Title registration for a systematic review: Testing frequency and student achievement

Jens Dietrichson, Trine Filges, Julie Kaas Seerup, Bjørn Arleth Viinholt

Submitted to the Coordinating Group of:

☐	Crime and Justice
☒	Education
☐	Disability
☐	International Development
	☐ Nutrition
	☐ Food Security
☐	Social Welfare
☐	Methods
☐	Knowledge Translation and Implementation
☐	Business and Management
☐	Other:

Plans to co-register:

☒	No
☐	Yes ☐ Cochrane ☐ Other
☐	Maybe

Date submitted: 27 November 2018
Date revision submitted:
Publication date: 11 December 2018

Title of the review

Testing frequency and student achievement: a systematic review

Background

The main objective of the review is to examine the effects of interventions where the testing frequency – i.e., how often student achievement is tested during a given period – is changed.

Testing students relatively often may have beneficial effects on achievement and learning as tests have the potential to provide teachers with information about gaps in students' learning. This information could then be used by teachers to provide better feedback to students about what areas they need to work more on (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; McDaniel, Roediger, & McDermott, 2007; Rawson & Dunlosky, 2012). Students who are prone to procrastination may study more when there are frequent examinations focused on small issues (Cid, Cabrera, & Bernatzky, 2017). The information provided by tests may also be used to better adjust the instruction to the students' level, either individually or for the whole class (Black & Wiliam, 2009). Lastly, testing may in itself be a way for students to learn. Several studies argue that when it comes to learning it may in fact be better to perform a test than to repeat the material (Carpenter, 2012; Carpenter, Pashler, & Cepeda, 2009; Glover, 1989; Karpicke & Aue, 2015; Rowland, 2014).

Frequent testing of students may on the other hand increase stress and de-motivate students (e.g., Cheek, Bradley, Reynolds, & Coy, 2002). Another argument against frequent testing is that testing takes time away from instruction (Crooks, 1988; Denny, Paterson, & Feldhusen, 1964; Hancock, 2001; Marso, 1970; National Research Council, 2011). Pushed to the extreme, if students are only tested and never instructed, this would have negative effects on achievement. It therefore seems reasonable that there is a limit where an increased frequency of student testing starts having negative effects on achievement. This frequency is not known.

Policy relevance

There is an ongoing policy debate about what impact testing of students has on their academic achievement (Bergbauer, Hanushek, & Woessmann, 2018; National Research Council, 2011; World Bank, 2017). Testing have become a much debated topic within policy making and educational research in the last decade. With the expansion of standardized testing, and cross-country comparative tests such as Programme for International Student Assessment (PISA), the debate of "how much testing" has been brought to attention.

On one side of the debate you can find educational experts voicing their concerns, and asking policy makers to "slow down the testing juggernaut" (Andrews et al., 2014). This critical stance towards testing is prevalent, according to Adesope, Trevisan & Sundarajan (2017, p.

688): “Indeed, in a review of three policy-oriented journals, Buck, Ritter, Jensen, and Rose (2010) found that 90% of articles were critical of testing. As standardized testing has skyrocketed in recent years, educators may be understandably opposed to more testing”. On the other side of the debate however, we find the World Bank (2017), which recently have argued for more measurement of learning. In their perspective many countries and school systems still have “too little focus on learning – not too much” ((World Bank, 2017, p. 17). Over one third of the countries examined by the World Bank did not have sufficient data to report on levels of reading and math proficiency of children leaving primary school.

According to Bergbauer, Hanushek & Woesmann (2018) the question of whether or not testing is a viable way to improve student learning has led to a confused debate, where “both critics and proponents of international and national testing often fail to differentiate among alternative forms and uses of testing” (Bergbauer et al., 2018, p. 1). They recommend, that the issue of student testing should be considered carefully and with attention to “how assessments are used and what incentives they create” (Bergbauer et al., 2018). To the best of our knowledge no studies that advocate for abandoning tests completely have been published. The interesting question is how often students should be tested and with what type of tests – not whether or not students should be tested at all. This is why we find it relevant to conduct a systematic review of testing frequencies and their potential effects.

Objectives

Our primary research question is: What are the effects of different testing frequencies on student achievement?

The discussion above indicates two additional research questions of interest. Examining them are the secondary objectives of this review:

First, does the type of test matter for the effect sizes? In particular, we are interested in comparing the use of tests with a formative purpose (i.e., to adjust instruction) to tests with a summative purpose (i.e., where students are tested primarily to measure how much has been learnt). Summative assessments are often more comprehensive and, by their nature, more focused on comparing students with each other.

Second, we also want to examine moderators related to the age/grade of students, subject (e.g., math or reading), measurement timing (e.g., spacing between tests), and the at-risk status of students.

Existing reviews

Currently, there are no related title registrations, protocols or reviews in the Campbell Library (database checked 20 November 2018).

