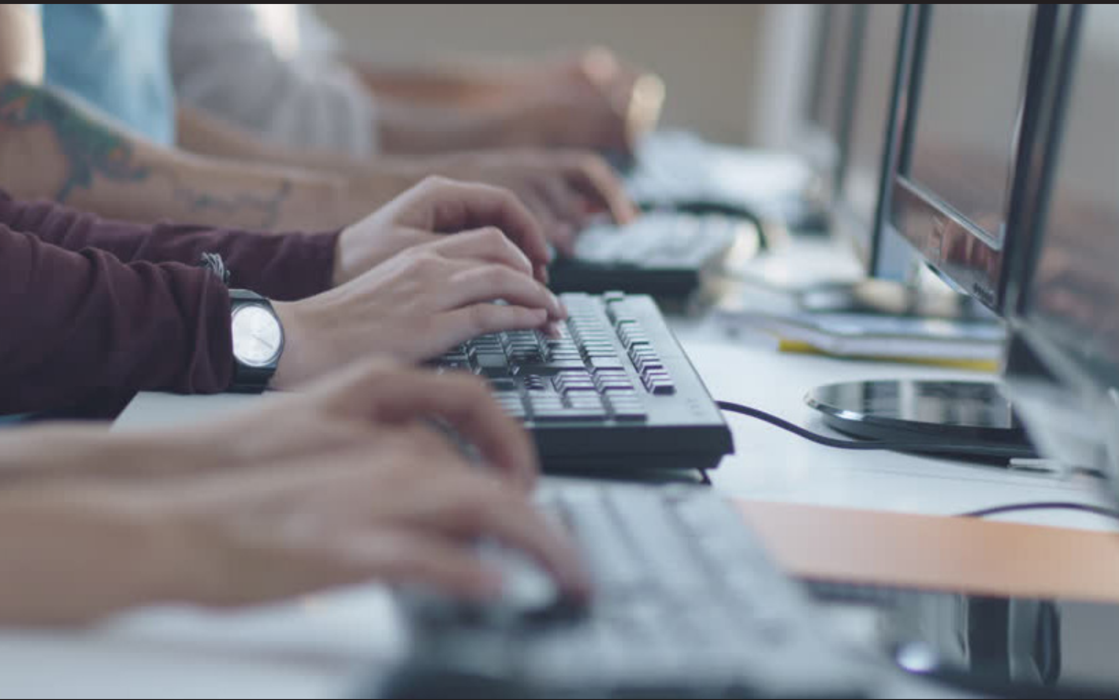
Annual research digest 2016-17

Evidence on uses of technology in education

CENTRE FOR EDUCATION ECONOMICS



23 Great Smith Street, Westminster, London, SW1P 3DJ • www.cfee.org.uk

Annual research digest 2016-17

CENTRE FOR EDUCATION ECONOMICS

First edition published 2017 by the Centre for Education Economics Ltd.

Typeset and design by Peter Johnson, Great White Designs Ltd

Centre for Education Economics Ltd
23 Great Smith Street, Westminster, London SW1P 3BL
www.cfee.org.uk
comms@cfee.org.uk
@cmr_ed

About the Authors

John Blake is Head of Education and Social Reform for the think-tank Policy Exchange. Prior to that, he was a senior school leader and history teacher for ten years in a variety of schools in London and Essex.

Daisy Christodoulou is the Director of Education at No More Marking, a provider of online comparative judgement. She works closely with schools on developing new approaches to assessment. Before that, she was Head of Assessment at Ark Schools, a network of 35 academy schools. She has taught English in two London comprehensives and has been part of government commissions on the future of teacher training and assessment. Daisy is the author of *Seven Myths about Education* and *Making Good Progress? The Future of Assessment for Learning*, as well as an influential blog. You can also find her on Twitter @daisychristo.

James Croft is Founder and Executive Director of the Centre for Education Economics (CfEE) and the author and co-author of several of its reports, including Optimising autonomy: a blueprint for education reform (2017) and Taking a lead: how to access the leadership premium (2016) and Collaborative overreach: why collaboration probably isn't key to the next phase of school reform (2015).

Nick Hassey is Associate Director of Research and Evaluation at Teach First where he undertakes long term research projects looking at the key drivers of lower attainment for young people from disadvantaged areas, what changes to government policy are likely to mean for schools, and the communities they serve, and the role of Teach First in addressing the issues these schools and communities face. Previously he worked at PwC and began his career at the polling company ComRes.

