


Toward Reducing Anxiety and Increasing Performance in Physics Education: Evidence from a Randomized Experiment

François Molin^{1,2}  · Sofie Cabus^{3,4} · Carla Haelermans¹ · Wim Groot⁴

Published online: 8 May 2019

© The Author(s) 2019

Abstract

This study evaluates the effectiveness of an intervention of formative assessments with a clicker-based technology on anxiety and academic performance. We use a randomized experiment in physics education in one school in Dutch secondary education. For treated students, the formative assessments are operationalized through quizzing at the end of each physics class, where clickers enable students to respond to questions. Control students do not receive these assessments and do not use clickers, but apart from that, the classes they attend are similar. Findings from multilevel regressions indicate that the formative assessments significantly reduce anxiety in physics and improve academic performance in physics in comparison with traditional teaching. Furthermore, a mediation effect of anxiety in physics on academic performance is observed. In sum, this implies that an easy to implement technique of formative assessments can make students feel more at ease, which contributes to better educational performance.

Keywords Formative assessment · Physics · Clicker devices · Secondary education · Anxiety · Academic performance

Edited by: Melissa Hall

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11165-019-9845-9>) contains supplementary material, which is available to authorized users.

✉ François Molin
f.molin@maastrichtuniversity.nl

Sofie Cabus
sofie.cabus@kuleuven.be

Carla Haelermans
carla.haelermans@maastrichtuniversity.nl

Wim Groot
wim.groot@maastrichtuniversity.nl

Extended author information available on the last page of the article

Introduction

Anxiety is becoming more and more a problem in education (Rothman 2004). Data from the Program for International Student Assessment (PISA) 2015, which tests the academic performance of 15-year-olds, showed that more than half of the students feel very >anxious for an exam, even if they are well prepared (Organisation for Economic Co-operation and Development (OECD) 2016). It is not surprising that students experience anxiety before and during exams, because passing or failing a course is often based on a limited number of exams. With so much emphasis on these results, a poor grade in a single exam can have large consequences and have a negative effect on a student's final course grade (Burns 2004) and can even affect admittance to university, course enrolment, career choices, and future employment (Goetz et al. 2013; Udo et al. 2001). A high level of anxiety can distract students during evaluation and prevent them from recalling relevant information, resulting in a lower than expected performance (Hong 2010; Maloney et al. 2013). Anxiety arises when students recognize that their cognitive abilities are overwhelmed by academic demands (González et al. 2017). It is regarded as a serious learning difficulty that hinders students in their learning, students who, by intelligence and hard work, should otherwise perform well (Hong 2010; Mallow 2006). Anxiety is even more a problem in science than in other fields, as science-related fields often have specific prerequisites in grades or passing of science courses. Anxiety of science is described as fear of science subjects and science-related situations (Mallow 1986). It affects students' interest for science lessons and might eventually even act as a serious career filter, preventing students from entering certain science-related fields (Hong 2010; Udo et al. 2004). According to Hong (2010) and Batchelor (2015), only a few studies have examined the effectiveness of interventions aimed at diminishing students' science anxiety. After the review study by Hong (2010), only a few studies have been added to this body of evidence. For example, Brady et al. (2013) and Yu et al. (2014) found that anonymity in classrooms, by polling methods where results are not visible to other learners, releases students' anxiety and nervousness. Anonymity provides more involvement and active participation of shy or anxious students without peer pressure while outcomes improve. McDaniel et al. (2011) reported that anxiety reduces, or disappears, when offering students a series of formative assessments or no-stake assessments. By testing the same course material several times, students become more familiar with testing and enhance their retention and learning (Roediger and Karpicke 2006). These assessments improve students' metacognitive skills about what they do and do not know, without consequences for their grades (Kornell and Son 2009; McDaniel et al. 2011). Formative assessment is usually accompanied by immediate explanatory feedback. Arkin and Schumann (1984) and Rocklin and Thompson (1985) found that immediate explanatory feedback is associated with less anxiety. Furthermore, by explaining why an answer is correct or incorrect, students will improve their understanding of the course material, which reduces their anxiety and helps them to achieve a higher score in a subsequent attempt (Sullivan 2017).

Theoretical Background

Quizzing is a technique of formative assessment which is easy to implement in classrooms. Formative assessment generated by quizzing can be supported by low-tech or high-tech methods (Fallon and Forrest 2011). In low-tech classroom settings, students

answer multiple-choice questions by, for example, raising their hands. According to Caldwell (2007), this method has disadvantages, because it is difficult for a teacher to estimate the number of votes in a limited amount of time. In addition, due to a lack of anonymity, students will not always give their answers honestly (Kay and LeSage 2009). In high-tech classroom settings, small-handheld electronic student response systems are used. Students can then reply to multiple-choice questions projected on a screen in front of the classroom using, for example, clicker devices (Blasco-Arcas et al. 2013; Mayer et al. 2009). As follows, answers are collected by IT-software and, depending on the use and purposes of the teacher, the number of correct and false answers can be summarized and displayed on a screen. Collecting information on students' responses in an efficient and fast way can facilitate the teacher to provide formative feedback, while, at the same time, students remain anonymous.

Previous research has examined the effectiveness of clicker use on improving several aspects of learning, among which anxiety. One area of research focuses on general attitudes toward clicker use, but also on self-reported anxiety during exams in classes these devices are used in formative assessments. McDaniel et al. (2011) showed that frequent clicker use in science classes reduces anxiety during exams, because clickers stimulate active learning. The immediate feedback of clicker use helps students develop confidence in their science skills (Batchelor 2015) and improves students' hopes for success (Fallon and Forrest 2011). Agarwal et al. (2014) found similar results, because students become familiar with taking formative assessments with clickers and gain insight into which course material matters. Stowell and Nelson (2007) reported that students are more inclined to answer questions in classes where clickers are used. Because students can vote anonymously, instructors create a safe learning environment where students can respond without embarrassment (De Gagne 2011). Clickers can also encourage students to join peer discussion (Yu et al. 2014), so that students may feel less anxious when they first discuss with peers rather than with the teacher before voting anonymously. According to Kay and LeSage (2009), an additional advantage is that students see their own answers positioned in relation to answers of their fellow students, allowing them to monitor their own progress, or get confirmation that they are not alone in their misconception (while still being anonymous).