The review most relevant to ours is Bangert-Drowns & Kulik (1991), which is 27 years old. The authors conducted a meta-analysis on the effects of frequent classroom testing, based on 35 studies. Their population consisted of students attending secondary school or college. Our proposed review will include primary and secondary school students and will exclude students in higher education. Furthermore, all of the studies included in the review by Bangert-Drowns & Kulik were performed in the United States, whereas we aim to include studies from all over the world.

Two other related (and more recently published) reviews are written by Adesope, Trevisan & Sundarajan (2017) and Phelps (2012). The review by Adesope, Trevisan, & Sundarajan (2017) focused on the testing effect of low-stakes practice tests (which are close to the definition of formative tests). Our review would also include summative tests, such as high stakes standardized tests, as the effects of the increased use of such tests is an interesting policy question. Phelps (2012) included both high-stakes and low-stake tests, but no clearly formulated inclusion criteria can be found in his review article. None of the two reviews reported a fully adequate answer to our primary research question. Adesope, Trevisan, & Sundarajan (2017) reported effect sizes for 'one', and 'two or more' practice tests compared to none. Phelps (2012) reported one effect size where the treatment group is 'tested more frequently than control group'. Neither Adesope, Trevisan & Sundarajan (2017) nor Phelps (2012) have explicitly performed a risk of bias assessment in their reviews. Both reviews only perform one-by-one moderator analyses, which leave several questions unanswered (i.e., the effect size of interventions in primary school is not reported by either review). A large share of studies in Adesope, Trevisan, & Sundarajan (2017, p. 175) use identical practice tests and outcome tests, which may capture only rote learning and is a study design we will exclude. Most studies in Adesope, Trevisan, & Sundarajan (2017) are laboratory experiments which we also do not intend to include in the present review.

Other researchers who reviewed related topics concerning test-enhanced learning are Fuchs & Fuchs (Fuchs & Fuchs, 2001), Rawson & Dunlosky (2012), Karpicke & Grimaldi (2012), McDaniel et al. (2007), Rowland (2014), Black and Wiliam (2009) and Kingston and Nash (2011). Except for Rowland (2014) and Kingston and Nash (2009), none of the aforementioned researchers performed a meta-analysis, and therefore did not answer our primary research question. Rowland (2014) covered the psychological literature on the testing effect and did not focus on educational contexts. Kingston and Nash (2012) focused on formative assessment only and did not analyse the testing frequency.

Intervention

In order to be eligible for inclusion the intervention should manipulate the testing frequency and also provide information about the benchmark testing frequency. That is, it must be possible to extract information about how many tests were performed in a given period in both the treatment and comparison/control group, and the number of tests must differ between these groups. We will only include interventions performed in a school/classroom setting, and exclude laboratory experiments.

Interventions should furthermore only manipulate the testing frequency and not include additional components. If the intervention therefore combines changes in the testing frequency (e.g. by introducing curriculum-based measurement) with another component (e.g. peer-assisted learning strategies) it will not be eligible for inclusion.

Population

Our population will be students attending either primary or secondary school, which in most countries mean kindergarten to the end of high school (grade K to grade 12). It is worth noting that in some countries kindergarten is not a part of primary school but a form of preschool. Interventions with children in kindergarten in countries where kindergarten is not a part of primary school will be excluded, as well as studies performed in preschools. We will exclude interventions in preschool settings because the forms of interventions are difficult to compare to school interventions in this area. Formal tests of achievement are rarely used pedagogical tools in preschool settings, because e.g., the feedback that can be given to preschool children from these tests is limited. Preschool is furthermore almost always voluntary and in many countries only a small share of all children attend, which makes the population different from the one found in primary and secondary school.

Studies performed in higher education, such as universities, will also be excluded. We believe that there are differences between students in grade K to 12 and students attending university regarding, e.g., maturity and attendance of higher education (i.e., attendance is far from universal in higher education). A wide range of studies has been performed about testing frequency at undergraduate level in higher education, but we lack a clear overview of what the effects are in secondary and – even more so – in primary education.

Outcomes

Our primary outcome will focus on tests of academic achievement. The effect of the intervention must be tested using a test that is not identical to the test used during the intervention. Using identical tests may inflate effect sizes due to familiarity and recognition rather than learning (Adesope et al., 2017). We will include both formative and summative tests, low stakes and high stakes tests etc. We will include achievement tests in all academic subjects.

As a secondary outcome measure, we will include tests of socio-emotional outcomes and well-being. Examples of socio-emotional measures are reported levels of self-esteem, self-efficacy, stress, or test anxiety.

Study designs

Included studies should use a treatment-control/treatment-comparison group design where the treatment group is tested with a different frequency than the control/comparison group. Eligible research designs will be randomized field experiments/randomized controlled trials or quasi-experimental designs, where the assignment creates treatment and control/comparison groups.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Andrews, P., Atkinson, L., Ball, S. J., Barber, M., Beckett, L., Berardi, J., ... Zhao, Y. (2014). OECD and Pisa tests are damaging education worldwide. Retrieved November 15, 2018, from https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

Bangert-Drowns, R. L., & Kulik, J. A. (1991). Effects of frequest classroom testing. *Journal of Educational Research*, *85*(2), 85–99.