Carl Hendrick is the head of learning and research at Wellington College where he teaches English. He is also completing a PhD at King's College London in education. He has taught for several years in both the state and independent sectors where has worked on several cross sectoral collaborations and is a co-director of the Telegraph Festival of Education.

Gabriel Heller Sahlgren is Research Director at the Centre for Education Economics (CfEE), affiliated research fellow with the Research Institute of Industrial Economics, and a PhD student at the London School of Economics. He is the author of numerous publications on issues relating to applied microeconomics, including 'Retirement Blues' in the *Journal of Health Economics* and *Incentivising Excellence: School Choice and Education Quality* (CMRE and IEA 2013).

Contents

Introduction

by: Gabriel Heller Sahlgren

In the past decades, interest in the role of technology in education has surged worldwide, with many hoping it could become a game changer that radically improves education productivity. In recent years, therefore, various technology companies have begun to expand their offerings in the field. However, they rarely present independent research supporting the grandiose claims of the usefulness of these offerings – and whatever evidence is presented is generally of poor quality.

This research digest seeks to shed more light on the potential of technology in education by providing commentary on selected studies released in the past academic year, which analyse the effects of such technology in different settings.

The first paper, with commentary by John Blake, analyses the effects of classroom computer use on academic performance at a liberal-arts college in America. The paper finds that students who are induced to use computers perform considerably worse and that these effects are concentrated among males and low-performing students. In other words, according to this research, technology is more likely to be a bane than a boon when used indiscriminately in the classroom.

However, the effects of technology are likely to depend strongly on how it is introduced – and, more generally, what it is used for. In the second paper, I comment on a recent randomised experiment in India, which analyses the effects of a computer-led after-school programme. The programme exploits a software that precisely targets the learning level of each student and responds to changes in performance dynamically. The results display radical improvements in learning compared with the control group in just a few months, thus highlighting the potential of education technology for personalising instruction to the right learning level.

The third paper, discussed by Nick Hassey, provides further evidence that the impact of technology depends on how it is used, analysing whether different uses of education technology in Italian classrooms have different effects on pupil performance. Using technology for instructional purposes appears to raise pupil performance when it focuses on increasing pupils' awareness in technology use and when it enhances communication – but decreases performance when it requires students to take an active role. This supports prior evidence of negative effects of pupil-led learning compared with whole-class instruction, further indicating the way in which technology is used matters greatly for its effects.

This conclusion is further evidenced by the fourth paper, analysed by Carl Hendrick, which studies the relationship between computer use and performance at the university level. The researchers objectively measure the use of computers in the classroom and show that non-academic use of computers, which was highly prevalent in the study population, is negatively related to performance – while academic use of computers is unrelated to it. While the methodology is not perfect, it does indicate that education technology may harm performance if it acts as a distraction.

The penultimate paper, discussed by James Croft, offers important findings in relation to the claim that ICT-enabled personalisation improves productivity. The paper reports results from a randomised experiment analysing whether adaptive practicing is more effective than traditional means of test preparation. The authors evaluate a recently developed computerised practicing tool made available to pupils in a static version, which resembles the traditional approach in which the sequence of exercises is predetermined, and an adaptive version, which offers a sequence of exercises that is contingent on how they perform. The results display little difference between the two approaches. If anything, the adaptive version may generate slightly lower pupil achievement. However, contrary to expectations, static practicing improves test scores for higher-ability pupils somewhat. These findings appear contrary to those in the Indian experiment, indicating that the type and quality of software are crucial for the potential of learning personalisation via technology.

Of course, it is important to note that the potential of education technology is not only restricted to the instructional and learning process – but can be used for assessment purposes as well. The final paper, discussed by Daisy Christodoulou, analyses whether the use of 'comparative judgement', a method of assessment relying on an algorithm developed in the 1920s, could be used to assess students' conceptual understanding of mathematics – tasks which are normally very difficult to mark reliably. A key problem has been logistical challenges of using comparative judgement for large-scale assessments, but the creation of online comparative judgement engines has allowed for the algorithm to be used quickly and efficiently in ways that have not been possible in the past. The results provide evidence that comparative judgement is indeed reliable for this purpose, suggesting that technology has provided an opportunity to disrupt the marking process to make it more efficient and enable new types of pupil assessment.

Overall, therefore, while it is clear that technology is no panacea in education, the papers discussed here indicate that it may play an important role if it is implemented well and for the right purposes. It is up to future research to analyse more precisely in what ways technology could be a boon in the educational process – and which applications may undermine it.

