CERGE
Center for Economic Research and Graduate Education
Charles University Prague

Essays on Incentives and Information in Schools

Dagmara Celik Katreniak

Dissertation

Prague, August 2016

Dissertation Committee:

Michal Bauer (CERGE-EI, Co-Chair)

Randall K. Filer (Hunter College, City University of New York, Co-Chair)

Peter Katuscak (CERGE-EI)

Daniel Munich (CERGE-EI)

Nikolas Mittag (CERGE-EI)

Referees:

Karna Basu (Hunter College, City University of New York)

Gil S. Epstein (Bar-Ilan University)

Contents

Acknowledgements

First and foremost, I would like to thank my main supervisors, Michal Bauer and Randy Filer. I was introduced to experimental economics in Michal's lecture back in 2010, which changed my career fundamentally. His encouragement and guidance helped me to develop and implement a large-scale field experiment, which essentially combined my professional work with my personal interest. Randy patiently and systematically believed in me and provided me with guidance without which I would not have been able to put this project into place. I am also grateful to other members of my committee, namely, Nikolas Mittag, Daniel Munich, and Peter Katuscak, whose comments contributed to significant improvements in my dissertation. I would also like to thank Barbara Forbes and Deborah Novakova who helped me to turn this dissertation into a readable and understandable text. My gratitude also goes to my external referees, Gil S. Epstein and Karna Basu, for their helpful comments.

I would also like to thank the wider community of CERGE-EI faculty members who interacted with me during my studies – Kresmir Zigic, Patrick Gaule, Stepan Jurajda, Eva Vourvachaki, Fabio Michelucci, Andreas Ortmann, Jan Zapal, Filip Matejka, Honza Hanousek, Alena Bicakova, Jakub Steiner and Byeongju Jeong – I am most grateful to them for their helpful input.

I was lucky to have amazing classmates/colleagues around me. Special "thanks" goes to Eva Hromadkova, Jana Cahlikova and Pavla Vozarova for being there for me at any time. Thank you Andrea Majekova, and Branislav Zudel for motivating me to work hard during painful study-time and helping me to stay focused. Thank you Nata Shestakova, Klara Kaliskova, Vojtech Bartos, Lubomir Cingl, Tomas Lichard, Volha Audzei, Tomas Miklanek and Mirka Federicova for your critical comments that helped to shape my ideas.

I would not have been able to implement my project without my local team in Uganda – thank you Winifred Candiru, Yaseen Nsubuga, Sandra Basemera, Semei Mukisa, Hanifa Zawedde for hours spent on boda-boda and for your patience with me in the field. Thank you, Juliana Bukirwa, for endless administrative support and making matatu drivers

deliver missing exams to schools on time. Thank you, Ramjet Banura, Remmy Nambowa and Isma Nyombi for joining our team for follow-up testing. And thank you, Grace Mboizi, for not letting me get lost in Ugandan villages.

Special thanks go to Misa Chatrna, Evicka Lakoma, Mirka Dolezalova, Pavli Danhelkova, Zuzka Kuranova, Baska Silharova, Klara Janotova and Zuzka Kazdova for unforgettable moments in Uganda. Girls, you made every day brighter - literally.

This work would not have met its deadline without endless support of Kresimir Zigic, Lenka Pavlikova, Iva Havlickova, and Tereza Kulhankova as well as the members of the CERGE-EI Academic Skills Center. Thank you very much for your help, hard work and for being patient with me.

Finally, I would like to thank my parents for their unlimited support and for being on my side when I decided to spend two years in Uganda (and for stopping their questions about the date of my dissertation defense).

This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network (GDN) and Grant Agency of Charles University (GAUK). All opinions expressed and all errors are mine. The project was implemented in close cooperation with Charitas Prague and Uganda Czech Development Trust.

I dedicate this work to my husband, Levent Celik and our daughter Maya, who helped me to fulfill this goal. Without your support I would be still half-way through. Sizi cok seviyorum.

Prague, Czech Republic Dasa

August, 2016

Abstract

The question posed in this dissertation is whether the quality of education can be improved in a developing country by means of incentives for students to learn. This complex topic has been subject to a plethora of research studies in economics, psychology, and sociology using data from developed countries, but comparatively few studies have been conducted in the developing world. I discuss evidence from an extensive randomized control trial (RCT) employing a variety of incentive mechanisms, which I designed and implemented in primary and secondary schools in Southern Uganda. This study involved more than 5,000 students aged 11 through 25 who were repeatedly interviewed and tested between 2011 and 2013. I collected data and analyzed the effects of different incentive schemes on students' performance on Math and English tests, and also on their well-being, measured by perceived happiness and stress. The latter is a unique contribution to this field of study.

