same time, the computer-based learning initiative expanded to 6th-8th graders at 57 other local schools in the UAE.

In this paper, we conduct quasi-experimental comparisons to analyze the impact of doubling class size on student time on task and achievement. We compare the students in the large class both to another class in the same school, and separately match the students to comparable students in two other, larger schools. The hypothesis of this study is that larger classrooms using a well-implemented computer-based learning curriculum will not have poorer student learning than smaller classrooms.

2. Methods

2.1 Learning System

We study this difference in class size in the context of the Alef computer-based learning environment. By the 2018-2019 school year, Alef was in use by over 25,000 6th-8th grade students in public sector schools in the UAE, spanning 6 core subjects: Mathematics (English), Science (English), English Language Arts (English), and Arabic Language Arts (Arabic), Social Studies (Arabic), and Islamic Studies (Arabic). Alef follows the UAE Ministry of Education's national standards, aligned to elements of CCSS (Common Core State Standards) and NGSS (Next Generation Science Standards).

Alef lessons cover specific sets of skills through an instructional process of Explore, Apply, Relate. Students acquire their skills by moving from tasks associated with the initial recall and recognition of foundational concepts and procedures to their application and extension. Within a lesson, the platform detects students' weaknesses through continuous formative assessment. It analyzes student performance on summative assessments as well, and identifies skills that the students are lacking. The assessment-performance data drives automated remediation and practice, and teachers receive real-time data to target support to students with learning differences and needs for offline intervention.

Alef's content includes both content instruction (in the form of multimedia, videos, and text) and skills application, as shown in Figure 1. Skills practice takes the form of sets of problems to complete. Teachers are also supported in this process through the provision of a series of offline experiential learning kits, usually consisting of manipulatives, simulations, and other hands-on activities. All data collected in this study was drawn from the English language subjects.

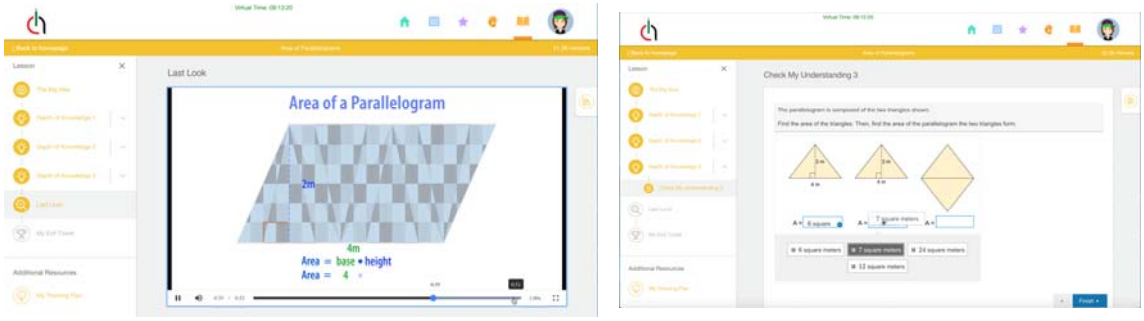

Figure 1. Alef video instruction (left) and assessment of problem-solving skill (right).

2.2 Participants

Data was obtained for three public seventh-grade classes in girls' schools in Abu Dhabi, UAE. The schools studied in this paper were selected for similar demographics, socioeconomic status, and academic performance profiles, and the students are majority Emirati.

In school A, students were divided into two classes: a large class of 61 students (observation data was obtained for all students; some test data was unavailable for 6 students), and a smaller class of 31 students (test data was obtained for all students; 4 students were absent during the observation days). Students were assigned to different classes by each school; random selection was not used. In schools B and C, 8 additional smaller classes participated, with a total of 204 students (ten absent during the observation days). Each student in the study used Alef as their primary curriculum throughout the course of the school year. As with school A, student-level demographic data is unavailable. Student-level demographic data was unavailable, creating limitations for applying algorithms such as propensity score matching to compare conditions.

2.3 Procedure

We obtained the following data for each of the students in our sample:

Two mathematics diagnostic tests, given in October 2018 and May 2019, using the benchmarked math computer-adaptive test STAR-MATH. On average, students performed 0.62 academic years better on the second test than the first test. STAR-MATH provides two different types of test

scores: Criterion-referenced scores (performance relative to absolute criteria) and Norm-referenced scores (performance relative to other students). In this study, we use Grade Equivalent (GE) Criterion-referenced scores for both diagnostic tests, indicating a grade level for a student in terms of the average knowledge a student would be expected to have in the subject at that grade. The GE score has a range from 0 to 13. These Grade Equivalents are in terms of a US context, but are appropriate for use in the UAE due to the commonalities between the topics taught in each country. Students in the UAE and US need not reach exactly the same point at the same time as students for this measure to differentiate each student's overall progress.