Computers and Productivity: Evidence from Laptop Use in the College Classroom

- Richard W. Patterson, Robert M. Patterson

Economics of Education Review (April 2017)

Published version

Working paper version (free)

Commentary by John Blake

Are you involved in education? Do you like laptops? Well, I've some good news and some bad news for you, derived from this recent paper by two American academics on the impact of laptop use in the university classroom. The good news is that word laptop is used 532 times in this paper, so fill your boots with that; the bad news is that this is done in order to show that laptop use in class had a negative effect on student outcomes.

Given the ubiquity of educational technology evangelism this may well be something of a surprise. Certainly, university students do not seem to share this view: the use of ICT by students in American universities is increasing. In 2011, 57 per cent of university students in a paper cited here reported using a laptop, smartphone, or tablet in class; in the current study, 72 per cent of students used a laptop, and another study conducted at much the same time had 79 per cent of its cohort using laptops. This, however, is a trend at odds with the present state of the evidence on the impact of such technology on student grades. The authors reference several studies which examined correlations between outcomes and laptop use and came up negative. A randomised control trial conducted at the United States Military Academy (USMA) found that allowing laptops decreased student outcomes. Even those studies the authors cite which did not generate negative conclusions were not rich in support for the general use of laptops: one found no impact at all, and the other found a positive impact, but only for minority students.

However, these studies all present methodological problems. The correlational studies might well be explained by selection issues, in which the students who chose to use laptops were different in meaningful ways from those who did not. In addition, the USMA has, as the authors put it in a footnote, “a somewhat unique learning environment”, which seems a tad euphemistic for an academic institution where the students are army cadets and under military discipline. In other words, it’s not clear whether we can draw too strong conclusions for lap top use in general from this paper.

To overcome these problems, this paper takes advantage of an unnamed private liberal arts college’s flexible laptop policy. In this college, all students are required to possess a laptop, and all teachers may decide which of three rules will apply to their classroom environment: (1) laptops are required; (2) laptops are optional; or (3) laptops are prohibited. In general, the staff of the institution shared a positive view of the utility of laptops to learning: 57 per cent said they felt they increased learning, whilst only 25 per cent said they did not. Whatever their view, faculty members were, in general, not willing to enforce their convictions on their students: 20 per cent of them required laptops, 67 per cent allowed them, and only 4 per cent prohibited them.

Given these timetabling arrangements, the authors surveyed 229 of the more than 5000 students to discover in which classes the students were using laptops, and whether they had other classes that day which either required or prohibited laptops. They use this data to conclude that laptop policies influence the use of laptops in laptop-optional classes: having a laptop-required class on the same day increased the probability that a student would use their laptop in a laptop-optional class by 21 per cent, whilst having a laptop-prohibited class decreased the probability by 49 per cent.

They then examine the students' academic outcomes in laptop-optional courses and conclude that the impact of having at least one laptop-required course on grades in laptop-optional courses is consistently negative – whilst the impact of having at least one laptop-prohibited course is positive. This they explain as the effect of laptop use since requirement increases the probability of laptop use in other classes, whilst prohibition decreases it. They examine the data through prisms of race, gender and prior attainment, and conclude that whilst race is not a strong predictor of response to laptop use, gender is: male students account for much of the observed effect, which the authors suggest may be consistent with other research suggesting young males tend to have weaker non-cognitive skills. As well as having a strong gender component, the study suggests that the impact of laptop use is only really observable amongst weak students (as defined by their prior attainment, measured here by their Grade Point Average).

Overall, then, the authors conclude that their results suggest that laptop use directly worsens academic outcomes for students who choose to use them. They acknowledge that there are some weaknesses in their approach: the methodology does not permit direct measurement of laptop impact on academic performance. Further, the focus is on students who may or may not choose to use a laptop, not on those who would never pick one up nor those who could never bear to be parted from them. Thus, although the results suggest a classroom-wide ban on laptop use would have some academic benefits, the impact of the dynamics of a class may go well beyond those hinted at here.

However, what is striking is the extent to which these results cut against the prevailing sense identified by the authors amongst both the students and the teachers that laptop use was a positive feature. This is a subtle tug on the sleeve to consider whether the growing ubiquity of technology in the professional educational environment is a trend to be welcomed.

Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India

- Karthik Muralidharan, Abhijeet Singh, and Alejandro J. Ganimian

NBER Working Paper No. 22923 (December 2016)

Published version (free)

Commentary by: Gabriel Heller Sahlgren

In the past decade, interest in the role of technology in education has surged worldwide, with many believing it could be a game changer that disrupts education and radically increase productivity. Consequently, large technology companies have begun to expand their education offerings. Most often, no independent research supporting the grandiose claims of the usefulness of these offerings is presented – and in the few cases evidence is presented, its quality is generally poor.

