International Journal of Science and Mathematics Education (2022) 20:1411–1429
https://doi.org/10.1007/s10763-021-10207-9

Implementing Confidence Assessment in Low-Stakes, Formative Mathematics Assessments

Colin Foster[1]

Received: 25 November 2020 / Accepted: 20 July 2021 / Published online: 3 September 2021

Abstract

Confidence assessment (CA) involves students stating alongside each of their answers a confidence rating (e.g. 0 low to 10 high) to express how certain they are that their answer is correct. Each student's score is calculated as the sum of the confidence ratings on the items that they answered correctly, minus the sum of the confidence ratings on the items that they answered incorrectly; this scoring system is designed to incentivize students to give truthful confidence ratings. Previous research found that secondary-school mathematics students readily understood the negative-marking feature of a CA instrument used during one lesson, and that they were generally positive about the CA approach. This paper reports on a quasi-experimental trial of CA in four secondary-school mathematics lessons (N = 475 students) across time periods ranging from 3 weeks up to one academic year, compared to business-as-usual controls. A meta-analysis of the effect sizes across the four schools gave an aggregated Cohen's d of –0.02 [95% CI –0.22, 0.19] and an overall Bayes Factor B_{01} of 8.48. This indicated substantial evidence for the null hypothesis that there was no difference between the attainment gains of the intervention group and the control group, relative to the alternative hypothesis that the gains were different. I conclude that incorporating confidence assessment into low-stakes classroom mathematics formative assessments does not appear to be detrimental to students' attainment, and I suggest reasons why a clear positive outcome was not obtained.

Keywords Confidence assessment · Formative assessment · Low-stakes assessments · Mathematics education · School mathematics

✉ Colin Foster
c.foster@lboro.ac.uk

[1] Mathematics Education Centre, Schofield Building, Loughborough University, Loughborough LE11 3TU, UK

Introduction

Confidence assessment (CA) is a pedagogical practice involving a modification to the usual ways of conducting low-stakes (i.e. not-for-credit) formative assessments during school mathematics lessons. Students are asked to state a confidence rating (e.g. 0 low to 10 high) alongside each of their answers to express how certain they are that each answer is correct (Foster, 2016). Each student's score is then calculated as the sum of the confidence ratings for the items that they answered correctly, minus the sum of the confidence ratings for the items that they answered incorrectly.[1] The purpose of this scoring system is to incentivize students to give confidence ratings that are as truthful as possible. In the long term, students cannot systematically 'game' CA scores by consistently over- or under-stating their true confidence levels, so CA provides the possibility of accessing students' genuine beliefs about their degree of confidence at the level of individual items on an assessment. Students' *calibration* refers to the correlation between their confidence rating and their mean facility (the mean number of questions that they answered correctly) (Fischhoff et al., 1977), and it might be expected that students using CA over an extended period of time would gradually become better calibrated.

Previous research testing a simple CA instrument over the course of one mathematics lesson on the topic of directed (positive and negative) numbers (Foster, 2016) found that secondary-school students were generally well calibrated in this topic, giving higher confidence scores on average for items that they answered correctly. Most students also readily understood the negative-marking aspect of CA, and were positive about the CA approach, alleviating concerns that students would find such an unfamiliar approach alien and unacceptable. Since then, CA has gained prominence amongst teachers (e.g. Foster et al., 2021), and, anecdotally, at teacher conferences and on social media, mathematics teachers tend to express considerable enthusiasm for the idea of CA, and it has been described enthusiastically on teacher-oriented podcasts (e.g. Barton, 2019; see also Baker, 2019).

However, the extent to which CA 'works', in the sense of improving students' mathematics attainment in *summative* assessments when it is used over an extended period of time in *formative* assessments, is not known. Here, by 'summative' I mean an assessment which is used primarily as a retrospective evaluation of what has been learned up to that point, whereas a 'formative' assessment is integral to the teaching process and is "used to make changes to what would have happened in the absence of such information" (Wiliam, 2006, p. 284; Wiliam, 2017).

Clearly, students' mathematics attainment, as measured by general summative assessments, will depend on myriad factors, but literature from the use of CA in higher education (e.g. Gardner-Medwin, 1995, 1998, 2006, 2019; Gardner-Medwin & Gahan, 2003; Gardner-Medwin & Curtin, 2007) suggests that it could have potential in school mathematics. Helping school students to consider how sure they are of the answers that they give might encourage them to self-check and to develop higher levels of self-awareness, which could enable them to target areas of weakness more effectively, which could increase their overall mathematics attainment. It could also be that being better calibrated is a desirable educational goal in its own right, since "knowing what

[1] Other scoring rules and tariff matrices are possible; for simplicity, this was the one used in this research.

you know and what you do not know", and therefore what you do or do not need to look up or seek support with, is an essential part of becoming a more educated person.

One of the advantages of CA is that it is easy to implement, as it does not require redesigning assessment instruments (Barton, 2019; Foster, 2016, Foster et al., 2021). Any classroom formative assessment method in which students write their answers (on paper, or even on mini-whiteboards [see McCrea, 2019]) can easily be modified by asking the students to write a confidence rating from 0 (low) to 10 (high) alongside each answer to indicate how sure they are that they are correct. This suggests that it could be possible to conduct a 'naturalistic' trial of CA in schools that minimally interferes with schools' existing assessment processes. If successful, such an intervention would be a low/no-cost 'easy win' for schools to implement (see Wiliam, 2018). Such an approach to trialling increases ecological validity (Robson & McCartan, 2015) and respects schools' autonomy and teachers' professionalism (Newmark, 2019), since the researcher does not attempt to take over and impose new systems and materials from the outside, without reference to the context of the students and teacher, and this may also increase compliance. Additionally, since it is easy for schools to agree to participate in such a trial, a closer-to-random sample of schools could likely be obtained, with fewer refusals than with standard educational trials. CA would seem to provide an ideal opportunity to test out such a 'naturalistic' trial.