- Final mathematics examination grade comes from a national exam to measure pupils' knowledge in all public schools, conducted by the UAE Ministry of Education (MOE).
- Classroom observation data for the students in the sample. BROMP (Baker Rodrigo Ocumpaugh Monitoring Protocol; Baker et al., 2020) was used to collect observations of whether students were on-task or off-task, and whether they were experiencing frustration, boredom, engaged concentration, delight, or confusion. BROMP is a momentary time-sampling method where trained observers make holistic assessments of learner behavior and affect based on a combination of facial expression, posture, and other factors. BROMP has been used in research and practice by over 160 researchers in 6 countries.
- Interaction data from within the Alef platform.

For performance, we focused on mathematics, as diagnostic tests were only available for this subject. For time on task and affect, our observations (three sessions per student) cut across Mathematics, English, and Science, due to the applicability of the method in all cases.

3. Results

3.1 Performance on Exams

Overall, students appear to improve over the course of the year. Across all groups, students improve from an average of 4.45 (SD=1.69) at the first diagnostic test to an average of 5.17 (SD = 2.26) at the second diagnostic test, $t(286)=7.94$, $p<0.001$. This gain in performance has an effect size (Cohen's D) of 0.47, in the range that Hattie (2012) refers to as the "zone of desired effects" and – in Hattie's view – beyond what even an expert teacher can produce without special practices or programs.

Taking the groups individually, students in the large class had a statistically significant gain from the first diagnostic exam to the second diagnostic exam, $t(54)= 3.92$, two-tailed $p<0.001$, 0.93 average gain, Cohen's D = 0.53. In the smaller class in the same school, the difference in performance between the two exams was not statistically significant, $t(27)=0.927$, two-tailed $p=0.36$, 0.14 average gain, Cohen's D = 0.18. The average gain was significantly higher for students in the large class than students in the smaller class in the same school, $t(80.47)=2.79$, $p=0.007$, assuming unequal variances.

Students in the smaller classes in different schools had a statistically significant gain from the first diagnostic exam to the second diagnostic exam, $t(203)= 6.83$, two-tailed $p<0.001$, 0.74 average gain, Cohen's D = 0.48. The average gain was not statistically significantly higher for students in the large class than students in the smaller classes in different schools, $t(77.67)=0.75$, $p=0.46$.

As one possible confound, it appears that students were somewhat stronger at the time of the first diagnostic test in the large class (M=5.41) than in the smaller classes in the same school (M=4.11, $t(80.71)=4.18$, $p<0.001$) or the smaller classes in the other schools (M=4.24, $t(78.90)=4.29$, $p<0.001$). We consider how to address this possible confound in the next section.

On the final examination, students in the large class (M= 76.1, SD=16.6) were outperforming the smaller class in the same school (M=63.0, SD=21.0), $t(55.87)=2.28$, $p=0.026$, D=0.69, and students in the smaller classes in the other schools (M=67.4, SD=16.2), $t(105.42)=4.87$, $p<0.001$, D=0.53.

Overall, then, our findings do not suggest that being in a larger class hurt student performance.

3.1.1 Matching Analysis

Given that students in the large class had better pre-tests than students in the two comparison conditions, there is some risk that our results are due to a confound. We can address this possible confound, at least in part, through a matching analysis, where we match students in the large class to

Table 1

Performance by type of class. Averages given, with standard deviations in parentheses. Cases where large school's values are statistically significantly higher are in boldface.

Type of Class	First Diagnostic Test	Second Diagnostic Test	Gain	Final Examination
Large	5.41 (1.84)	6.34 (2.66)	0.93 (1.76)	76.1 (16.6)
Smaller (same school)	**4.11 (0.99)**	**4.26 (0.92)**	**0.14 (0.81)**	63.0 (21.0)
Smaller (other schools)	**4.24 (1.65)**	**4.98 (2.17)**	0.74 (1.54)	67.4 (16.2)

similar students in the smaller classes at other schools and then compare the matched students. In doing so, it would be ideal to have full demographic information enabling us to use a sophisticated statistical method such as inverse probability of treatment weighting or propensity score matching. Unfortunately, this information was not available, so we match students based on their performance on the first diagnostic test. We conduct this matching process 1000 times to see how robust the results are to the details of the match. Within this matching process, every student in the large class was represented in the matching and a subset of students in the smaller classes in other schools were matched to the students in the large class, without replacement. Specifically, we randomly shuffle the order of students in the large class, and then for each student in the large class, we find the student closest to them within the smaller classes, removing the students from the data set after matching (i.e. a student cannot be matched twice). In cases where multiple students in the smaller classes had pre-test scores equidistant to the student in the large class being matched, a student was randomly selected to create a matched set.