As most papers in this digest (and elsewhere) highlight, using technology in education is certainly no panacea for improving pupil outcomes. This is not necessarily because technology is bound to fail – but rather indicate that the exact details of how and what type of technology unsurprisingly matters for the results.

In this paper, the authors present evidence from a randomised experiment in India, which sought to test the effects of a technology-led instructional programme called “Mindspark”. In India, the programme is used by over 400,000 pupils and administers more than a million questions every day. It exploits these data to benchmark the learning level of each pupil and then match the content with the individual pupil’s level and progress. Also, it analyses the data to identify pattern of errors made by students and then targets the material to get over ‘bottlenecks’ that may be difficult to overcome in a classroom setting.

The authors focus solely on the effects of using the programme at after-school study centres, which scheduled six days of instruction per week (90 minutes per day). In each session, pupils spent 45 minutes of self-driven learning via the programme and 45 minutes in groups of 12-15 pupils where they received instructional support from teaching assistants. These centres target low-income neighbourhoods in Dehli and normally charge a small fee.

The sample was composed of 619 pupils from state schools in Dehli. About 50 per cent of these pupils were randomly assigned to receive a voucher so they would receive the after-school instruction free of charge for 4.5 months. The control group was told that they would be provided free access after these 4.5 months had passed if they participated in an assessment at the end of the period. Crucially, pupils in the control group were not allowed to access the programme during the 4.5 period.

The results first of all show that the pupils in the sample are on a level several years lower than what would be expected, suggesting that the level of classroom-instruction in their schools is not set at an appropriate level for them. This suggests that the computer programme could be useful for the purposes of setting an appropriate level of instruction more generally.

Second, the authors find exposure to the programme increased test scores by the equivalent to 59 PISA points in mathematics and 36 PISA points in Hindi, a radical and remarkable improvement in just 90 days. This impact does not vary depending on baseline test score, gender, or socioeconomic background. However, the relative improvement was much larger among low-performing pupils because this set of pupils barely made in any progress at all in the control group during the intervention period. Moreover, the effects were achieved with lower per-pupil expenditures than in the state school system. The programme was therefore extremely cost-effective at raising pupil performance.

Interestingly, the authors document that the questions presented to pupils are precisely targeted by the software to their individual-learning level – and updates its target in response to how the performance of pupils changes. This indicates that the software's ability to identify and adopt to different types of pupils dynamically is an important mechanism behind the impact. More broadly, it also indicates that the personalisation of learning levels may be key to reap the benefits of education technology. Indeed, this is also suggested by other papers cited by the authors, which do find a positive impact of education technology.

Still, it is important to also note the limitations of the paper, which analyses the impact of after-school studying in a controlled environment using the computer programme compared with no extra formal tuition in the same period. It may be that the impact of the programme to a certain extent reflects the extra studying in the centres rather than solely the software. The authors cite other randomised research in the same setting, and during the same period, which found no impact at all of group-based tutoring despite pupils getting more time of instruction than in the computer-programme intervention. It is unlikely, therefore, that the effects are solely or even partly due to the extra instruction regardless of the computer programme – but the concern cannot be disregarded entirely.

It is also important to note that it is not clear whether the programme would be equally effective in a school environment. More generally, it is unclear whether the software works only as a complement to regular classroom instruction or whether it could be used to replace such instruction, as well as whether it works if pupils work unsupervised rather than in study centres. Thus far we only know that the programme worked as part of a structured after-school study programme. Furthermore, as the programme focused on a poor segment in India, it is difficult to draw too strong conclusions for developed countries.

Nevertheless, the paper provides one of the most interesting and promising results in the history of research into the workings of education technology. At the very least, they indicate that well-designed interventions based on relevant education technology have the potential to radically increase productivity and quality – and this must be investigated further in future research. While much other research hitherto indicates disappointing results of technology for the purposes of raising pupil performance, this paper clearly shows it would be a mistake to write it off entirely.

Is it the Way they Use it? Teachers, ICT and Student Achievement

- Simona Comi, Gianluca Argentin, Marco Gui, Federica Origo, and Laura Pagani

Economics of Education Review (February 2017)

Published version

Working paper version (free)

Commentary by: Nick Hassey

In recent years, an increasing body of research has begun to evaluate the effects of education technology on pupil achievement. Yet it is likely that the impact depends on how such technology is utilised – not just whether or not it is used at all. Unfortunately, there's little research on whether or not the ways in which teachers exploit technology matter for outcomes.

