Short-term Efficacy and Usage Recommendations for a Large-scale Implementation of the Math-Whizz Tutor

Manolis Mavrikis[1], David Sebastian Schlepps[1] and Junaid Mubeen[2]

[1] UCL Knowledge Lab, University College London, UK
[2] Whizz Education, London, UK
m.mavrikis@ucl.ac.uk,
junaid.mubeen@whizzeducation.com

Abstract. This paper adds to the evidence of the efficacy of intelligent tutoring systems (ITS) in mathematics learning by evaluating a large-scale intervention at the state of Aguascalientes, Mexico. We report the results of a quasi-experimental study, addressing at the same a particular request of the decision-makers responsible for the rollout to provide, from early stages of the intervention, independent evidence of the efficacy of Math-Whizz Tutor beyond its internal metrics, and recommendations in terms of the expected weekly usage levels to guide the blended learning approach.

Keywords: Intelligent Tutoring Systems, Evaluation, large-scale

1 Introduction

Although there is mounting evidence that intelligent tutoring systems, under the right conditions, can offer a significant advantage in supporting students' learning [1],[8], understandably educators or other stakeholders responsible for their adoption require evidence of large-scale evaluations and specific recommendations about classroom integration in their context.

Our case study involves the rollout of the intelligent tutoring system, Math-Whizz, in the state of Aguascalientes, Mexico. While previous studies have demonstrated positive results (e.g. [9], [6]) and Whizz Education has developed global usage guidelines for implementations based on historical data, the decision-makers in charge of the state-wide adoption required (i) guidance on how much time students should spend on the Math-Whizz tutor each week, and (ii) independent evidence, at early stages of the roll-out, that demonstrates the intervention's potential in their context.

This paper presents our approach to providing weekly usage recommendations for Math-Whizz in Aguascalientes based on internal system metrics, and reports the results of a quasi-experimental study evaluating the efficacy of the overall approach using a mixture of standardized exams and locally appropriate tests. The late addressing the common concern in the field that the type of test affects evaluation results [4]. As such, beyond the results of interest to the specific study, the paper makes a methodological contribution and aims to add to the debate of efficacy of intelligent tutoring systems in general.

adfa, p. 1, 2011.

2 Math-Whizz

Math-Whizz is an intelligent online tutor for 5 to 13-year-olds. It comprises 1200 learning objectives which have been organised into 22 topics and sequenced within each topic based on a curriculum map developed by educationalists. Maths-Whizz is being used by hundreds of schools in eight international territories and currently serves over 150,000 students worldwide.

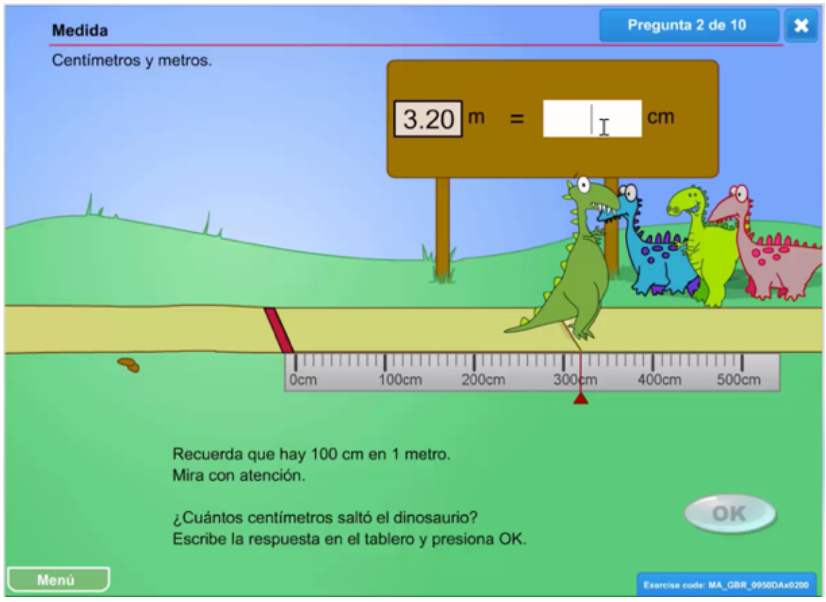

Figure 1 An example of a Maths-Whizz exercise

The student experience begins with an adaptive assessment that measures the student's knowledge across several maths topics. The tutor then guides students through an individualised learning pathway, privileging topics in which the student is behind with the goal of helping each student achieve a rounded learning profile.

Each lesson begins with a Teaching Page, which uses direct instruction to introduce the concept or method. This is followed by an interactive exercise, which use rich visual representations and scaffolded prompts to guide learners. During the exercise, a student receives hints (including the solution). If a student passes a certain threshold of correct answers, they are given a test where they can demonstrate what they learnt in a different context than the initial activity.