Confidence Assessment in Mathematics

Confidence of Response

There are many similar and overlapping constructs in the literature relating to confidence at the fine grain size of individual items on an assessment (see e.g. Clarkson et al., 2017; Dirkzwager, 2003; Marsh et al., 2019; Stankov et al., 2012). For the purposes of CA, a pupil's "confidence of response" may be defined as "how certain they are that the answer that they have just given is correct" (Foster, 2016, p. 274), and this may be represented on a scale from 0 (completely uncertain; i.e., just guessing) to 10 (absolutely certain). Since students' scores are calculated by *summing* these ratings (positively for correct answers and negatively for incorrect answers), it may be reasonable to treat this as a linear scale. This is plausible if we suppose that students are seeking to maximise their total score, leading to, for a 10-item formative assessment, a range of possible scores from –100 to 100. It may seem over-optimistic to expect students to discriminate their confidence on a 10-point scale (Preston & Colman, 2000); however, a 0–10 scale is likely to be considerably easier for students to use to calculate and interpret their scores in the classroom, since it is easy to calculate that for n questions the highest possible obtainable score will be $10n$, and this may be considered to be what their score is 'out of'.

Uses of Confidence Assessment

CA has been used in higher education, particularly in medicine and related disciplines where it is critically important to discourage guessing in life-and-death matters (Gardner-Medwin, 1995, 1998, 2006, 2019; Schoendorfer & Emmett, 2012).

University students are often found to be poorly calibrated and to tend towards overconfidence (Ehrlinger et al., 2008). However, with repeated use of CA over time, calibration tends to improve (Gardner-Medwin & Curtin, 2007). It has been found that CA, often implemented in a multiple-choice context, can encourage self-checking, self-explanation and higher-level reasoning (Gardner-Medwin & Curtin, 2007; Sparck et al., 2016), and improve test validity by reducing gender bias (Hassmén & Hunt, 1994). Consequently, it has been suggested (Foster, 2016, 2017; Foster et al., 2021) that CA might have considerable potential benefits for formative assessment in the school mathematics classroom. Although secondary-school students are younger and more diverse than university students, particularly those studying medicine and related areas, previous research (Foster, 2016; Foster et al., 2021) showed that they are capable of making valid judgments about their levels of confidence in a confidence assessment. So, it seems plausible that repeated use of CA over a period of time could have similar effects with secondary students to those reported for university students (Gardner-Medwin & Curtin, 2007).

The potential benefits of incorporating confidence assessment into low-stakes formative assessments in mathematics include:

1. *Discouraging guessing*, which adds noise to formative assessments and trivialises the learning of the subject (Foster, 2017).
2. *Reducing over-confidence*, which may inhibit students, through complacency, from learning and improving, and may lead them to embed errors and misconceptions without correcting them, thus hampering their future development in the subject.
3. *Reducing under-confidence*, which may prevent students from gaining as much satisfaction from the subject as they otherwise would, perhaps leading to lower motivation and a disinclination to pursue mathematics beyond the compulsory school years. Under-confidence may also trap students within repetitive cycles of practising content that is already secure, thus keeping them from accessing more demanding learning material.

It is important to stress that the pedagogical aim of using CA is not simply to raise all students' confidence, which would be highly undesirable. The aim is to encourage *appropriate* levels of self-confidence and self-awareness to help students engage in more effective future learning, since, "for secure development of procedural fluency it is important not only that a pupil can obtain the correct answer in a reasonable amount of time but that they have an accurate sense of their reliability with the procedure" (Foster, 2016, p. 272). This means that, in some circumstances, the intention of CA would be to *reduce* students' confidence, at least temporarily, to help them gain a clearer picture of their weaknesses and difficulties, with a view to more targeted and effective subsequent learning. Many authors have drawn attention to *illusory superiority effects*, such as the *Dunning–Kruger effect* (see Ehrlinger et al., 2008), in which someone with low ability at a task overestimates their ability because they 'do not know what they do not know'. Common quotes to this effect within popular culture include "It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so" (Mark Twain). CA has a potential role to play in recalibrating students to a more accurate assessment of what they do and do not currently know.

In theory, we would expect the same CA intervention to improve the calibration of both under-confident and over-confident students. In this way, it should address the fact that students have different personalities: some may be more optimistic and others more pessimistic in their general outlook. Whilst this may affect their scoring in the short term, the intention is that over time all students would become better calibrated.

The Hypercorrection Effect

Another potential benefit of CA would be harnessing the *hypercorrection effect*, which is the observation that errors made with high confidence are more easily corrected than are errors made with low confidence (Butterfield & Metcalfe, 2001). This effect is surprising, since it might be expected that when a student is incorrect but very sure of their incorrect answer this might indicate a firmly-held belief, which would be difficult to dislodge. However, the effect has been reliably demonstrated in studies involving people of varied ages and in a variety of contexts (e.g. Metcalfe & Finn, 2012; van Loon et al., 2015), including in an authentic mathematics learning situation (Foster et al., 2021). The hypercorrection effect has been attributed to several possible mechanisms, including the memorable nature of a shock or surprise (Butterfield & Metcalfe, 2006). To see benefits from the hypercorrection effect in the mathematics classroom, it may be necessary that students *reflect* on the fact that they are confident, which CA provides an opportunity for them to do.

Research Question

Foster (2016) found that secondary school students were overwhelmingly positive about the CA process, although, as this study was conducted over a single lesson, this may be at least partly a result of novelty effects (see Clark, 1983), and it was not possible to test for any long-term benefits over, say, time periods of up to a school year. Consequently, the research question for this study was: *Does repeated use of CA in formative assessments in school mathematics lessons over an extended time period result in improvements in students' mathematics attainment in summative assessments?*

The answer could be 'no' if the use of CA is distracting for students or the teacher, or takes valuable time away from teaching input, if it makes students become unduly risk averse (Kahneman & Tversky, 1984), or if students find the CA process discouraging, because their CA score is lower than the simple total of their correct answers. Less confident students may be repeatedly reminded of their low confidence by the CA method, and this could impair their performance on the questions. It may also be that it is unreasonable to expect CA to have a measurable impact on a distant outcome, such as overall mathematics attainment in a summative assessment. The CA process could also be contrary to equity if, as is plausible, high-attaining students are more likely to be better judges of their level of knowledge, and therefore more able to benefit (Ben-Simon et al., 1997). It may also be that teacher professional development is necessary in order to see students' attainment improve measurably from CA. Alternatively, the hope is that CA could be 'low-hanging fruit', which can offer schools an 'easy win' to benefit their students at little or no opportunity cost (*cf* Wiliam, 2018).