In this fascinating paper, the authors seek to remedy this situation by investigating what Italian teachers actually do with the technology at hand – and whether different practices have different effects on pupil learning. The researchers use a combination of surveys and standardised tests to study how different uses of technology corresponds to pupil learning.

Although observational, the study is well designed and the authors offer useful and proportionate advice to policymakers in their conclusions. To try overcome the risks of selection bias, the researchers exploit some interesting aspects of the Italian education system. For instance, although parents choose schools, pupils are usually randomly allocated to classes within schools – and this class is the same for all subjects and for the entire duration of upper-secondary school. The authors also manage to broadly control for unobserved pupil characteristics by comparing pupil performance across two different subjects, in this case Italian and mathematics. This approach partials out any unobservable characteristics that affect pupil performance in the two subjects equally.

The authors report high response rates – 94 per cent of schools contacted and all pupils asked within these schools responded to the survey – and set out how teachers were matched to the actual classes they taught. The researchers also collect standard demographic data on pupils and teachers, and even try to identify teacher motivation for, and competence in, technology use through their survey and standardised test scores. This goes some way to addressing potential issues around unobserved teacher factors, especially the likelihood that teachers who are good with education technology self-select into using it, but is not sufficient to control for it entirely.

Nevertheless, the authors use a neat robustness check to study whether self-selection of teachers biases the results: they study scores from TIMSS, an international mathematics and science survey where children were asked similar questions on technology use. Italian pupils are taught by the same teacher in mathematics and science when they sit the TIMSS test. This allows the researchers to compare pupil performance across subjects and taught by the same teacher. This entirely solves the potential problem of teacher quality confounding the effect of different types of technology use.

To investigate technology use, the researchers construct five 'factors' of technology teaching practice: using technology to deliver information to students more effectively (knowledge transmission), like the awareness of digital risks (evaluation of website content or how to avoid viruses); privacy rules or netiquette in social media (media education); active interaction between pupils, teacher, and ICT (active involvement); the use of ICT to assist in production of materials and other background components of teaching (backstage activities); and finally assisting the speed and ease of communications within and across school boundaries (communication).

After including the full-range of controls, the researchers find all the considered factors influence student achievement – but that backstage activities, such as helping a teacher print material for their lesson, has no statistically significant effects, and that active involvement is negative and statistically significant. Importantly, the authors also find that the impact of active involvement in TIMSS – where teacher quality and characteristics are held constant – is similar as in their main analysis, suggesting that teacher quality does not bias the findings in this respect at least.

The researchers then test both a simplified model to see if ICT use in general is associated with increased performance, and a more detailed test for heterogeneity in effectiveness of each of the different factors by subject taught. In these tests, the researchers find that ICT in general has no effect either way on pupil performance. This aligns with the wider literature into ICT in schools and implies that increasing access to, or spend on, technology indiscriminately won't increase learning.

Importantly, the results also reveal some heterogeneity: ICT teaching practices that enhance communication and knowledge transmission (e.g. through well-chosen video or audio aides) significantly increase student performance only in maths – while the effect in Italian is not statistically significant. However, this does not necessarily reflect that the impact of ICT is concentrated in one subject; the negative effect of practices requiring the active involvement of pupils in the use of ICT is much larger and statistically significant in the case of Italian language.

The authors speculate that new technology may be more effective when used to transmit knowledge in scientific fields, and that the active use of technology by pupils in the humanities could be particularly bad because 'the contents and approach of the subject are more likely to favour an unproductive use of the Internet'.

However, these are only speculations and would not support policies that ban the use of technology in humanities classes. Instead, the researchers position the paper merely as a caution against policymakers rushing into spending fortunes on ICT without 'well-designed pilot interventions assessed through RCTs to get accurate information about the conditions that render students' active use of ICTs beneficial'. Given the variation across different approaches and subjects – and the potential for some ICT practices to apparently harm pupil performance – it is easy to agree with this conclusion.

Logged In and Zoned Out: How Laptop Internet Use Relates to Classroom Learning

- Susan M. Ravizza, Mitchell G. Uitvlugt, Kimberly M. Fenn

Psychological Science (December 2016)

Published version (free)

Commentary by: Carl Hendrick

As technology becomes ever more powerful, an increasing number of schools and universities are adopting a BYOD (bring your own device) policy. This is especially prevalent at university level where many courses even recommend bringing a laptop to lectures and seminars. Despite the evidence that taking notes on a laptop is less effective than note taking and the distracting influence of technology in academic settings, many students underestimate their deleterious effects. For example, in one study, 62 per cent of students did not see an issue with texting in class so long as other students were not impacted.