Another area of research investigates whether academic performance increases in classes where clickers are used, where researchers compare clicker groups to comparable non-clicker groups. In general, these studies found positive effects of improving academic performance by clicker use. Mayer et al. (2009), for example, showed that students score significantly higher on exams when they use clickers to answer two to four questions per lesson. Similar results are also found by McDaniel et al. (2011), namely that clicker use for formative assessments in middle school science classes can be extremely effective in increasing academic performance on summative tests.

However, there are also some drawbacks of formative assessments facilitated with clicker devices. The high initial costs might be an economic barrier to integrate clickers into class (Blasco-Arcas et al. 2013), while there is always a possibility of technical malfunctions (Guse and Zobitz 2011). Lantz and Stawiski (2014) state that the time to set up these devices and software, or preparing multiple choice questions before class, may discourage teachers. Other authors, such as Liu et al. (2017) and Hoekstra (2015), consider the difficulty of stimulating higher level thinking with multiple choice questions during formative assessments as a drawback, but also indicate that these questions develop critical thinking of students from seeing and analyzing the various responses of their fellow students.

The Present Study

In sum, the previous literature has shown that formative assessments facilitated with clickers may have the potential to reduce students' anxiety while simultaneously improve academic performance, although one should not overlook the potential drawbacks when implementing clickers.

In the present study, we use a randomized experiment to examine the effects of formative assessments on physics performance and anxiety in physics in secondary education, taking the potential drawbacks into account. For its implementation in daily teaching practice, the formative assessments in the treatment group are facilitated with clickers, while the control group does not use in-class questioning or clickers, but, instead, follows traditional physics teaching.

In the sequel of this paper we will only use the term formative assessments when referring to formative assessments with clicker use.

Research Questions

The effectiveness is evaluated for two particular outcomes, namely anxiety and academic performance. Our research questions are

1. Do formative assessments reduce anxiety in physics compared to traditional teaching?
2. Do formative assessments improve academic performance in physics compared to traditional teaching?
3. Does anxiety work as a mediating factor for the effect of formative assessments on academic performance?

Contribution

We contribute to the literature in at least three ways. First, to the best of our knowledge, there have been no previous studies that investigated the causal relationship of formative assessments and anxiety in physics, or tested a potential mediation effect of anxiety in physics on the relationship between these assessments and academic performance. Second, most previous studies do not control for selection issues, because there is no randomized trial (e.g., Bachman and Bachman 2011; Keough 2012; Shaffer and Collura 2009), or do not take student characteristics into account (e.g., Bartsch and Murphy 2011; Fortner-Wood et al. 2013). The few studies that do use an experimental design are not able to distinguish the effect of formative assessments from other effects, such as class attendance in the experiment. For example, in the studies of Mayer et al. (2009) and Morling et al. (2008) students of the treated (clicker) group are motivated to attend class by earning course credits for answering clicker questions. Students in the control (non-clicker) group do not receive these extra credits for class attendance. In this case, we cannot distinguish whether the estimated improvement in academic performance can be attributed to formative assessments or to higher class attendance. In the study at hand, we carry out a randomized experiment that solely focuses on the effectiveness of formative assessments and do not use reward systems for students' attendances or students' responses, or other potential confounding treatments. Third, we study the effect of formative assessments over a traditional teaching approach in secondary school,

whereas the systematic literature review of Kay and LeSage (2009) showed that most of this research was done at university-level. Only few studies have analyzed the effectiveness in secondary education (Vital 2011). Nevertheless, this knowledge is deemed necessary, since clicker usage is rapidly increasing in secondary education.

Methods

The Intervention

The intervention consists of formative assessments, in which we used a method similar to Mayer et al. (2009). There, students in the treatment group used clickers to answer two to four multiple-choice questions per lesson, while identical students in the control group did not use in-class questioning or clickers. In our study, we similarly apply clickers for formative assessment in class, although it should be noted that we do not use reward systems for students' responses, in contrast with the study of Mayer et al. (2009).

In this study, a total of 73 treated students used a clicker-supported questioning method as a form of formative assessment, and 66 untreated students did not use in-class questioning or clickers, but, instead, followed traditional physics teaching. The treatment group and the control group of each education level were both taught by the same teacher (Fig. 1), to minimize potential teacher effects that might otherwise influence the results. The three physics teachers collaborated voluntarily in this study. In line with other studies (e.g., Mayer et al. 2009), the teachers taught both sections identically: students of each education level received the same lecture contents, notes, assignments, and exam questions. The main (and only) difference between the two sections was the way the teacher interacted with the students at the end of each lesson. Three times per week, at the end of each lesson, students in the treatment group were formatively tested for about 10 to 15 min for a period of 17 weeks. In particular, the clicker was used by treated students to answer four multiple-choice questions each lesson. Each question covered a part of the homework in the textbook, which provided the teacher with valuable information about students' understanding, and the students an insight in the level of their comprehension of the course material (Beatty et al. 2006). The multiple-choice questions had four possible answers, while the fifth option was "I don't know". This latter option should minimize guessing of students and inform the teacher when students really did not know the answer to the question (Caldwell 2007).