Method

A 'naturalistic' quasi-experimental trial (Christensen et al., 2015) was used, in which summative assessment data was obtained from schools whilst they incorporated CA into their regular in-class formative assessments. The intervention was minimally invasive, and aimed to preserve most features of schools' typical classroom and assessment practices. Rather than providing schools with researcher-designed CA tools and instruments, or requesting that schools develop their own, schools were instead asked to continue to use whatever assessment systems they had currently in place, both for formative assessment and for summative assessment. The only requested change was a simple modification to schools' formative assessment procedures, which did not require the re-design of any of the materials, in which students were asked to write a confidence rating next to each of their answers as they completed them during low-stakes, in-class tests. This small change to normal practice could be easily accommodated within lessons, constituting a minimal intrusion into schools' existing routines. This had the advantage of being easy for schools to commit to doing, as well as preserving as much ecological validity (Robson & McCartan, 2015) as possible. All teachers contacted agreed that the intervention would be easy to implement in their schools. For example, one head of department commented, "Students won't need any extra time to write in a confidence level and it'll be interesting for the teachers."

Intervention

Schools were asked to identify groups of "parallel classes" (as determined by them) of similar age and attainment, and to assign one of these to a business-as-usual control, where students would continue to experience the usual formative assessment practices currently operating in the school, without any modification. Schools were asked to modify the experience of the other, parallel, groups in only one way: for their regular formative, low-stakes in-class tests, schools were asked to continue to use their existing assessments but to ask students to also write a confidence rating (0 low to 10 high) beside each of their answers to indicate how sure they were that their given answer was correct (Foster, 2016). On completing their formative assessments, students were asked to calculate their scores by adding up the confidence ratings for the questions that they answered correctly and subtracting from this the total of the confidence ratings for the questions that they answered incorrectly. The scoring process, and the students' involvement in this, was an important aspect of the intervention, in order to incentivize truthful confidence ratings, by rewarding accurate confidence and penalising students for being over- or under-confident in their answers (see Foster, 2016). It also maintained the 'low-stakes' aspect of existing classroom culture in relation to in-class assessments, in which the teacher does not formally collect in the tests, mark them and record the results.

The outcome measure was to compare students' marks in their normal termly or half-termly summative assessments, before and after a period of time in which CA was used in the students' low-stakes in-class formative assessments. (None of the schools' summative assessments contained any CA, and these were not modified in any way for either condition in this study.) Because schools used different summative assessments (often produced in-house) and administered these at different frequencies and times

during the school year, data points were not synchronised between schools, or directly comparable between schools. Schools provided data on each student from the most-recent summative assessment point before embarking on CA, and this (converted to a percentage) was used as the pre-test score. They also provided data from the most recent summative assessment point before conclusion of the trial period, and this (also converted to a percentage) was used as the post-test score. (All schools were intending to continue using CA beyond the timescale of the trial, so CA was still in use at the conclusion of the study.)

The intention was to test whether using CA in formative assessments over an extended period of time (half a term or so) might raise students' general mathematics attainment, as measured by their normal (non-CA) summative assessments, by comparing students' pre-test to post-test gain scores between students in the two conditions (confidence assessment and business-as-usual control). The CA formative assessments were the intervention, but not the data sources; the pre- and post- test summative scores were the data sources, and these were ordinary, non-CA summative assessments.

Participants

Emails were sent to schools across several mathematics teacher networks in the UK, and an invitation was posted on *Twitter* and widely retweeted by several contacts with large teacher followings:

> Now recruiting schools to trial 'confidence assessment', where students put a rating next to their answers to say how sure they are that their answer is right. Lots of potential benefits – see https://tinyurl.com/ydb2lhjx Please email me if you could help.

Following this, 55 schools contacted me, and I replied to each, enclosing a 1-page explanation of what I was requesting (available in full in Appendix A). All of these teachers replied positively, saying that they wanted to try out the technique, but not all were able to take part fully in the trial. The most commonly-stated reasons for this were:

(1) no suitable parallel classes: often classes of similarly-attaining students were available, but they were not taught by teachers deemed to be comparable (e.g. one was a mathematics specialist and the other was not) or the school was too small to have parallel classes;
(2) anticipated or actual lack of continuity in some classes during the period of the trial (e.g. extended periods of teacher absence or teachers leaving the school);
(3) the teacher who was taking a lead on CA leaving the school or obtaining an unforeseen promotion to a more senior role;
(4) a desire to implement CA with *entire* year groups, and not a subset of a year group.

The last-mentioned reason (4) was the most common. Although schools appreciated that simple controlled experiments (i.e. A/B tests) would be the only way to directly

measure the effectiveness of CA, some had ethical concerns about denying students an intervention that they felt sure would be beneficial, and a small number felt that "experimenting on children" was morally wrong and did not wish to "use children as guinea pigs". In some cases, these schools rolled out CA to all of their mathematics students in a particular year group. In other schools, the reason for wanting to implement with full year groups was more practical, relating to efficiency, equity and simplicity of provision, and communication of assessment systems and outcomes to parents.

In the end, complete data was obtained from four co-educational secondary schools in England, who participated fully in the trial, and this involved a total of 475 students over time periods ranging from 3 weeks to an entire school year (see Table 1 for details of the participating schools). A total of 11 students were not present at both the immediately-preceding summative assessment (the pre-test) and the summative assessment following the trial (the posttest), and data from these students were excluded from the analysis (see Table 1). All test marks were converted to percentages for analysis.

Analysis

Some schools provided data on considerably more participants in the business-as-usual control classes than in the intervention classes. This was because schools C and D were trialling CA with only one class each, but each school had two other classes in the same year group who were not trialling the CA, and it was convenient for the schools to provide data for all of the cohort together. In School A, two classes trialled CA, and control data was provided for a large number of students from all of the other parallel classes. Consequently, the opportunity presented itself to create a dataset matched by pre-test scores, in order to compare more accurately the gain scores for students beginning at a similar attainment level at pre-test.