Bergbauer, A. B., Hanushek, E. A., & Woessmann, L. (2018). *Testing* (NBER Working Paper No. 24836). Cambridge, MA. Retrieved from http://www.nber.org/papers/w24836

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760–771. https://doi.org/10.1002/acp.1507

Cheek, J. R., Bradley, L. J., Reynolds, J., & Coy, D. (2002). An intervention for helping elementary students reduce test anxiety. *Professional School Counseling*, *6*(2), 162–164.

Cid, A., Cabrera, J. M., & Bernatzky, M. (2017). *Frequency of testing. Lessons from a field experiment in higher education.* (MPRA No. 84760). Montevideo. Retrieved from https://mpra.ub.uni-muenchen.de/84760/

Crooks, T. J. (1988). The impact of Eevaluation practices on students. *Review of Educational Research*, *58*(4), 438–481. https://doi.org/10.1207/s15326985ep2202_4

Denny, T., Paterson, J., & Feldhusen, J. (1964). Anxiety and achievement as functions of daily testing. *Journal of Educational Measurement*, *1*(2), 143–147. Retrieved from https://www.jstor.org/stable/1433684

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Fuchs, L. S., & Fuchs, D. (2001). What is scientifically-based research on progress monitoring? *National Center on Student Progress Monitoring*. https://doi.org/10.2320/jinstmet.75.406

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Hancock, D. R. (2001). Effects of test anxiety and evaluative threat on students' achivement and motivation. *The Journal of Educational Research*, *94*(5), 284–290. https://doi.org/10.1080/00220670109598764

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401–418. https://doi.org/10.1007/s10648-012-9202-2

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Marso, R. N. (1970). Classroom testing procedures, test anxiety, and achievement. *Journal of Experimental Education*, *38*(3), 54–58. https://doi.org/10.1080/00220973.1970.11011197

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, *14*(2), 200–206. https://doi.org/10.3758/BF03194052

National Research Council. (2011). *Incentives and test-based accountability in education*. (Committee on Incentives and Test-Based Accountability in Public Education, M. Hout, & S. W. Elliot, Eds.). Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington D.C: National Academies Press. https://doi.org/10.1080/0969594X.2013.877873

Phelps, R. P. (2012). The effect of testing on student achievement, 1910-2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

World Bank. (2017). *World development report 2018: learning to realize education's promise*. Washington DC: The World Bank. https://doi.org/10.1596/978-1-4648-1096-1

Review authors

Lead review author:

Name:	Jens Dietrichson
Title:	PhD in Economics, Senior Researcher
Affiliation:	VIVE – The Danish Center for Social Science
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen
Post code:	1052
Country:	Denmark
Phone:	+4533697797
Email:	jsd@vive.dk

Co-authors:

Name:	Trine Filges
Title:	PhD in Economics, Senior Researcher
Affiliation:	VIVE – The Danish Center for Social Science
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen
Post code:	1052
Country:	Denmark
Phone:	+4533480926
Email:	tif@vive.dk

Name:	Julie Kaas Seerup
Title:	MSc in Sociology, Analyst
Affiliation:	VIVE – The Danish Center for Social Science
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen
Post code:	1052
Country:	Denmark
Phone:	+4533480968
Email:	jkp@vive.dk

Name:	Bjørn Arleth Viinholt
Title:	MLISc, Information Specialist

Affiliation:	VIVE – The Danish Center for Social Science
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen
Post code:	1052
Country:	Denmark
Phone:	+4533480862
Email:	bcn@vive.dk

Roles and responsibilities

The roles and responsibilities in this review will be as follows:

- **Content:** Jens Dietrichson will be in charge of the review content. Jens has both practical and research experience with education interventions.
- **Systematic review methods:** Jens Dietrichson, Trine Filges and Julie Kaas Seerup are experienced in conducting systematic reviews. All members in the review team have experience with Campbell review methods.
- **Statistical analysis:** Both Jens Dietrichson and Trine Filges has conducted statistical analyses on other Campbell reviews, and will be in charge of the statistical meta-analysis.
- **Information retrieval:** Bjørn Arleth Viinholt is information specialist at VIVE and has experience in conducting systematic information retrieval for Campbell reviews. Bjørn Arleth Viinholt will be in charge of conducting information retrieval for this review.

Funding

Funding is received from VIVE – The Danish Center for Social Science Research.

Potential conflicts of interest

None of the review authors have conflicts of interest related to this review.

Preliminary timeframe

- Date you plan to submit a draft protocol: Within 6 months after title registration.
- Date you plan to submit a draft review: Within 18 months after protocol approval.