This paper is one of the first that uses objective measures of classroom internet use for academic and non-academic purposes to evaluate the relationship between actual internet use and intelligence, motivation, and interest. One of the problems in this field has been the use of self-reported measures of internet use, which often means that actual use is underestimated – thus leading to measurement error and unreliable results. To deal with this issue, this study monitored the behaviour of a number of students taking an introductory psychology class at

Michigan State University, which involved logging onto a proxy server when the students used laptops. Of all participants, 83 logged into the proxy server in more than half of the 15 course sessions during the semester and were included in the final report.

Students were also asked to report their use in a survey administered after all lectures were completed. Intelligence was measured by ACT scores and motivation to succeed in class was measured by an online survey after the classes were completed.

The central finding is that non-academic use of the internet in classes was highly prevalent and inversely related to performance in the final exam, regardless of interest in the class, motivation to succeed, and intelligence. In addition, using the Internet for academic purposes during class did not yield a benefit in performance.

The results showed that participants spent a median of 37 minutes per class browsing the internet for non-class-related purposes with their laptops – and "spent the most time using social media, followed by reading e-mail, shopping, watching videos, chatting, reading news, and playing games" – while they spent a total of 4 minutes browsing class-related websites. Internet use was classified to seven categories according to type of site visited and then correlated with final exam score. Although all correlations were negative, only two were statistically significant: those for social media and video sites.

Additionally, the authors compared participants' estimates of their non-academic internet use during a typical class with their actual use. They found that the participants were actually quite accurate in their estimations. Students were also observed using laptops to access college websites to access academic material and to access other 'academic' websites such as Wikipedia – but researchers claim that this type of use was "limited to passive processing of information". Furthermore, they found no association between this type of use and actual classroom learning.

Interestingly, the study also found little support for the oft-repeated claim that more intelligent students are better able to multitask; internet use predicted lower final-exam scores even when ACT scores were included as controls.

Certainly, an important limitation of the study is that internet use is not random, which means that the results may be sensitive to self-selection; unobservable variables may correlate with both internet use and lower achievement. While the inclusion of ACT scores ameliorates this problem, it is not sufficient to conclusively solve it. In addition, the sample size is rather small (84 undergraduate students) and restricted to one university only. It is therefore important not to draw too strong conclusions from this study alone.

Nevertheless, in combination with other papers in this research digest, the study does raise questions about the potentially distracting nature of education technology – and more specifically whether or not students should have unfettered internet access in their learning environments. The participants in this study were college-age students and it's reasonable to assume the impact may be greater for younger students in compulsory education. In some instances, schools are paying significant sums of money for accessing technology tools, such as iPads, in whole classes at the primary and secondary levels, although there are sometimes more stringent measures used to limit the kinds of websites students can access.

Over the last decade, there has been an increasingly loud call from policymakers to educate students for an unknowable future that will require ‘21st Century skills’, which are often closely aligned with technology. It may well be the case that 21st Century solutions to 20th Century problems, such as access to information, might be solved by technology, but the results in this study certainly sound a note of caution for the use of technology for technology’s sake.

The Effect of Adaptive versus Static Practicing on Student Learning - Evidence from a Randomized Field Experiment

- Chris van Klaveren, Sebastiaan Vonkb, and Ilja Cornelisz

Economics of Education Review (June 2017)

Published version

Working paper version (free)

Commentary by: James Croft

Proponents of personalisation in education, placing diverse learners' needs at the heart of the pedagogical process, have long regarded education technology as a key enabler of their reform agenda. Yet just as evidence has increasingly cast doubt on the adequacy of this generalised theory of change, so too researchers have become aware that deploying technology in ways consistent with it to raise performance is more difficult than imagined.

One example of a potentially viable alternative to traditional means of test preparation is computerised adaptive practicing, which has been widely adopted in schools worldwide. Yet its effects on pupil performance have thus far remained unclear. As the authors of this new paper on the subject argue: 'Few randomized controlled experiments with testable outcomes have been conducted.' Research that has rigorously examined impacts for learning 'show no consistent pattern, yielding a mixture of positive, non-significant and even negative results', which they attribute to the fact that the mechanisms through which improvement is meant to take place rarely are made explicit. The authors thus argue that much greater specificity as to the particular uses of different technologies is required.