Consequently, a matched sample (Table 2) for each of schools A, C, and D was produced, using the *MatchIt* package in *R* (Ho et al., 2011) and using the "nearest-neighbour" method. (*R* code for all of the analyses described in this paper is available in

Table 1 Schools participating in the study

School	School description	Year group[1]	Duration of trial[2]	*N* control	*N* intervention
A	comprehensive secondary school	7	3 weeks	111 (+ 5 excluded)	55
B	independent day school	7	1 term	57 (+ 1 excluded)	77 (+ 3 excluded)
C	comprehensive secondary school	8	2 terms	59	28 (+ 1 excluded)
D	comprehensive secondary school	7	3 terms	56	32 (+ 1 excluded)

(Students were excluded because they did not complete either the pre-test or the post-test.)

[1] Year 7 is aged 11–12; Year 8 is aged 12–13

[2] There are 3 terms in a school year and each term lasts about 13 weeks in total

Appendix B.) Matching was *not* used for school B, because there were fewer control students than intervention students. Distribution plots for raw and matched samples (Appendix C) show that the matching process worked well for schools A and C and satisfactorily for D.

Ethical approval was obtained from the University of Leicester Ethics and Integrity Committee, and the full and matched datasets are freely available at https://doi.org/10.6084/m9.figshare.15027966.v1.

Results

Pre-test and post-test scores were all converted to percentages, and gain scores were calculated for each student as (post-test score – pre-test score) (Table 3). Analysis was conducted for each school using Welch's independent samples *t* tests on matched datasets for schools A, C, and D, and the original dataset for school B, where matching was not carried out. In schools A and B, students improved more in the intervention condition than in the control condition, but in schools C and D the reverse was the case, as shown by the sign of the *t* values and Cohen's *d* effect sizes in Table 4. Separate *t* tests for each school revealed that none of the effect sizes was significantly different from zero, providing no evidence for any effect of the CA intervention.

Bayesian *t* tests were also conducted on the gain scores, comparing the fit of the data under the null hypothesis (that the gain scores are equal under the CA and control conditions) and the alternative hypothesis (that the gain scores are *not* equal under the two conditions). A Bayes factor *B* indicates the relative strength of evidence for two hypotheses (Dienes, 2014; Lambert, 2018; Rouder et al., 2009), and the interpretation is that the data are *B* times as likely under the null hypothesis as they are under the alternative hypothesis. (For a recent example of a similar use of Bayes factors, including a brief explanation of their meaning and interpretation, please see Foster, 2018.)

Table 2 Original and matched datasets. Empty lines under 'matched dataset' indicate where matching was not used. Intervention datasets were not altered

		Original dataset			Matched dataset		
School	Condition	*N*	pre-test Mean	pre-test *SD*	*N*	pre-test Mean	pre-test *SD*
A	control	111	61.17	18.60	55	59.91	17.97
	intervention	55	59.49	18.04			
B	control	57	76.35	14.45			
	intervention	77	74.35	14.33			
C	control	59	50.94	16.01	28	55.78	14.88
	intervention	28	56.69	15.25			
D	control	56	62.61	9.81	32	68.75	6.41
	intervention	32	75.78	11.22			

Table 3 Pre-, post- and gain scores for each school

School	Condition	*N*	pre-test Mean	pre-test *SD*	post-test Mean	post-test *SD*	Gain Mean	Gain *SD*
A	control	55	59.91	17.97	45.06	19.07	–17.15	11.33
	intervention	55	59.48	18.04	43.15	19.72	–16.33	10.99
B	control	57	76.35	14.45	76.18	13.55	–0.18	14.08
	intervention	77	74.35	14.33	74.60	15.83	0.25	12.04
C	control	28	55.78	14.88	50.15	12.70	–1.44	10.13
	intervention	28	56.69	15.25	53.98	12.89	–2.71	9.36
D	control	32	68.75	6.41	81.60	8.99	14.97	8.35
	intervention	32	75.78	11.22	89.06	6.22	13.28	10.75

Bayes factors B_{01} in favour of the null hypothesis of equal gains under the two conditions, relative to the alternative hypothesis of unequal gains, were calculated using the default settings in *JASP* (JASP Team, 2020), with a Cauchy prior width of .707 (see Table 4). All of the estimated Bayes factors fell in the 3–10 range described by Jeffreys (1961) as providing "substantial" evidence for the null hypothesis of no difference between the classes doing CA and the control classes. Note that this is not the same as the *inconclusive* result obtained from *p* values greater than .05 in the frequentist *t* tests, where we cannot conclude that either group outperformed the other. The Bayesian result provides *positive evidence of no difference*, not merely lack of evidence of a difference (see Dienes, 2014; Lambert, 2018).

Standardised effect sizes from each school (Table 4) were combined, using both frequentist and Bayesian meta-analysis. The reason for treating each school as a separate study was that each was using different summative assessment measures, and pooling the raw data would not have been meaningful. Frequentist meta-analysis of the four effect sizes, using the *metafor* package in *R* (Viechtbauer, 2010) and the random-effects model ($k = 4$; τ^2 estimator: restricted maximum likelihood), gave an overall effect size not significantly different from zero ($d = –0.02$ [95% CI –0.22, 0.19], $p = .870$ with $Q(3) = 0.878$, $p = .831$) (see Fig. 1). Note that, in the conventional interpretation of frequentist hypothesis testing, it is not possible to conclude from this that the intervention had no effect, only that it was not possible to detect any effect.