Treatment group	73 students	66 students	Control group
Physics teaching AND 10-15 minutes formative assessment with a clicker-supported questioning method. (3 times a week)	<div>Class 1</div> <div>10th grade general secondary education</div> <div>Class 3</div> <div>10th grade pre-university education</div> <div>Class 5</div> <div>11th grade pre-university education</div>	<div>Teacher 1</div> <div>Class 2</div> <div>Teacher 2</div> <div>Class 4</div> <div>Teacher 3</div> <div>Class 6</div>	Physics teaching with in class homework time and feedback opportunity (but WITHOUT in-class formative assessment). (3 times a week)

Fig. 1 Course and design of the intervention

All multiple-choice questions were inserted into PowerPoint slides and projected on a screen in front of the class. In line with the difficulty of the questions, students were given limited time for considering the question individually or discussing with their peers, after which they answered the question individually by choosing the corresponding button on their clicker. The responses of all clickers were registered and recorded by the TurningPoint software. After each round of answering a multiple-choice question, the software presented the distribution of answers as a bar graph to all students on the screen in front of the class. Next, they heard the correct answer and received the teacher's feedback on the most common mistakes made by students (Duncan 2005). Depending on the variety of answers, the teacher could decide whether to spend limited or more elaborated time on feedback. On average, students spent around 1 min per question to provide an answer and around 2 min per question receiving feedback from the teacher on their answers. The purpose of these sessions was not only to provide feedback to students about common misunderstandings or misconceptions, but also to evaluate students' understanding of the course material and visualize academic progression (Premuroso et al. 2011). The treatment students' answers on the multiple-choice questions were not graded. In fact, the response to the questions using clickers was completely anonymous, so that there were no drawbacks for students to answer the questions.

Untreated students in the control group did not receive in-class questioning using clickers at the end of each lesson. Instead, they followed a traditional teaching approach by completing their homework independently or with peers, where they had the opportunity to ask questions to the teacher, and received additional feedback via that way. Figure 1 gives a visual representation of the intervention.

Note that the intervention of formative assessments consisted of an inseparable combination of frequent multiple-choice questions and feedback on the responses to these questions. Researchers have compared various ways to implement formative tests, resulting in fairly detailed instructions about effective instructional design on the matter, e.g., regarding feedback (Larsen and Butler 2013). In this regard, it has been shown that feedback is most effective if given immediately, and substantively elaborate (Roediger and Butler 2011). The current consensus in the literature is that the learning process is more effective when, already during the instruction period, students receive feedback on their progress, both cognitively (knowledge and skills) and metacognitively (their learning techniques). Moreover, the literature stresses that the effect of formative testing may not be limited to its information value ('feedback effect') and related learning incentive (Roediger and Karpicke 2006). Testing also contributes to learning by itself, which is often called the 'testing effect'. Roediger and Butler (2011, p. 20) describe the latter as "the finding that retrieval from memory (i.e. as during a test) produces better retention than restudying the same information for an equivalent time". Roediger and Karpicke (2006), for example, found that an intervention group that received testing performed better than a control group that studied the topic three times (once in class, twice afterwards).

One potential threat for the design of the intervention is that treated and untreated students share the didactic material. This would violate the independence assumption. In order to prevent students from sharing the exact clicker questions with untreated students, students of the intervention group made their notes with pen and paper and left these in the classroom at the end of each lesson. In addition, they were not allowed to use their mobile phones during lessons, in order to avoid that students would take pictures. Here, the supervision by the teacher was crucial to make sure that no learning material of the clicker questions was taken home. If the independence assumption is violated, and

untreated students would have had the same information and clicker questions, it is likely that this would underestimate the true effect.

Assignment to the Treatment

The experiment was conducted in one school in the southern part of The Netherlands. The school is a typical, average mid-sized school outside the highly urbanized, central region of the Netherlands, offering secondary education to 1500 students at three ability levels, theoretical pre-vocational education (4 years), general secondary education (5 years), and pre-university education (6 years). The participating school had been invited by contacting the principal by email. In this email, the purpose and set-up of the study were explained. In a follow-up personal meeting, the researchers explained the importance of randomization of the participants and suggested randomizing before the timetable was made, to prevent timetabling issues interfering with the possibility of randomization. The principal was convinced by the importance of the study and reassured that the study would not interfere with other issues at school. Additionally that it would not be likely that the study would harm students or their performance, and thus agreed to participate. After agreement of the principal, the involved teachers were informed about the aims of the study.

One hundred and thirty-nine physics students participated in the study over a period of 17 weeks at the start of the school year 2016–2017. The students belonged to six different classes; two of 10th grade general secondary education, two of 10th grade pre-university education, and two of 11th grade pre-university education. All the participants were first randomly assigned by the scheduling software Zermelo to one of two classrooms of each education level, after which the school timetable was made. Next, the classes of each education level were allocated randomly to a treated or control group by the researchers. This assignment procedure successfully constructed a comparable intervention- and control group and accounted for potential selection-into-treatment effects, as will be discussed in the next section.

Empirical Strategy

We use a series of two-level hierarchical regression models to test whether formative assessments affect academic performance and anxiety. This multilevel analysis can be formulated as follows:

$$Y_{ij} = \alpha + \beta_0 \theta_{ij} + \sum_{k=1}^k \beta_k x_{k,ij} + (\mu_j + \varepsilon_i) \quad (1)$$

where Y_{ij} denotes the outcome variable anxiety in physics or academic performance on a final exam of student $i \in \{1, 2, \dots, N\}$ attending physics class $j \in \{1, 2, \dots, 6\}$; θ_{ij} the intervention dummy (0,1) with 0 for the control group and 1 for the treatment group; X_{ij} a vector of observed student characteristics prior to the experiment; and ε_i the standard error (note we include control variables to increase the precision in estimation of the effects of our intervention on differences in the outcome measures between students). We also included μ_j , a parameter which denotes unobserved information at the class level. Previous literature has pointed out that teachers' teaching style and interactions between students in class are important factors that may influence the results of the students, regardless of the intervention (Chetty et al. 2011; Koth et al. 2008). Because we are

interested in the causal impact of the treatment on students' outcomes Y_{ij} , it is not desirable that teachers are influencing student outcomes. Therefore, one class of each teacher is randomly assigned to the control group and the other to the treatment group. Furthermore, the intervention takes place in several classes, so there is an individual learning process and a group learning process. The group learning process is an important outcome of the intervention. Because treated students were allowed to have peer discussions on the clicker questions, it is possible that interactions between students in the classes arose. These interactions may also play a significant role in students' anxiety in physics (Guarascio et al. 2017). Therefore, we introduce a multilevel random effects model in order to control for unobserved class (and thereby teacher) level variance in the regression.