In their own study, the authors evaluate a recently developed computerised practicing tool made available to students in two different versions: (1) a static version that resembles the traditional approach in which the path of exercises offered is predetermined in terms of topics, question types, and difficulty levels, and (2) an adaptive version that offers a sequence of exercises contingent on a student's performance on these exercises.

Both the treatment and control conditions are analysed in detail in order to facilitate clear interpretation of the mechanisms at work. The essential difference between them lies in the sequence of the exercises offered, with the static pre-determined, and the adaptive taking account of performance. In the latter version, the path is based on students' individual needs and level of readiness.

To evaluate their relative effectiveness for student learning, the authors conducted a randomised field experiment over one year, with 1,021 Dutch secondary-school pupils randomly and blindly assigned to either the static or the adaptive practicing version. Pupil participants, and their teachers, were surveyed to collect information on background characteristics, before the students were randomly assigned within their classes to either the static or the adaptive version, with no knowledge of whether they had been assigned to the treatment or control groups. Pupils chose a subject from a choice of Dutch, biology, history or economics, but had no further control over the content they were exposed to by the programme. Students were taught in instructional blocks of six weeks prior to test administration, and encouraged to practice freely, with a minimum requirement of a quarter of an hour a week. Familiarity with the medium was ensured through regular use of the digital learning environment in class, and exclusively when doing homework.

The authors describe features of the experiment’s design that proof it against attrition bias, behavioural effects, substitution bias, spill-over effects, and problems of test validity. Issues of external validity do arise in connection with pupils’ familiarity with the digital learning environment – which means that the estimated effects may not translate to the average Dutch secondary school. The results may also not be relevant for those interested to apply the treatment for subjects other than those offered in the experiment.

Interestingly, the results show that test scores are not significantly higher for pupils assigned to the adaptive version, as proponents of such practising strategies might expect. In fact, though not statistically significant, the estimates indicate that the adaptive treatment condition is less effective than the static version with more intensive practicing; pupils working with the adaptive version need more time for the same amount of exercises because the difficulty level of these exercises is higher. Accordingly, the number of correct answers given by students in the adaptive group is smaller than those given by students in the static group also. The authors find no average test score effects, but, again contrary to expectations, static practicing does improve test scores by the equivalent of 8 PISA points for higher-ability pupils.

Caution is therefore warranted, the authors argue, when considering introducing adaptive practice software to address individual learning needs: static formative test preparation may in fact be more time-efficient and more likely to generate improved test scores for some learners.

Overall, this also suggests that the potential gains from exploiting education software to achieve greater personalisation of instruction to pupils' individual knowledge levels may be far lower than often is believed. This conclusion contrasts greatly to the Indian experiment analysed in this digest, suggesting that the type and quality of software are crucial for the potential of learning personalisation via technology. Further research on what type of software works and does not work is therefore necessary before implementing large-scale reforms in this respect.

Measuring Conceptual Understanding Using Comparative Judgement

- Marie-Josée Bisson, Camilla Gilmore, Matthew Inglis and Ian Jones

International Journal of Research in Undergraduate Mathematics Education (July 2016)

Published version

Pre-print version (free)

Commentary by: Daisy Christodoulou

Disclaimer: the lead author of the paper under commentary, Ian Jones, is an advisor to No More Marking, an online comparative judgement engine. The author of this digest, Daisy Christodoulou, is shortly to begin working for No More Marking as Director of Education.

Over the last decades, there has been increasing concern in both the US and the UK about the effects of 'teaching to the test'. Researchers in both countries have identified the existence of a 'sawtooth pattern' in the results of national, high stakes assessments such as GCSEs. The sawtooth pattern sees test results rise gradually as schools, teachers, and pupils become familiar with a particular test specification, only for results to fall when a new specification is introduced. This suggests that the initial increase in test results is not being driven by genuine improvements in teaching and learning, but instead by familiarity with the structures and typical patterns of a particular specification. This thesis is borne out in other ways, for example the fact that increases in pass rates in high-stakes examinations are often not matched by increases in similar, but lower-stakes tests. Similarly, when looking at the types of revision materials and advice given to schools and students, it's clear that lots of it is designed with very specific exam structures in mind. For example, in history, many revision guides give over almost a third of their pages to exam-board specific instruction in dealing with the 4-mark, 8-mark and 16-mark question. In maths, one school improvement organisation have produced an approach to revision called 'Beyond the Staples', which identifies the types of questions in the second half of the exam paper (i.e., those beyond the staples of the exam booklet) which are not as difficult as some of the others and should therefore be targeted by low-attaining pupils.