Table 4 Analysis of gain scores for each school. Note: A matched sample was not used for school B. Welch independent-sample *t* tests were used

School	*N* control	*N* intervention	*t*	*df*	*p*	*d*	95% CI for *d*		Bayes Factor B_{01}
A	55	55	0.384	107.90	.701	0.073	–0.301	0.447	4.63
B	57	77	0.182	109.41	.856	0.032	–0.310	0.375	5.26
C	28	28	–0.485	53.665	.630	–0.130	–0.653	0.395	3.36
D	32	32	–0.702	58.426	.486	–0.175	–0.666	0.316	3.18

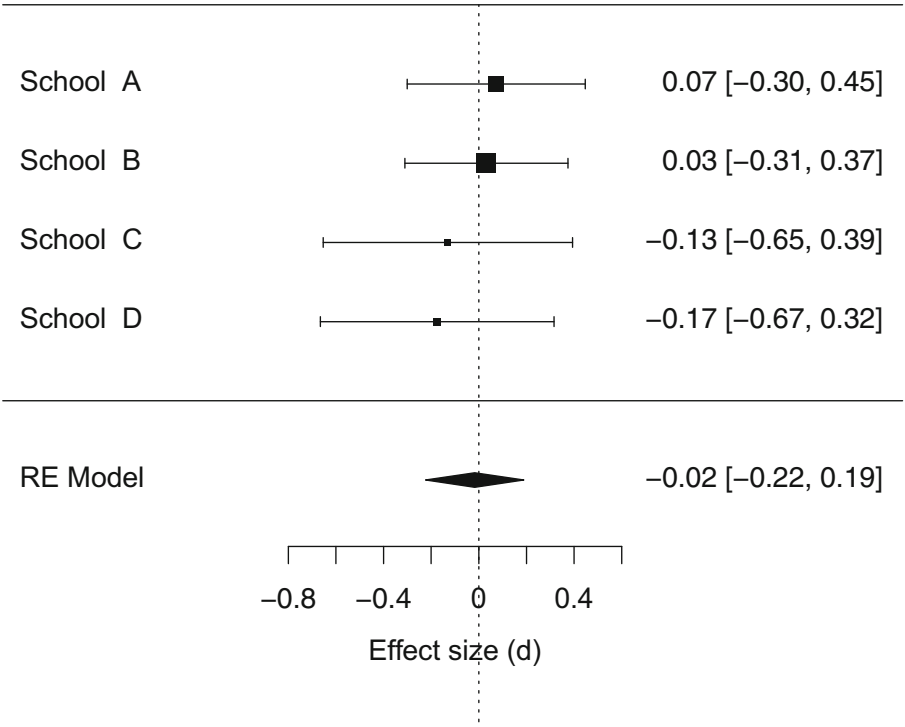


Fig. 1. Forest plot of Cohen's *d* effect sizes for each school. Bars indicate 95% confidence intervals

Consequently, Bayesian meta-analysis was conducted using the *BayesFactor* package in *R* (Morey et al., 2015). This gave an overall Bayes Factor B_{01} in favour of the null hypothesis of 8.48, meaning that there was "substantial" (Jeffreys, 1961) evidence for the null hypothesis that the intervention group improved the same amount as the control group, relative to the alternative hypothesis that it did not. Note again that this is *not* the same as an inconclusive result, where we cannot say whether one group did significantly better than the other, such as was obtained from the frequentist meta-analysis. The Bayesian meta-analysis indicates positive evidence that the two groups did equally well.

Schools were not explicitly asked to comment on their experiences of using CA with some of their classes. However, a few schools did comment when providing the data, and this was invariably positive. For example, one school noted, "The classes enjoyed doing the confidence scores on the Low Stake Quizzes and we found it useful to identify categories of student in the group and so how to help them: Confident and correct, Confident but incorrect, Not confident but correct & Not confident and incorrect." Another stated that "[the students] seemed intrigued by the scoring system so I'm hopeful it may have an impact on their self-reflection and thus on the follow up consolidation work they do … They are very keen to not get any negatives so may be having the desired effect." No schools reported any problems implementing CA with their classes.

Discussion

The research question for this study was: *Does repeated use of CA in formative assessments in school mathematics lessons over an extended time period result in improvements in students' mathematics attainment in summative assessments?* To answer this, a 'naturalistic' quasi-experimental trial of CA involving four secondary schools ($N = 475$ students) was conducted over time periods ranging from 3 weeks up to one academic year, with students experiencing CA compared to students in business-as-usual control groups. A frequentist meta-analysis of the effect sizes across the four schools (Cohen's $d = –0.02$ [95% CI –0.22, 0.19]) is consistent with positive or negative effect sizes of small or moderately important size, and so, by itself, is inconclusive (see Lortie-Forgues & Inglis, 2019, for discussion of this phenomenon). However, a Bayesian meta-analysis (Bayes Factor B_{01} of 8.48) revealed "substantial" (Jeffreys, 1961) evidence for the null hypothesis that there was no difference between the attainment gain of the intervention group and that of the control group, relative to the alternative hypothesis that the gains were different. This means that we can conclude that CA was not detrimental to students' attainment, but neither was it beneficial.

There were reasons to think that CA might have turned out to have been detrimental to students' mathematics learning, through distracting them or their teachers from more effective aspects of the lesson and taking time away from teaching, but we did not find any evidence for this opportunity cost (see Wiliam, 2018). Similarly, CA could have had an adverse effect if it had led to students becoming damagingly risk averse (Kahneman & Tversky, 1984) and thereby leaving questions unanswered and supplying a zero confidence rating for them. Additionally, the requirement to think about confidence scores could have disturbed students' flow when answering the questions and distracted them from their learning. It might also have been anticipated that the students could have found their confidence scores discouraging, as these are likely to be lower than a simple total of questions answered correctly, and this could have led to disengagement from learning, but again this does not appear to have happened to any measurable degree.

There were also good reasons for supposing that CA might have had a beneficial effect on students' mathematics attainment – indeed, this was the intention motivating this programme of research, and is frequently expressed enthusiastically by teachers (e.g. Baker, 2019; Barton, 2019). However, again, this does not appear to have happened. It may be that the attainment measure was too distant from the intervention, and that it was too optimistic to expect that CA used, from time to time, in formative assessments in mathematics lessons would lead to a detectable effect in termly/half-termly summative assessments, when myriad other factors are likely to be important to students' overall attainment. It may also be that without a strong imperative from the senior leadership in the school backing CA there was a lack of focus and the teachers' agendas were dominated by other promoted strategies and approaches that were given a higher status within the school. This is potentially one of the disadvantages of the 'naturalistic' nature of this trial, in which minimal disturbance is made to the school system. In this scenario, we are unlikely to benefit from Hawthorne-like effects (Robson & McCartan, 2015) and the general enthusiasm deriving from a concentrated focus on a high-profile intervention trial.