After matching based on the first diagnostic tests, the data sets are almost identical in terms of that variable, with an average of 5.411 in the larger class and an average of 5.431, across runs (cross-run SD of average = 0.067), in the matched cohorts. The standard deviations for the first diagnostic test are also almost identical, 1.820 for the larger class and an average standard deviation of 1.856 across runs for the matched cohorts (cross-run SD of SD = 0.136).

On the second diagnostic test, the larger class averaged 6.344 and across runs the matched cohort averaged 6.432 (cross-run SD of average = 0.081). This led to very similar levels of gain between the two groups: 0.933 for the larger class and 1.001 for the matched cohort. The difference in gain between the two groups was not statistically significant in any run.

However, the larger class had substantially better performance on the final exam, 70.82, than the average final exam performance of students in the matched cohort, 62.65 (cross-run SD of average = 0.154). The difference in final exam performance between the two groups was statistically significant in 997 of 1000 cases. The average t value across runs was 2.81 ($p = 0.007$); the standard deviation of the t value across runs was 0.0148. a statistically significant difference, paired $t(54)=3.59$, $p<0.001$, average Cohen's D across runs =0.466 (cross-run SD of average = 0.028).

Hence, even after we control for pre-test differences through matching, we still find that the students in the large class obtain higher final examination scores than the students in the smaller classes.

3.2 Affect and Time on Task

The first step to conducting BROMP observations in a new country is to create and refine a locally-appropriate coding scheme (Baker et al., 2020). The finalized coding scheme is tested through an inter-rater reliability round where two coders label student behavior and affect separately but

Table 2

Performance by type of class, for matched cohorts. Averages across runs given, with cross-run standard deviations of averages in parentheses. Cases where large school's values are statistically significantly higher are in boldface.

Type of Class	First Diagnostic Test	Second Diagnostic Test	Gain	Final Examination
Large	5.41	6.34	0.93	70.8 (15.4)
Smaller (matched cohort)	5.43 (0.07)	6.43 (0.08)	1.00 (0.15)	**62.7 (0.15)**

simultaneously. Three coders conducted observations at this study's schools. The coders labeled 1,747 observations of up to 20 seconds duration, in 10 classes in 3 schools, including the large classroom, the smaller classroom in the same school, and 8 more comparison classrooms in the two other schools. The coders established inter-rater reliability at School A, two months prior to this study's data collection. Coders 1 and 2 obtained Kappa of 0.91 for both behavior and affect. Coders 1 and 3 obtained Kappa of 0.72 for behavior and 0.71 for affect.

We compare the proportion of affect and behavior between the larger classroom and the smaller classrooms – focusing on engaged concentration, frustration, boredom, and off-task behavior -- using a two-sample t-test (assuming unequal variance), and using a Benjamini & Hochberg (1995) post-hoc control to adjust for conducting multiple comparisons. The students in the large class displayed significantly more frequent engaged concentration (76%) than the students in the other schools (51%), $t(118.96)=6.23$, $p<0.001$, adj $\alpha = 0.00625$, and also displayed marginally more frequent engaged concentration (76%) than the students in the smaller class in the same school (60%), $t(39.56)=2.14$, $p=0.038$, adj $\alpha = 0.025$. The students in the large class displayed significantly less frequent boredom (5%) than the students in the other schools (15%), $t(165.80)= -4.38$, $p<0.001$, adj $\alpha = 0.0125$. However, there was not a significant difference in boredom between the large class (5%) and the smaller class (2.3%), within the same school, $t(72.82)=1.15$, $p=0.252$, adj $\alpha = 0.0375$. The students in the large class displayed significantly less frequent frustration (0.4%) than the students in the other schools (2.4%), $t(243.87)= -2.58$, $p=0.010$, adj $\alpha = 0.01875$. However, there was not a significant difference in frustration between the large class (0.4%) and the smaller class (1.2%), within the same school, $t(33.62)= -0.65$, $p=0.519$, adj $\alpha = 0.0435$. No significant or marginally significant differences were seen between classes in terms of off-task behavior.

Table 3

Proportion of affect and behavior by type of class. Boldface indicates statistically significantly worse than large class; Italics indicate marginally worse than large class.