These tactics exploit the fact that large-scale national assessments have to be reliable, which leads to consistency and, to a certain extent, predictability. More innovative and surprising exam tasks would be harder to coach and might be better at teasing out pupils with genuine understanding as opposed to those who have been coached 'beyond the staples'. However, such assessments are extremely hard to mark reliably. Past attempts to use such tasks have ended up creating even worse problems than those outlined above. For example, in maths and science, coursework tasks designed to do this have in fact ended up relying on narrowly written rubrics that constrain a student's approach and encourage stereotyped and gamed responses.

A method of assessment called comparative judgement promises to transcend some of these difficulties. Comparative judgement is not a new idea: it was developed in the 1920s by the psychometrician Louis Thurstone. It relies on the fact that whilst humans are not very good at the kinds of absolute judgements required in traditional marking, they are much better at making comparative judgements, about which of a pair of essays or scripts is the better one. Thurstone's algorithm allows for many such judgements to be aggregated to form a scale. In the last few years, the creation of online comparative judgement engines has allowed for Thurstone's algorithm to be used quickly and efficiently in ways that have not been possible in the past; technology has provided an opportunity to disrupt the marking process to make it more efficient and enable new types of pupil assessment.

In this paper from 2016, a group of professors from the University of Loughborough investigated whether comparative judgement could be used to assess students' conceptual understanding of mathematics. They report on three separate studies: two with undergraduates, and one with year 7 pupils. The undergraduates were given the following two questions:

> *Explain what a p-value is and how it is used to someone who hasn't encountered it before. You can use words, diagrams and examples to make sure you explain everything you know about p-values. Write between half a page and one page.*
>
> *Explain what a derivative is to someone who hasn't encountered it before. Use diagrams, examples and writing to include everything you know about derivatives.*

The year 7 pupils were given the following task:

> *Explain how letters are used in algebra to someone who has never seen them before. Use examples and writing to help you give the best explanation that you can.*

The researchers recruited groups of mathematics PhD students to do the judging. Traditionally, tasks like this would be fiendishly hard to mark reliably. However, using comparative judgement, this was not the case. For all three tasks, reliability – as measured by inter-rater reliability and the Scale Separation Reliability statistic – were high. There was also a moderate correlation with pupils' prior achievement on traditional tests.

From an educator's perspective, the interesting implication is that detailed rubrics or criteria are not required in order to make such a judgement. Instead, judgers rely on their tacit, expert knowledge of the construct in question. As the researchers note: 'the shift from rubrics to a reliance on collective expertise for measuring understanding can be an uncomfortable notion'. Yet, in fact, it is this reliance on collective expertise which is the great strength of such an approach, as it removes the distortions that are often brought about by rubrics. For those who worry about what a nebulous concept such as 'expert judgement' really means, it is important to note that it's measurable. Unlike traditional methods of marking and moderation, which often depend on the personal authority of a chief examiner, the decisions and conclusions reached by a comparative judgement process are distributed and externally verifiable.

The researchers conclude that comparative judgement offers a quick and efficient way of testing conceptual understanding. They compare it with more traditional types of assessment and note that each has its advantages: 'traditional instruments lend themselves well to the fine-grained mapping of misconceptions, whereas comparative judgement is better suited to testing the relative effectiveness of interventions for improving understanding of a given concept more broadly'.

The researchers do not specifically mention the issue of teaching to the test, but given the salience of this issue, and the robustness of the results in the paper, there is clear potential in developing comparative judgement tasks that attempt to assess the same constructs as those in traditional maths papers – but in a way which is much harder to game or subvert. The existence of a suite of low-stakes tasks of this type could provide teachers and pupils with a quick and effective way of teasing out the differences between those pupils who have been coached 'beyond the staples', and those who have acquired understanding beyond the rubric.

CfEE monthly research digest

CfEE produces a monthly Research Digest (roughly 8 issues a year) which is intended to give interested parties a direct view of what the rigorous economic research suggests works in education, both from a macro-policy perspective as well as from the point of view of teachers and head teachers who are looking for more effective classroom strategies. The digest provides abstracts and snippets of such research, with comment and analysis of selected studies that are especially interesting from the point of view of educators and policy-makers.

There is no charge for subscription. If you wish to subscribe, please email us.

£8.99

ISBN 978-1-5136-2373-3

CENTRE FOR EDUCATION ECONOMICS

23 Great Smith Street, Westminster, London, SW1P 3DJ

www.cfee.org.uk
comms@cfee.org.uk