Another factor may be that some of the schools may have been implementing CA infrequently and without sufficiently careful attention. The constraints within the naturalistic nature of this study meant that there was no opportunity for monitoring the fidelity of the implementation in a process evaluation, and it may be that heads of department were overestimating the compliance with CA within their schools when reporting back to me. It is also conceivable that there was 'leakage' from the intervention condition to the control condition, with teachers talking about CA in the staffroom, leading to teachers of control classes also trying it. This seems unlikely, however, given that heads of department were clear about the purpose of the trial, had ownership of it, and believed that their colleagues were keen to support this.

The research question for this study refers to "repeated use of CA in formative assessments" and "over an extended time period", and neither of these terms can be defined precisely, due to the naturalist nature of this trial. This means that it is possible that more intensive use of CA, such as every lesson, rather than once a week, might be needed in order to see a measurable improvement in attainment in summative assessments. Alternatively, or additionally, it may be that the "extended time period" necessary is greater than the 3 weeks up to one academic year used in this study, although an intervention which shows no effect even after a year is probably of little practical value to schools.

A clear overriding factor may be that teacher professional development is likely to be necessary in order for students to benefit significantly from CA. We know that, in general, for teaching strategies to be implemented effectively teachers need time to think through the pedagogical rationale and discuss approaches that they will use in the classroom (Joubert & Sutherland, 2009; Timperley et al., 2008; Yoon et al., 2007). In the present study, professional development was completely absent, and the only guidance to schools was a single sheet of paper outlining the approach (the entirety of this is presented in Appendix B). It is highly plausible that important features of CA might consequently have been 'lost in translation' or that teachers might not have sufficiently 'bought in' to the strategy.

Conclusion

A 'naturalistic' quasi-experimental trial of CA use within regular, low-stakes mathematics formative assessments across four secondary schools (N = 475 students) over time periods ranging from 3 weeks up to one academic year was conducted. A Bayesian meta-analysis of the effect sizes revealed substantial evidence of no effect on students' overall mathematics attainment, meaning that CA was not detrimental to students' attainment, but neither was it beneficial. Like a number of other high-profile interventions (see Lortie-Forgues & Inglis, 2019), CA, at least in this simple format, does not appear to be a quick, easy win for schools. To see benefits of CA, a closer-to-intervention measure may be needed and/or a more comprehensive implementation package, involving at least some professional development for teachers to set out the rationale for the process and generate some commitment. The sample of schools for this study was self-selecting, and consequently at least the lead teacher at each school was likely to be enthusiastic about the idea of CA. This might have been expected to have led to a stronger effect than if CA were implemented across 'typical' schools, but it may

be that, without some professional development, enthusiasm by itself is insufficient to lead to effective implementation. Alternatively, it may simply be the case that use of CA does not in fact raise student attainment, and further studies would be needed to investigate these different possibilities.

A novel, 'naturalistic' minimal-intervention approach to trialling was employed in this study, with mixed success. Some schools found it easy to find 'parallel' classes and to set up alternative conditions for the students, and using schools' existing data from summative assessments was unproblematic and very low-cost. All schools found conducting such a trial minimally invasive and fully compatible with their normal working processes. However, some schools were unwilling or unable to offer the intervention to a subset of the students, and in some cases this was because of uneasiness over the quasi-experimental methodology, as has been reported previously (Meyer et al., 2019).

It is important that the educational research literature reports null findings, to combat the file-drawer problem of bias in the literature (Chambers, 2017), and this would seem to be particularly relevant for pedagogical approaches such as CA that currently command increasing attention in teacher-oriented literature. In future research, I plan to collect further data from schools who are persisting with CA and examine whether students' *calibration* (in addition to attainment) improves over extended periods of CA use. I also plan to interview students and teachers in schools who are persisting with CA to try to understand their perspectives on the approach. I also intend to develop associated professional development materials to support use of CA and to test the effectiveness of these. The ease with which CA can be adopted in practice, with minimal disturbance to existing classroom routines, makes it seem an attractive option to teachers, and it addresses the important issue of students' confidence. However, how to make it effective in practice remains an unsolved problem.

Note

The full dataset is available at https://doi.org/10.6084/m9.figshare.15027966.v1.

A. Appendices

A: Instructions Sent to Schools

Confidence Assessment in Mathematics: School-Based Trial

Confidence Assessment

The idea is for students to give a confidence rating on a scale of 0 to 10 (0 is just guessing; 10 is absolutely certain) alongside their answers in their school mathematics assessments. Then, when they mark it, instead of simply counting the number of correct answers, they calculate their mark as the sum of the confidence ratings for the questions they got right ***minus*** the sum of the confidence ratings for the questions they got wrong.

It shouldn't be possible to 'game' this system, as it rewards accurate assessment of confidence and penalises both over-confidence and under-confidence.

There are potentially four benefits of building this into formative assessments:

- It improves students' calibration—being more confident about the ones they get right and less confident about the ones they get wrong—so that they 'know what they know and what they don't know' better, enabling better metacognition, more targeted revision, and more secure future learning;
- It promotes self-checking—students often correct their answer when asked how sure they are;
- It discourages guessing, which adds 'noise' to formative assessments;
- It capitalises on the 'hypercorrection effect'—if a student states that they are very sure about something, and then they discover that they are wrong, they remember the correct answer better.

These benefits are plausible, and there is small-scale and anecdotal evidence for them, but we don't know if/how these things will work out in practice over time in real schools.

What I would like you to do

I would like you to trial confidence assessment with some classes across a year group. The only way to get convincing evidence whether confidence assessment 'works' is to do it with some classes and not with others and compare their progress in their school assessments. For example, if your Year 8 cohort were set in two bands, could you have the classes in one band continue as normal and those in the other band try confidence assessment for a term, or longer? Then see how their scores on half-termly assessments, say, compare between the two?