Previous literature indicates that formative assessments may not only improve academic performance, but may also reduce anxiety. This immediately raises the question if the relation between these assessments and academic performance are mediated by anxiety in physics. A mediation effect will be assessed using Baron and Kenny's (1986) test for mediation and is present as (a) the intervention significantly impacts academic performance; (b) the intervention has a significant effect on the presumed mediator anxiety in physics; (c) the presumed mediator anxiety in physics is significantly associated with academic performance; and (d) the intervention is no longer significant (complete mediation), or is reduced (partial mediation), when the post-test of anxiety in physics is included in a model that tests the causal effect of the intervention on academic performance.

To reduce the risk of type I statistical error, for all estimates in our multilevel random effects models, we use Bonferroni's adjustment alpha level of 0.025 (Grove and Andreasen 1982).

Outcome Measures

Pre-treatment Information

While a number of clicker-related studies only compare post-outcomes of participants in treatment groups and control groups (Hunsu et al. 2016), this study also uses pre-treatment characteristics on demographic, academic-, and non-cognitive performance skills of the students. First, we collected demographic data and pre-treatment physics grades of all (treated and untreated) students who participated in the intervention from administrative data of the school (Fig. 2). The physics grades are computed as an average of all exam grades in the school year before the intervention. In the Dutch education system, the traditional grading scale is a 90-point scale from 1.0 through to 10.0, with 5.5 being the minimal pass grade. This scale is subdivided with intervals of one decimal place. All physics grades are converted to z-scores in order to calculate effect sizes. By standardizing, we ensure that a reader who is not familiar with the Dutch grade system is also able to interpret the effects.

Furthermore, we collected data on five non-cognitive components, namely: motivation (extent to which students like studying), concentration (extent to which students can concentrate during homework), study approach (extent to which students study efficiently), task approach (extent to which students tackle study components systematically), and memory (extent to which students learn until they know 'everything') (Table 1). To measure these components, we used a validated, self-reported questionnaire called 'Study Conditions Questionnaire' (Vragenlijst Studievoorraarden) (Crins 2002) consisting of 38 three-point Likert scale items ranging from never to always. In further analyses, we use a standardized (z-) score

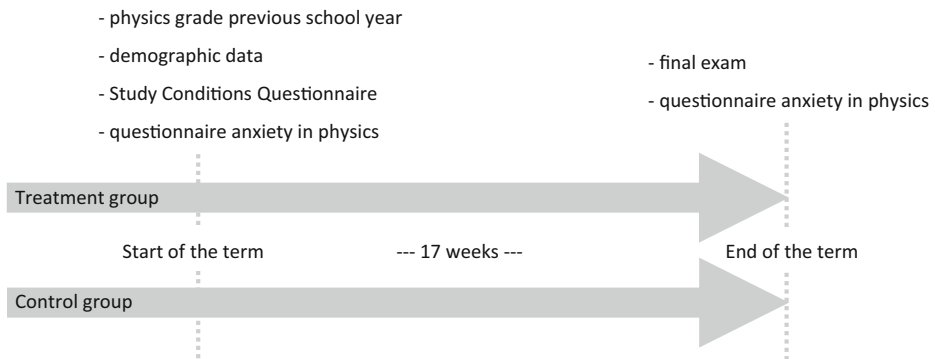


Fig. 2 Research design

of each of the values on the components in order to facilitate the interpretation of results. Higher z-scores indicated more motivation, a better concentration, a better study approach, a better task approach, and more memorization. The internal consistency reliabilities, as measured with Cronbach's alpha, for all of the five components were acceptable (Tavakol and Dennick 2011): motivation $\alpha = 0.65$, concentration $\alpha = 0.80$, study approach $\alpha = 0.68$, task approach $\alpha = 0.73$, and memory $\alpha = 0.79$.

For the pre-treatment outcome variable 'anxiety in physics', we collected data from the students using the Mathematics Anxiety Scale of Betz (1978) adjusted for physics. The scale was intended to assess feelings of anxiety and nervousness related to doing physics and consisted of 6 four-point Likert scale items ranging from 'strongly disagree' to 'strongly agree'. Positively worded items were reversed, so that higher scores indicated more anxiety in physics. The scores on this scale were converted to z-scores (Table 1). The internal consistency reliability of this anxiety scale was 0.68.

Post-treatment Information

In both treated and untreated groups, the academic performance was measured after 17 weeks in a final exam (score out of 50 points). This exam incorporated none of the multiple-choice questions from the clicker sessions, but consisted solely of open

Table 1 Mean differences on standardized pre-treatment characteristics (measured at T0) between treatment and control group using the independent sample *t* test

	<i>Treatment group (N = 73)</i>		<i>Control group (N = 66)</i>		<i>Diff D(1)-D(0)^a</i>	<i>p value</i>
	<i>Mean D(1)^a</i>	<i>St. Dev.</i>	<i>Mean D(0)^a</i>	<i>St. Dev.</i>		
Male	0.58	0.50	0.45	0.50	0.13	0.16
Age	-0.023	0.12	0.025	0.12	-0.048	0.77
Physics grade	0.055	0.12	-0.061	0.12	0.12	0.50
Anxiety pre-score	0.0063	0.12	-0.0069	0.12	0.013	0.18
Motivation	0.12	0.12	-0.13	0.12	0.25	0.14
Concentration	0.11	0.11	-0.12	0.13	0.23	0.18
Study approach	0.20	0.11	-0.22	0.13	0.42	0.014
Task approach	0.093	0.13	-0.10	0.11	0.19	0.25
Memory	0.048	0.12	-0.053	0.12	0.10	0.55

^a Mean standardized scores

questions of standardized Dutch national physics exams constructed by the Central Institute for Test Development (the Dutch abbreviation of this organization is CITO). Each teacher marked the exams of their colleague according to a uniform correction model, not knowing which students belonged to the treatment group and the control group. Hereby, we pursued an increased reliability and internal validity of the test scores on the exam. The scores on these exams were converted to z-scores.