The tutor uses real-time learner data to offer remedial support when students are stuck and advancing them when they have demonstrated a good understanding of each learning objective. The topic sequencing policy thus relies on the assumption that a student will be able to solve the exercises of the selected difficulty level and only advances students when they have acquired the relevant prior knowledge. As such it implements a type of mastery learning approach, guided by its internal metric, 'Maths Age'™, which has a natural interpretation: a Maths Age of nine corresponds to the knowledge expected of a nine-year old according the curriculum map. Maths Age is calculated for each topic, and an arithmetic mean is then assigned as an overall Maths Age for each student. Maths Age is aimed at informing teachers and parents about students' mathematics ability [9]. In addition, the overall platform offers live reports as a monitoring tool for teachers, as well as a collection of instructional resources (including the bank of 1200 lessons) for classroom use.

3 Methodology

The overall evaluation approach in Aguascalientes followed a mixed methods approach that included both qualitative and quantitative data from a range of stakeholders. The qualitative part relies on observations in a range of schools, interviews with teachers, parents and students themselves as well as observation of teacher training sessions. In this paper, we focus on the quantitative analysis that relied on a quasi-experimental design, particularly a non-equivalent control group study [2] in two conditions: the Math-Whizz condition (MW) with schools that are implementing the intervention (and take part in the teacher training), and the non-users (NU) condition that included a range of state schools throughout Aguascaliantes. We do not refer to the latter as 'control' group because, for reasons outside the control of the first author, the design of the evaluation started after schools were already selected for the government Math-Whizz rollout. The NU schools are still potentially future MW schools for a second round of the roll-out. With the understandable threats to internal validity and generalisation of the results, although the group assignment (MW vs NU) was not explicitly controlled, the initial selection process to take part in the roll-out did not seem to introduce any selection bias and other factors like students' socioeconomic status, other government metrics were the same. All schools were also following the same curriculum and the main difference between the MW and NU schools were that the MW schools took time out of the normal class for the students to interact with Math-Whizz.

The focus of the work reported here are the primary school students aged between 8 and 9 years old. This is because of the availability of state-wide data from the Mexican PLANEA test (see http://www.planea.sep.gob.mx/), which we used as an achievement 'snapshot'. The need to provide an independent evaluation of the intervention at early stages meant that we could not rely on a state-administered test to measure students' levels of achievement as these are run at the start and end of the year. Due to other factors, including holidays and other school priorities we also had a limited time window (7 weeks) in which to apply a test. We used the corresponding PLANEA that runs every September as a pre-test (with relevant permissions granted). For the post-test (end of October) we selected 10 questions from the 'numbers and counting' and 'addition' problems, as these were among the topics covered by Maths-Whizz at this period. We will refer to these tests as Sep and Oct respectively from now on. Internal consistency for the Sep test was α=0.85 and for Oct α=.76. Note that Whizz Education did not make any changes to their adaptive algorithm, nor were they aware of the exact contents of the test, which was the responsibility of the first author.

4 Results

Due to the short duration of the implementation, we were pragmatic and did not expect large effect size in learning gains between the two conditions, particularly given the difficulty of isolating the effects of a complex socio-technical intervention to just the introduction of an ITS system. Nevertheless, despite the limitations (discussed in Section 5), the results are promising and warrant further research.

For a sample of 3407 4th grade students (N_{MW} = 2188, N_{NU}=1219), results were obtained using a linear regression model that resembles the ANCOVA method for measuring change in time and using cluster robust and heteroscedasticity-consistent standard errors [3,7]. Accounting for the differences in students' test achievement in September, the predicted achievement score for Maths-Whizz users in the sample is 0.659 points on a scale 0 to 10 higher than non-Whizz users ($\beta = 0.659$, $p < 0.05$).

If the change is not due to unobserved variables, this significant difference in progress seems associated with the Whizz intervention. Accounting for previous

achievement and for whether students belong to the Whizz user group explains around 18.6% of the differences in students' scores in October. A corresponding multilevel model suggests very similar values with an effect size of about *d=0.22*, commensurate with others in the area (e.g. [5]). Analysing the interaction effect of students' condition and their initial achievement in September, we found no significant relationship (β = -0.086 β = 0.659, $p > 0.05$), suggesting that the relationship between students' improvement and their membership in the Whizz user group did not depend on their previous achievement.

To derive both a recommendation for teachers and parents in relation to usage levels and a way to group students for analysis, we conducted a linear regression of time in the system again the internal Math Age metric of the system for the Mexico cohort.