The great thing about confidence assessment is that it should be very easy to implement. You wouldn't need to alter your school-based assessments at all, and students in the 'intervention' half of the year would just be asked to write down a confidence rating beside each answer and then work out their confidence score themselves. There would be no additional tests or marking, and I would be happy to crunch the numbers for the comparison of the students' half-termly/termly marks in the 'intervention' and 'control' groups if you sent me an anonymised spreadsheet. (I wouldn't need data on the confidence assessments themselves.)

Obviously, you would need to decide what is feasible in terms of the number of classes and which year group, and the number of assessment points would be controlled by what you do in your school. I don't want to try to impose constraints on these things, because I don't want to increase anybody's workload and I want to see how confidence assessment might work in the 'natural' setting of real schools.

I am extremely grateful for any help with this that you can give. If the results from this are promising, I intend to apply for funding to do a larger-scale, more robust trial.

I have written a couple of articles about confidence assessment, if you want to read more—available free at https://tinyurl.com/ydb2lhjx and https://tinyurl.com/trrnu2v4 – and I spoke about it on the Mr Barton Maths podcast: https://tinyurl.com/ybpbb68v.

Please let me know if you can help, and get back to me with any questions.

B: *R* Code Used in the Analysis

Creating the Matched Data Sets

```
install.packages("MatchIt")
  library(MatchIt)
  Data <- read.csv("Fulldata.csv", header = TRUE, sep=",")
  Data <- subset(Data, School=="A")
  m.out <- matchit(treat ~ pre, data = Data, method = "nearest")
  summary(m.out)
  head(m.out)
  plot(m.out, type = "hist")
  m.data <- match.data(m.out)
  write.csv(m.data, file = "matched.csv")
```

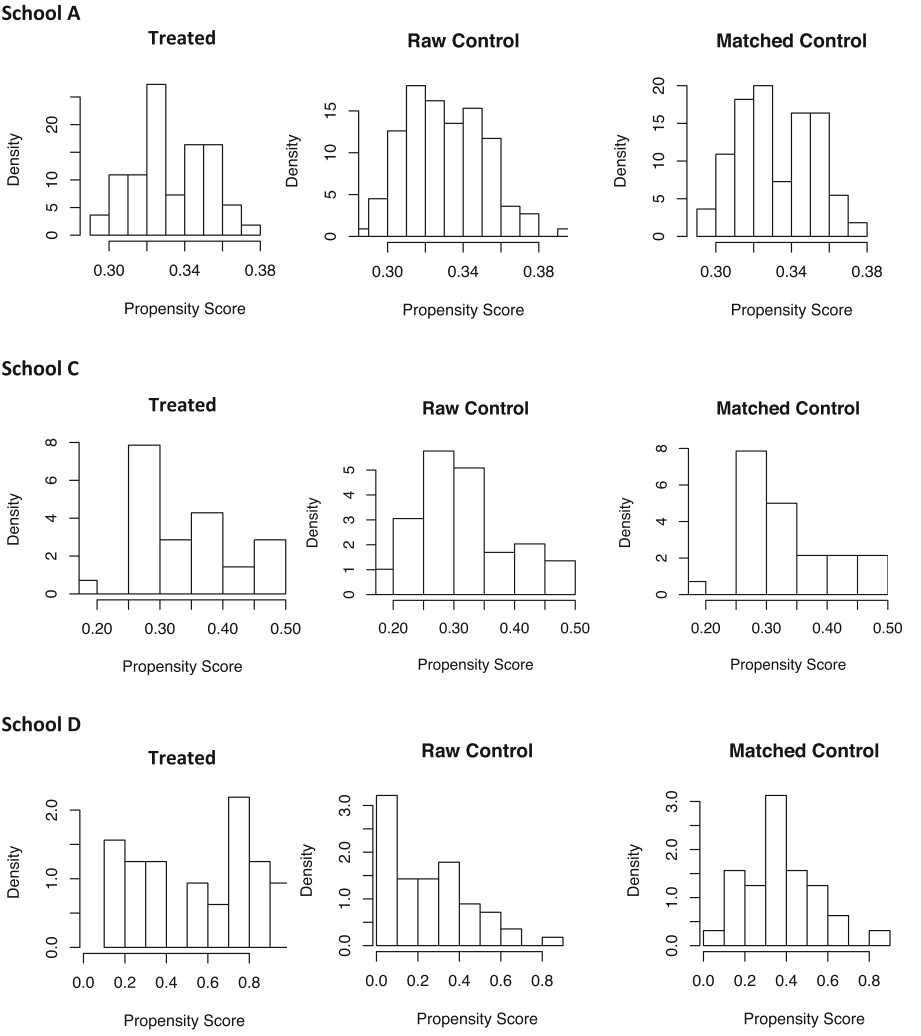
Frequentist Meta-Analysis

```
install.packages("metafor")
  library(metafor)
  Data <- read.csv("Metadata.csv", header = TRUE, sep=",")
  Data$effectsize <- as.numeric(as.character( Data$Cohend))
  Data$var <- as.numeric(as.character( Data$Var))
  res <- rma(yi=effectsize, vi=var, data=Data, slab=paste(School, Ntotal, sep=", "),
method="REML")
  res
  forest(res, xlab="Effect size (d)")
  mtext(bquote(paste("Summary: (Q = ",
  .(formatC(res$QE, digits=2, format="f")), ", df = ", .(res$k - res$p),
  ", p = ", .(formatC(res$QEp, digits=2, format="f")), "; ", I^2, " = ",
  .(formatC(res$I2, digits=1, format="f")), "%)")))
```

Bayesian Meta-Analysis

```
install.packages("BayesFactor")
  library(BayesFactor)
  bf <- meta.ttestBF(Data$t, Data$Ncontrol, Data$Ntreat, rscale=.7071)
  bf[1]
```

C: Raw and Matched Distributions



Note. Treated distribution is included for comparison.

References

Baker, J. (2019). Using low-stakes quizzes with confidence assessment. *Mathematics Teaching, 268*, 11–14.