For the measure of anxiety in physics, the students completed the same questionnaire ‘Mathematics Anxiety Scale’ adjusted for physics at the end of the intervention (Fig. 2). The internal consistency reliability of this anxiety scale was 0.79.

Descriptive Statistics

In total, 139 students from 6 different classes participated in this study. At the start of the intervention, the students were on average 16.2 years old ($SD = 0.69$). Fifty-two percent of the participants were male, while students had an average grade of 6.66 points ($SD = 1.02$) on exams in the previous school year.

Table 1 presents a comparison of the observable characteristics of the treatment group and the control group, as well as the statistics of the (significance of the) mean differences. The quality of the randomization was examined using independent two sample t tests. The independent two sample t test shows that students in the treatment group and control group scored, on average, the same on these exams. However, the score for study approach significantly differed between the treatment group and control group; students in the treatment group studied more regularly and more efficiently. Except for the variable study approach, students of the treatment group were, on average, similar with students of the control group. A joint F -test on all the characteristics also does not show a significant difference; $F(9, 129) = 1.09, p = 0.37$.

In the next section, we will control for all of these student background variables in multilevel regressions, when we analyze if formative assessments affect anxiety in physics and improve academic performance.

Results

Anxiety in Physics

The first two models of Table 2 present the results of the multilevel analyses with the outcome measure ‘anxiety in physics’. Model 1 only includes the intervention dummy. Next, we gradually add observed pre-treatment characteristics. The results of model 1 indicate a positive effect of the intervention on anxiety in physics with $\hat{\theta}$ equal to -0.37 points of one standard deviation, significant at 2.5% level (i.e., with Bonferroni correction applied). This corresponds to a small to medium effect size (Cohen 2013) and indicates that the intervention significantly reduced anxiety in physics. In the second model, the control variables gender, physics grade of previous year, anxiety pre-score, and a set of five non-cognitive variables are added. Adding these pre-treatment control variables in the analysis increases the precision of our estimates, as they can predict the differences in the outcome (Bloom et al. 2007; Raudenbush 1997) and thereby make the

Table 2 Multilevel regression analysis predicting post-score on anxiety (model 1 and model 2) and academic performance (model 3 and model 4) and mediation analysis predicting academic performance (model 5 and model 6)

	<i>Model 1</i> <i>anxiety</i> <i>post-score</i>	<i>Model 2</i> <i>anxiety post-</i> <i>score</i>	<i>Model 3</i> <i>academic</i> <i>performance</i>	<i>Model 4</i> <i>academic</i> <i>performance</i>	<i>Model 5</i> <i>academic</i> <i>performance</i>	<i>Model 6</i> <i>academic</i> <i>performance</i>
Intervention ($\hat{\theta}$)	− 0.37** (0.17)	− 0.28** (0.13)	0.46*** (0.16)	0.34** (0.17)	—	0.24 (0.16)
Control variables	No	Yes	No	Yes	Yes	Yes
Anxiety post-score	—	—	—	—	− 0.34*** (0.078)	− 0.33*** (0.078)
ρ	0.00	0.00	0.00	0.034	0.068	0.035
Observations	139	139	139	139	139	139

Significance level denoted at * $p < 0.05$, ** $p < 0.025$, *** $p < 0.01$. Note: Standard errors are in parentheses. Control variables (measured at T0): gender, physics grade previous school year, anxiety pre-score, study approach, motivation, concentration, task approach, and memory. Anxiety post-score has been measured at T1

model better performing. Even though most control variables were not significantly different between treatment and control students before the intervention, they still influence the outcome and thereby add precision to the model. Model 2 indicates that the effect of the intervention on academic performance ($\hat{\theta} = -0.28$) remains significant at 2.5% level. Furthermore, gender ($\hat{\beta}_{\text{gender}} = -0.36$, $p < 0.01$), physics grade of last year ($\hat{\beta}_{\text{physics}} = -0.27$, $p < 0.01$), anxiety pre-score ($\hat{\beta}_{\text{anxiety pre-score}} = 0.37$, $p < 0.01$), and motivation ($\hat{\beta}_{\text{motivation}} = -0.30$, $p < 0.01$) are significant predictors of the post-anxiety score.¹ This means that students who already experienced more anxiety before the intervention also experience more anxiety in physics after the intervention. On the other hand, post-test anxiety is lower if students are more inclined to study or willing to commit themselves to their study (variable motivation). The variable study approach, in which the treatment group and control group differed significantly before the intervention, does not significantly predict the post-anxiety score, nor does it influence the significance and magnitude of the intervention.

Academic Performance

Models 3 and 4 of Table 2 present the results for the outcome variable academic performance. In model 3, the estimate of $\hat{\theta}$ is equal to 0.46 points of standard deviations significant at the 1% level and corresponds to a medium effect size (Cohen 2013). When additional pre-treatment control variables are included in model 4, the effect of participation in the treatment group remains robust with $\hat{\theta}$ equal to 0.34 points of standard deviations, significant at 2.5% level. Furthermore, gender ($\hat{\beta}_{\text{gender}} = 0.30$, $p < 0.05$), physics grade of last year ($\hat{\beta}_{\text{physics}} = 0.57$, $p < 0.01$), and motivation ($\hat{\beta}_{\text{motivation}} = 0.24$, $p < 0.01$) are significant predictors of academic performance. After controlling for a couple of other factors, we see that men still perform better than women on physics exams. Therefore, we also estimated our model with interaction

¹ See Table X in Supplementary Material for the full regression tables.

effects between intervention and gender, to see if the intervention might have a differentiating effect by gender, but we did not find a significant effect.