This showed that a student needs to use the system approximately 33 mins per week to achieve a progress rate of one (which corresponds to an expected increase of Maths Age of 1.00 over one year) and 45 mins for a Progress Rate of 1.50. These findings are consistent with Whizz's global recommendations, suggesting that the effort required by Aguascalientes students to achieve learning gains on Maths-Whizz is comparable to the rest of the world. Accordingly, we create groups of high usage (45 min or more), minimum recommended usage (34-44 min), low usage (5-33 min) and very low (less than five mins) and conducted an additional cluster-robust and heteroscedasticity-consistent regression analysis [7] of the September and October test scores.

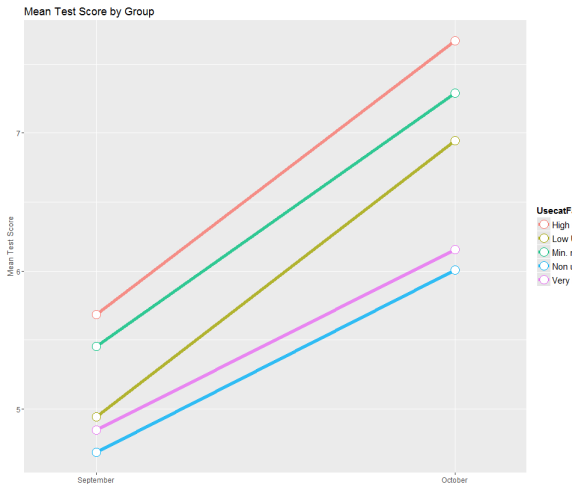

Figure 2 Mean score in Sep. and Oct. by usage group

Among other interesting results, we observe, first no significant difference between non-users and those who used the software less than five minutes (β = 0.081, $p > 0.05$). The high usage group is associated with an additional 1.26 point progress compared to the non-user group (β=1.255, $p<.001$). Similarly, a significant difference to the progress of the non-user group was found for minimum usage (β = 0.969 , $p < 0.01$) and low usage (β=0.831, $p<.01$). We briefly discuss these below.

5 Discussion

The results from the evaluation described in this paper add to the evidence of the efficacy of intelligent tutoring systems (ITS) in mathematics learning in a large-scale implementation at the state of Aguascalientes in Mexico. Of course, the disentanglement of causal relationships between the use of any technology and learning out-

comes and other factors that may distort the view on such relationships, is a well-known problem in the field[1]. As the data are from a real live context, some factors were outside the scope of this study and of course, this raises some limitations here. For example, there could be systematic differences between schools or homes that, unknown to us, led implicitly to initial intervention selection or self-selection in usage groups. The lack of any significant difference in the learning gains of non-users and those Whizz users with minimal usage, speaks against a selection bias that arises from prior differences between the groups. However, further studies should investigate the relationship of previous achievement and usage.

Lastly, qualitative and teacher survey data (not discussed here), paint a positive picture for the overall implementation attributing to its success other factors such as the intense professional development offered to the teachers and strong parental involvement at home. Taking also into account the novelty of the intervention and findings from meta-analysis such as [8] that short-term interventions appear generally more successful than more lengthy ones, future work should look at large-scale and long-term experimental evaluation that takes into account government initiatives on pupil testing, a robust sampling procedure and testing instruments and a systematic way to include student and teacher opinions and their role in the intervention.

Acknowledgements

This evaluation was possible with co-funding and support from the State of Aguascalientes. We would like to thank the teachers and students who took part in the studies and colleagues at UCL and Whizz Education for the support with the data analysis.

References

1. du Boulay, B. (2016). 'Artificial intelligence as an effective classroom assistant'. IEEE Intelligent Systems. 31(6) 76-81.
2. Fife-Schaw, C. (2006). Quasi-experimental designs. In Breakwell GM, Hammond S, Fife-Schaw C, Smith JA (2006) Research Methods in Psychology. SAGE
3. Graham, N. Arai, M. and Hagströmer, B. (2016). multiwayvcov: Multi-Way Standard Error Clustering. R version 1.2.3. https://CRAN.R-project.org/package=multiwayvcov
4. Kulik JA, Fletcher JD (2016) Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. Review of Educational Research 86:42–78
5. Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online Mathematics Homework Increases Student Achievement. AERA Open, 2(4).
6. Schlepps, D.S. (2015) Intelligent Tutoring Systems in K-12 Education – An Evaluative Study of Maths- Whizz and Maths Age. Unpublished MA dissertation. UCL
7. Snijders, T., and Bosker, R.. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second edition. London etc.: Sage Publishers, 2012
8. Steenbergen-Hu, S. & Cooper, H., 2013. A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), pp.970–987
9. Whizz Education Ltd., 2015a. FAQs. Parents Help [Online]. Available at: <http://www.whizz.com/help/parents-help/> Accessed Feb 2, 2017
10. Yearly, S., 2014. Report by Simon Yearley. In Pedagogical Foundations and Evidence of Impact for Maths-Whizz.

[1] c.f. OECD http://www.oecd.org/education/students-computers-and-learning-9789264239555-en.htm