Barton, C. (2019). *Conference takeaways: ResearchEd Blackpool 2019*. [Audio podcast episode]. Accessed August 4, 2021 from http://www.mrbartonmaths.com/blog/conference-takeaways-researched-blackpool-2019/.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88. https://doi.org/10.1177/0146621697211006.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1*(1), 69–84. https://doi.org/10.1007/s11409-006-6894-z.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press. https://doi.org/10.1515/9781400884940.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2015). *Research methods, design, and analysis*. Pearson.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. https://doi.org/10.3102/00346543053004445.

Clarkson, L. M. C., Love, Q. U., & Ntow, F. D. (2017). How confidence relates to mathematics achievement: A new framework. In A. Chronaki (Ed.), *Mathematics Education and Life at Times of Crisis, Proceedings of the Ninth International Mathematics Education and Society Conference* (Vol. 2, pp. 441-451). University of Thessaly Press. Retrieved from https://muep.mau.se/bitstream/handle/2043/24043/MES9_Proceedings_low_Volume2.pdf?sequence=2&isAllowed=y#page=103

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. https://doi.org/10.3389/fpsyg.2014.00781.

Dirkzwager, A. (2003). Multiple evaluation: A new testing paradigm that exorcizes guessing. *International Journal of Testing, 3*(4), 333–352. https://doi.org/10.1207/S15327574IJT0304_3.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology, 3*(4), 552–564. https://doi.org/10.1037/0096-1523.3.4.552.

Foster, C. (2016). Confidence and competence with mathematical procedures. *Educational Studies in Mathematics*, *91*(2), 271–288. https://doi.org/10.1007/s10649-015-9660-9

Foster, C. (2017). The guessing game. *Teach Secondary*, *6*(8), 85. https://www.foster77.co.uk/Foster,%20Teach%20Secondary,%20The%20guessing%20game.pdf

Foster, C. (2018). Developing mathematical fluency: Comparing exercises and rich tasks. *Educational Studies in Mathematics, 97*(2), 121–141. https://doi.org/10.1007/s10649-017-9788-x

Foster, C., Woodhead, S., Barton, C., & Clark-Wilson, A. (2021). School students' confidence when answering diagnostic questions online. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-021-10084-7

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology, 3*(1), 80–85. https://doi.org/10.3402/rlt.v3i1.9597.

Gardner-Medwin, A. R. (1998). Updating with confidence: Do your students know what they don't know? *Healthcare Informatics, 4*, 45–46.

Gardner-Medwin, A. R. (2006). Confidence-based marking: Towards deeper learning and better exams. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 141–149). Routledge.

Gardner-Medwin, T. (2019). Certainty-based marking: Stimulating thinking and improving objective tests. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education: A handbook for academic practitioners* (2nd ed., pp. 141–150). Routledge. https://doi.org/10.4324/9780429506857-13.

Gardner-Medwin, A. R., & Curtin, N. A. (2007). *Certainty-based marking (CBM) for reflective learning and proper knowledge assessment*. Paper presented at the REAP International Online Conference on Assessment Design for Learner Responsibility. Accessed August 4, 2021 from https://ewds.strath.ac.uk/REAP/reap07/Portals/2/CSL/t2%20-%20great%20designs%20for%20assessment/raising%20students%20meta-cognition/Certainty_based_marking_for_reflective_learning_and_knowledge_assessment.pdf.

Gardner-Medwin, A. R., & Gahan, M. (2003). Formative and summative confidence-based assessment. In J. Christie (Ed.), *Proceedings of the 7th international computer-aided assessment conference* (pp. 147–155). Loughborough University.

Hassmén, P., & Hunt, D. P. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurement, 31*(2), 149–160. https://doi.org/10.1111/j.1745-3984.1994.tb00440.x.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08.

JASP Team (2020). *JASP (Version 0.12.2)* [Computer software]. Accessed August 4, 2021.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford University Press.

Joubert, M., & Sutherland, R. (2009). *A perspective on the literature: CPD for teachers of mathematics*. NCETM. Accessed August 4, 2021 from https://www.ncetm.org.uk/media/1y2dv0zx/ncetm-recme-final-report.pdf.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE Publications Ltd.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166. https://doi.org/10.3102/0013189X19832850.

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353. https://doi.org/10.1037/edu0000281.

McCrea, E. (2019). *Making every maths lesson count: Six principles to support great maths teaching*. Crown House Publishing Limited.

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction, 22*(4), 253–261. https://doi.org/10.1016/j.learninstruc.2011.10.004.

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., & Chabris, C. F. (2019). Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences, 116*(22), 10723–10728. https://doi.org/10.1073/pnas.1820701116.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). *Package 'BayesFactor'*. Accessed August 4, 2021 from ftp://192.218.129.11/pub/CRAN/web/packages/BayesFactor/BayesFactor.pdf.

Newmark, B. (2019). *Why Teach?* John Catt Educational Ltd.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5.

Robson, C., & McCartan, K. (2015). *Real world research* (4th ed.). John Wiley & Sons.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Schoendorfer, N., & Emmett, D. (2012). Use of certainty-based marking in a second-year medical student cohort: A pilot study. *Advances in Medical Education and Practice, 3*, 139-143. https://doi.org/10.2147/AMEP.S35972.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications, 1*(1), 3. https://doi.org/10.1186/s41235-016-0003-x.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747–758. https://doi.org/10.1016/j.lindif.2012.05.013.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2008). *Teacher professional learning and development* (Vol. 18). International Academy of Education. Accessed August 4, 2021 from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.524.3566&rep=rep1&type=pdf.

van Loon, M. H., Dunlosky, J., van Gog, T., van Merriënboer, J. J., & de Bruin, A. B. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology, 42*, 39–48. https://doi.org/10.1016/j.cedpsych.2015.04.003.

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*(3-4), 283–289. https://doi.org/10.1080/10627197.2006.9652993.

Wiliam, D. (2017). *Embedded formative assessment: Strategies for classroom assessment that drives student engagement and learning* (2nd ed.). Solution Tree Press.

Wiliam, D. (2018). *Creating the schools our children need*. Learning Sciences International.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement. Issues & answers* (REL 2007-No. 033). Regional Educational Laboratory Southwest (NJ1). Accessed August 4, 2021 from https://files.eric.ed.gov/fulltext/ED498548.pdf.