Mediation Analysis

Anxiety in physics might mediate the estimated effects of the intervention on academic performance. Baron and Kenny (1986) proposed four steps for testing this kind of mediation effect. First, consider again the multilevel analysis showing that the intervention had a significant positive effect on academic performance (model 4 in Table 2; $\hat{\theta} = 0.34$, $p < 0.025$). This effect is presented in Fig. 3, path A. Second, consider the multilevel analysis showing that the intervention had a significant negative effect on anxiety in physics (model 2 in Table 2; $\hat{\theta} = -0.28$, $p < 0.025$). This effect is presented in path B in Fig. 3. Then, third, the post-score on anxiety in physics is included into a regression with outcome academic performance. If not controlled for the intervention dummy, the results indicate that the presumed mediator anxiety in physics significantly correlates to academic performance (model 5 in Table 2; $\hat{\rho} = -0.34$, $p < 0.01$). This effect is presented in Fig. 3, path C. It is now intuitive that anxiety in physics might mediate the estimated effect of the intervention on academic performance, and that this mediation effect can be revealed by including the post-scores on anxiety in physics into the regression (Table 2; model 6). Doing so, however, shows that the estimate of $\hat{\theta}$ no longer significantly predicts academic performance; $\hat{\theta} = 0.24$, ns (path D in Fig. 3). Therefore, it is concluded that anxiety in physics mediates the effects of the intervention on academic performance. We should note that we cannot disregard issues with statistical power, so the results of the significances should be interpreted with caution. However, apart from the insignificant coefficient, the decrease in effect size (from 0.34 to 0.24) is also an indication that the effect of the intervention on performance is mediated by anxiety in physics.

To conclude, we also calculated the intraclass correlation coefficients (ICC) to estimate the percentage of variance of the outcomes anxiety in physics or academic performance explained by unobserved class effects. In most models, the ICCs are quite low, varying from $\rho = 0.00$ to $\rho = 0.07$. This means that almost all the variance is explained by student differences and not by unobserved class effects (Peugh 2010). However, although in most models, we find that the percentage share of variance of the outcome variables explained by unobserved class effects is less than 0.05 (which is the rule of thumb from Hox (1998) for deciding against the use of the multilevel model), we still opt for the multilevel random effects model. The most important

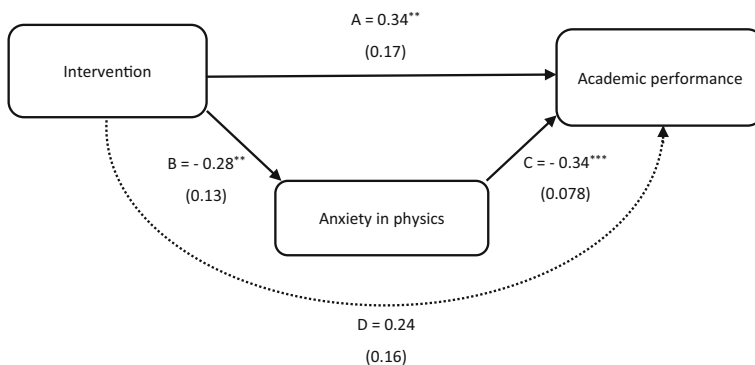


Fig. 3 Mediation analysis

reason is that in model 5, and to a lesser extent in models 4 and 6, we do observe class effects. Furthermore, the multilevel model also allows us to account for class differences (McNeish 2014), whereas using fixed effects for class (which also accounts for class differences) adds coefficients to the models, which lowers the degrees of freedom and is not a good idea given our number of observations.

Conclusion and Discussion

The aim of this paper was to evaluate the effects of using repeated formative assessments compared to traditional teaching. A randomized experiment was carried out over a period of 17 weeks among 139 secondary students and analyzed at student level, while controlling for class in a multilevel setting. The study answers our three research questions:

First, the results show that formative assessments improve academic performance in physics compared to traditional teaching; treated students have significantly higher grades on the post-test than untreated students. It corresponds to an effect of 0.34 points of standard deviations, significant at 2.5% level, a medium effect size. This finding is in line with the studies of, e.g., Bachman and Bachman (2011), Lin et al. (2011), and Mayer et al. (2009), who find that students in a clicker group outperform students in a non-clicker group. Our result is comparable with average gains in effect size in active learning strategies, which is about 0.31 in physics (Freeman et al. 2014). Comparable effects are also found in studies that relate to the testing effect (e.g., Roediger and Karpicke 2006), where small weekly tests have a greater positive effect on future retention of course material than spending an equivalent amount of time restudying the material.

Second, compared to traditional teaching, formative assessments significantly reduce anxiety in physics. This effect is equal to 0.28 points of a standard deviation, significant at 2.5% level, a small to medium effect size. These findings differ from the findings in the scarce studies in the literature on this topic; Sun (2014) and Batchelor (2015) both show insignificant differences in anxiety between treated classes and untreated classes, although both studies suffer from a low power. Noteworthy is that the study of Batchelor (2015) measures a statistical increase in anxiety during the semester in both classes. However, given that these studies focused on university settings, it is hard to compare these effects with secondary education.

And third, a mediation analysis shows that anxiety in physics mediates the effects of formative assessments on academic performance. This means that this form of assessment significantly reduces anxiety, which in turn also significantly affects academic performance. Although we may have expected the class level to introduce bias in the results (since peer discussions between students may play a significant role in improving academic performance and reducing students' anxiety (Wiggs 2011)), the analysis does not show significant bias, neither on anxiety in physics nor on academic performance. We did not collect information from students that could explain this, but the teachers indicated that most of the treated students answered multiple choice questions individually and in silence, without consulting their peers. As a result of these findings, we are inclined to assume that treated students experience less anxiety and become more familiar with taking assessments, because they get more moments of performance feedback in a single lesson, than students in the control group. The repeated assessments divide the course material into small units and the real-time feedback from the teacher helps students to monitor their own understanding of the material.