Type of Class	Engaged Conc.	Boredom	Frustration	Off-Task
Large	76.0%	5.1%	0.4%	11.0%
Smaller (same school)	**60.3%**	2.3%	1.2%	12.6%
Smaller (other schools)	*51.4%*	**15.4%**	**2.4%**	6.5%

3.3 Usage Data

We can also compare students between the larger and smaller classes in terms of their usage of the Alef system, shown in Table 4. There were not statistically significant differences in terms of the time-on-task between classes, though the difference between the larger class and smaller classes in other schools approached significance, $t(94)=1.60$, two-tailed $p=0.113$ assuming unequal variance. There were also no statistically significant differences between groups in terms of the average length of each learning session. Furthermore, we found that there were no statistically significant differences in the number of days the system was used at home by students in each group (based on the day of the week and time of day).

4. Discussion and Conclusions

The results of this study give no evidence for negative impact on the students in the 60-student

Table 4

Proportion home use by type of class. Averages given, with standard deviations in parentheses.

Type of Class	Total Usage (hours)	Session Length (min)	Days Home Use
Large	52.3 (25.7)	43. 9 (6.6)	12.3 (9.9)
Smaller (same school)	47.6 (32.0)	51.0 (10.5)	6.2 (8.3)
Smaller (other schools)	47.3 (16.2)	38.2 (5.4)	16.5 (11.3)

classroom. The 60-student classroom exhibited similar or better classroom time on task and academic performance as students in 30-student classrooms. Overall, students in the 60-student classroom appeared to perform acceptably relative to students in other classrooms.

These results do not conclusively demonstrate that there are no negative impacts to having larger classrooms when using a computer-based learning environment. There are several limitations to the existing study. For one thing, given the relatively small sample, it is possible that unaddressed confounds may have artificially enhanced the performance of the students in the larger class. For example, it is possible that the students in the large class were more prepared than students in the other classes, even beyond the differences addressed by controlling for differences in the first diagnostic test.

Another possibility is that the teacher in the large class was more skilled than the teachers in the other classes, producing the positive results seen, or that differences in teachers' pedagogies and implementation strategies may have compensated for having larger classes. Fine-grained details of implementation and pedagogy can be key to the effectiveness of computer-based learning environments in classrooms (Feng et al., 2014) – as such, we recommend that future work studying the impacts of class size in schools using computer-based learning environments should collect observational data on teachers' pedagogical strategies and implementation decisions, to help understand the learning and behavior seen in classes of different sizes.

Beyond these possible confounds, the sample is too restricted to draw general conclusions. Our current sample only involves girls-only schools; it is not clear that this pattern will hold in boys-only or mixed-gender schools. It is also unclear whether these results will be seen in other countries, or for other computer-based learning environments. Despite these limitations, though, this paper's findings suggest that the negative impacts of large class size may be mitigated through the use of computer-based learning environments. Given this finding's potential economic impacts, it warrants further study.

Acknowledgements

We thank Guzelle Shahid, Rand Muhsen, and Xin Miao for their assistance in study implementation.

References

Altinok, N., & Kingdon, G. (2012). New evidence on class size effects: A pupil fixed effects approach. *Oxford Bulletin of Economics and Statistics*, *74*(2), 203-234.

Baker, R.S., Ocumpaugh, J.L., & Andres, J.M.A.L. (2020). BROMP Quantitative Field Observations: A Review. In R. Feldman (Ed.) *Learning Science: Theory, Research, and Practice,* pp. 127-156. McGraw-Hill, New York, NY.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, *57*(1), 289-300.

Blatchford, P., Chan, K. W., Galton, M., Lai, K. C., & Lee, J. C. K. (Eds.). (2016). *Class size: Eastern and Western perspectives.* Routledge, Abingdon, UK.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In *International Conference on Intelligent Tutoring Systems* (pp. 561-566). Springer, Cham.

Glass, G. V. (1982). School class size: Research and policy. Sage, Beverly Hills, CA.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge, New York.

Madalinska, M., O'Doherty, T., & Flores, M. (2018). Teachers and teacher education in uncertain times. *European Journal of Teacher Education*, *41*(5), 567-571.

Miller, W.L., Baker, R., Labrum, M., Petsche, K., Liu, Y-H., & Wagner, A. (2015). Automated Detection of Proactive Remediation by Teachers in Reasoning Mind Classrooms. *Proceedings of the 5th International Learning Analytics and Knowledge Conference,* 290-294.

Shin, I. S., & Chung, J. Y. (2009). Class size and student achievement in the United States: A meta-analysis. *Korean Educational Development Institute Journal of Educational Policy*, *6*(2).

Stigler, J. W., Lee, S. Y., Lucker, G. W., & Stevenson, H. W. (1982). Curriculum and achievement in mathematics: A study of elementary school children in Japan, Taiwan, and the United States. *Journal of Educational Psychology*, *74*(3), 315.