Simultaneously, the anonymity enabled by clickers and the general group feedback is less risky to expose students' weakness to peers and teacher, which could give rise to anxiety. On the other hand, anxiety in physics of untreated students is not addressed specifically, because they do not receive in-class questioning and receive only selective feedback when asked for it. These assumptions are in line with the literature that a series of anonymous formative assessments with multiple-choice questions and explanatory group-feedback of teachers enables students to grow their self-confidence and improves their metacognitive skills that reduces anxiety (Brown et al. 2014; Kornell and Son 2009; Sullivan 2017).

Although the present study provides evidence that treated students, who receive formative assessments, experience less anxiety and perform better compared to traditional teaching, this study does not allow us to determine whether this effect is mainly due to the repeated testing or to the provided feedback based on the answers to the clicker questions. Furthermore, it is unclear whether the chosen division of attention to these two aspects is most efficient. We suggest further research to focus on separate aspects, or on a different share of time spent on each aspect, to determine which one of those is most important for the significant positive results.

Despite our good preparations, at least two caveats should be considered when interpreting the results. First, the students, who participated in this study, had not used clickers before. It is therefore possible that students were excited about using new course material, were studying harder, or even had more commitment to the course material, then before the introduction of clickers. As such, due to the technology, the estimated effect of the intervention may be upwardly biased. We did not collect information that could unravel novelty effects; however, teachers have said that students became more accustomed to using clickers over time, accepting clickers as a standard study tool. Since this study lasted 17 weeks, we expect that novelty effects faded out over time and did not, or at least not substantially, bias our results. Second, the scope of this study was limited to physics teaching at only one secondary school in the Netherlands. Further research should indicate whether the findings in this study also apply to other science subjects taught in secondary education.

Although the intervention was carried out among only 139 students, it is important to indicate that substantial effects can be achieved in secondary school physics courses if students are formatively assessed for only 10 to 15 min each lesson. These positive outcomes may stimulate physics teachers to implement formative assessments with clickers in their lessons.

Acknowledgements The authors are grateful for the discussions with Joris Ghysels, Trudi Schils, Raoul Haenbeukers, Kevin Bamforth, Herman Franssen, Louis Lenders, Han Jeurissen, Wim Timmermans, and all students of College Den Hulster in Venlo who participated in this study.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, 3(3), 131–139.

- Arkin, R. M., & Schumann, D. W. (1984). Effect of corrective testing: An extension. *Journal of Educational Psychology*, 76(5), 835–843.
- Bachman, L., & Bachman, C. (2011). A study of classroom response system clickers: Increasing student engagement and performance in a large undergraduate lecture class on architectural research. *Journal of Interactive Learning Research*, 22(1), 5–21.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.
- Bartsch, R. A., & Murphy, W. (2011). Examining the effects of an electronic classroom response system on student engagement and performance. *Journal of Educational Computing Research*, 44(1), 25–33.
- Batchelor, J. (2015). Effects of clicker use on calculus students' mathematics anxiety. *PRIMUS*, 25(5), 453–472.
- Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics*, 74(1), 31–39.
- Betz, N. E. (1978). Prevalence, distribution, and correlates of math anxiety in college students. *Journal of Counseling Psychology*, 25(5), 441–448.
- Blasco-Arcas, L., Buil, I., Hernández-Ortega, B., & Sese, F. J. (2013). Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers & Education*, 62, 102–110.
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30–59.
- Brady, M., Seli, H., & Rosenthal, J. (2013). Metacognition and the influence of polling systems: How do clickers compare with low technology systems. *Educational Technology Research and Development*, 61(6), 885–902.
- Brown, P. C., Roediger, H. L., III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Boston: Harvard University Press.
- Burns, D. J. (2004). Anxiety at the time of the final exam: Relationships with expectations and performance. *Journal of Education for Business*, 80(2), 119.
- Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, 6(1), 9–20.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. NBER Working Paper No. 17699. *National Bureau of Economic Research*.
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Crins, J. (2002). *Vragenlijst studievoorwaarden*. KPC Onderwijs Innovatie Centrum, 's-Hertogenbosch, Nederland, 3.100.11.
- De Gagne, J. C. (2011). The impact of clickers in nursing education: A review of literature. *Nurse Education Today*, 31(8), e34–e40.
- Duncan, D. (2005). *Clickers in the classroom: How to enhance science teaching using classroom response systems (Vol. 1)*. San Francisco: Pearson Education.
- Fallon, M., & Forrest, S. L. (2011). High-tech versus low-tech instructional strategies: A comparison of clickers and handheld response cards. *Teaching of Psychology*, 38(3), 194–198.
- Fortner-Wood, C., Armistead, L., Marchand, A., & Morris, F. B. (2013). The effects of student response systems on student learning and attitudes in undergraduate psychology courses. *Teaching of Psychology*, 40(1), 26–30.
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.
- Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, 24(10), 2079–2087.
- González, A., Fernández, M. V. C., & Paoloni, P. V. (2017). Hope and anxiety in physics class: Exploring their motivational antecedents and influence on metacognition and performance. *Journal of Research in Science Teaching*, 54(5), 558–585.
- Grove, W. M., & Andreasen, N. C. (1982). Simultaneous tests of many hypotheses in exploratory research. *Journal of Nervous and Mental Disease*, 170, 3–8.
- Guarascio, A. J., Nemecek, B. D., & Zimmerman, D. E. (2017). Evaluation of students' perceptions of the Socratic application versus a traditional student response system and its impact on classroom engagement. *Currents in Pharmacy Teaching and Learning*, 9(5), 808–812.
- Guse, D. M., & Zobitz, P. M. (2011). Validation of the audience response system. *British Journal of Educational Technology*, 42(6), 985–991.

- Hoekstra, A. (2015). Because you don't realize how many people have different experiences than you: Effects of clicker use for class discussions in sociology. *Teaching Sociology*, 43(1), 53–60.
- Hong, Z. R. (2010). Effects of a collaborative science intervention on high achieving students' learning anxiety and attitudes toward science. *International Journal of Science Education*, 32(15), 1971–1988.
- Hox, J. (1998). Multilevel modeling: When and why. In *Classification, data analysis, and data highways* (pp. 147–154). Springer, Berlin, Heidelberg.
- Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, 94, 102–119.
- Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education*, 53(3), 819–827.
- Keough, S. M. (2012). Clickers in the classroom: A review and a replication. *Journal of Management Education*, 36(6), 822–847.
- Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, 17(5), 493–501.
- Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, 100(1), 96–104.
- Lantz, M. E., & Stawiski, A. (2014). Effectiveness of clickers: Effect of feedback and the timing of questions on learning. *Computers in Human Behavior*, 31, 280–286.
- Larsen, D. P., & Butler, A. C. (2013). Test-enhancing learning. In K. Walsh (Ed.), *Oxford textbook of medical education* (pp. 443–452). Oxford: Oxford University Press.
- Lin, Y. C., Liu, T. C., & Chu, C. C. (2011). Implementing clickers to assist learning in science lectures: The Clicker-Assisted Conceptual Change model. *Australasian Journal of Educational Technology*, 27(6), 979–996.
- Liu, C., Chen, S., Chi, C., Chien, K. P., Liu, Y., & Chou, T. L. (2017). The effects of clickers with different teaching strategies. *Journal of Educational Computing Research*, 55(5), 603–628.
- Mallow, J. V. (1986). *Science Anxiety: Fear of Science and How to Overcome It (revised edition)*. Clearwater: H&H Publications.
- Mallow, J. V. (2006). Science anxiety: Research and action. In J. J. Mintzes & W. H. Leonard (Eds.), *Handbook of college science teaching* (pp. 3–14). Arlington, VA: NSTA Press.
- Maloney, E. A., Schaeffer, M. W., & Beilock, S. L. (2013). Mathematics anxiety and stereotype threat: Shared mechanisms, negative consequences and promising interventions. *Research in Mathematics Education*, 15(2), 115–128.
- Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, 34(1), 51–57.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399–414.
- McNeish, D. M. (2014). Analyzing clustered data with OLS regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, 40, 11–16.
- Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. M. (2008). Efficacy of personal response systems ("clickers") in large, introductory psychology classes. *Teaching of Psychology*, 35(1), 45–50.
- OECD, PISA 2015 Results in Focus. PISA, OECD Publishing, 4–14 (2016).
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology* 48 (1):85–112
- Premuroso, R. F., Tong, L., & Beed, T. K. (2011). Does using clickers in the classroom matter to student performance and satisfaction when taking the introductory financial accounting course? *Issues in Accounting Education*, 26(4), 701–723.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185.
- Rocklin, T., & Thompson, J. M. (1985). Interactive effects of test anxiety, test difficulty, and feedback. *Journal of Educational Psychology*, 77(3), 368–372.
- Roediger, H., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, 15(1), 20–27.
- Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210.
- Rothman, D. K. (2004). New approach to test anxiety. *Journal of College Student Psychotherapy*, 18(4), 45–60.
- Shaffer, D. M., & Collura, M. J. (2009). Evaluating the effectiveness of a personal response system in the classroom. *Teaching of Psychology*, 36(4), 273–277.
- Stowell, J. R., & Nelson, J. M. (2007). Benefits of electronic audience response systems on student participation, learning, and emotion. *Teaching of Psychology*, 34(4), 253–258.

- Sullivan, D. (2017). Mediating test anxiety through the testing effect in asynchronous, objective, online assessments at the university level. *Journal of Education and Training*, 4(2), 107–123.
- Sun, J. C. Y. (2014). Influence of polling technologies on student engagement: An analysis of student motivation, academic performance, and brainwave data. *Computers & Education*, 72, 80–89.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55.
- Udo, M. K., Ramsey, G. P., Reynolds-Alpert, S., & Mallow, J. V. (2001). Does physics teaching affect gender-based science anxiety? *Journal of Science Education and Technology*, 10(3), 237–247.
- Udo, M. K., Ramsey, G. P., & Mallow, J. V. (2004). Science anxiety and gender in students taking general education science courses. *Journal of Science Education and Technology*, 13(4), 435–446.
- Vital, F. (2011). Creating a positive learning environment with the use of clickers in a high school chemistry classroom. *Journal of Chemical Education*, 89(4), 470–473.
- Wiggs, C. M. (2011). Collaborative testing: Assessing teamwork and critical thinking behaviors in baccalaureate nursing students. *Nurse Education Today*, 31(3), 279–282.
- Yu, Z., Chen, W., Kong, Y., Sun, X. L., & Zheng, J. (2014). The impact of clickers instruction on cognitive loads and listening and speaking skills in college English class. *PLoS One*, 9(9), e106626.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

François Molin^{1,2} • Sofie Cabus^{3,4} • Carla Haelermans¹ • Wim Groot⁴

¹ School of Business and Economics, Maastricht University, Po Box 616, 6200 MD Maastricht, The Netherlands

² Onderwijsgemeenschap Venlo and Omstreken, College Den Hulster, Hagerhofweg 15, 5912 PN Venlo, The Netherlands

³ Research Institute for Work and Society (HIVA), KU Leuven, Parkstraat 47 bus 5300, 3000 Leuven, Belgium

⁴ Maastricht Graduate School of Governance, Maastricht University, Po Box 616, 6200 MD Maastricht, The Netherlands