The Preface provides contextual informaton on the Ugandan education system and the experimental design, critical to understanding the choices made at every level of this study. In Chapter 1, "The Dark Side of Incentives in Schools," I discuss the effects of feedback, as well as monetary and non-monetary incentives on students' performance and well-being. This study contributes by explicitly accounting for the tradeoffs between performance and well-being introduced by incentives. I implement two types of social comparative feedback regimes, within- and across-class group comparisons, and two types of incentive regimes, financial and reputation rewards. The results show that rewards can improve performance but at a cost of higher stress and lower happiness, whereas comparative feedback alone

(without rewards) increases performance only mildly but without a negative impact on student's stress and happiness levels. Moreover, the results show that more highly stressed students exert less effort, perform less well and are more often absent than those who are minimally stressed. Finally, the results also help to identify gender-specific responses to incentives: boys react strongly to rewards, but girls do so only if they are also given feedback.

In Chapter 2, "Information Provision and Overconfidence," I investigate whether and how students calibrate self-assessment of their performance in response to feedback and contribute evidence to the debate regarding the existence of the unskilled-but-unaware phenomenon.

While previous studies have found performance to be related to subjects' confidence (Camerer and Lovallo, 1999), some subjects consistently overestimate their abilities (e.g., Ehrlinger et al., 2008). Although informing subjects about their performance has been shown to decrease their inflated beliefs (e.g., Ryvkin et al., 2012), they remain overconfident (e.g., Lipko et al., 2009). A possible explanation is that they lack information about others.

As described in Chapter 2, students in the current RCT, who were from primary and secondary schools in Southern Uganda (as opposed to a typical sample involving (under)graduate students from developed countries), were evaluated and incentivized in groups repeatedly during an academic year. Students received complex feedback about their own performance and the performance of other group members.

The results show that the overconfidence of students in the control group (who received no feedback) increased with repeated testing, whereas feedback received by the treatment groups lowered students' inaccurate estimates of their performance. Students reacted immediately after they received the first feedback, by improving their estimation about their own performance. Nevertheless, overconfidence remained. Although students improved continuously in every round, the most significant improvements were achieved after the first two feedback rounds. Girls updated significantly more compared to boys.

Consistent with the "unskilled-and-unaware phenomenon", the bottom-quartile performers grossly overestimated their performance, although, interestingly, so did top-quartile performers, though to a significantly lesser degree. It is worth noting that the current experimental design makes it possible to document that the "unskilled-and-unaware phenomenon" is a behavioral regularity rather than a statistical artefact.

Abstrakt

Otázka, kterou se v mé disertaci zabývám spočívá v tom, jestli kvalita vzdělávání v rozvojové zemi může být zvýšena pomocí použití motivačních nástrojů navázaných na jejich výsledky ve školách. Toto obšírné téma bylo a je v hledáčku zájmu velkého množství studií v oblasti ekonomie, psychologie i sociologie založených na datech z rozvinutých zemí, zatímco relativně menší pozornost byla věnována rozvojovým krajinám. V disertaci přináším výsledky obsáhlého experimentu založeného na náhodné alokaci studentů základních a středních škol v Jižní Ugandě do skupin s nebo bez použití motivačních nástrojů (takzvaný „randomized control trial“). Celkem se studie zůčastnilo více než 5 tisíc studentů ve věku 11 až 25 let, kteří byli opakovaně testováni a tázáni v letech 2011 až 2013. Dataset obsahuje údaje ohledně studijních výsledků z Matematiky a Angličtiny, zároveň subjektivní hodnocení vlastní spokojenosti měřeno pomocí vlastního vnímání radosti a stresu.

V Předmluvě disertace poskytuji informace ohledně ugandského vzdělávacího systému a detailní popis designu experimentu s cílem ulehčit čitateli porozumění kontextu experimentu a jednotlivých rozhodnutí v jednotlivých krocích. V první kapitole nazvané “Temné Stránky Motivačních Nástrojů ve Školách” se zabývám efekty poskytování zpětné vazby, finančních a nefinančních odměn na studijné výsledky a spokojenost studentů. Hlavním přínosem této studie je explicitní srovnání efektů incentív na studijní výsledky a na spokojenost měřenou pomocí radosti a stresu. Celkem jsem zavedla dvě motivační schémy založené na zpětné vazbě (v rámci skupinek ve třídě nebo mezi třídami) a dvě založené na rozdávání odměn vítězům (finanční nebo reputační odměny). Výsledky ukazují,

že odměny sice motivují studenty zlepšit jejich studijní výsledky, jde to ale na úkor zvýšení stresu a snížení radosti, zatímco zpětná vazba má slabší vliv na zlepšení studijních výsledků, ale neovlivňuje spokojenost studentů. Zároveň výsledky poukazují na to, že studenti vykazující vyšší úroveň stresu vynakládají menší úsilí, vykazují horší výsledky a jsou častěji nepřítomní ve srovnání se studenty s minimální úrovní stresu. Výsledky zároveň pomáhají rozlišit odezvy na motivační faktory podle pohlaví: zatímco kluci reaguji pozitivně na odměny, holky reagují na zpětnou vazbu. Holky reagují na odměny pouze v případě, že dostávaly zpětnou vazbu.

V druhé kapitole nazvané „Poskytování informací a přehnaná sebedůvěra“ zkoumám jestli a jakým způsobem studenti kalibrují sebehodnocení vlastních studijních výsledků v návaznosti na poskytnutí zpětné vazby. Výsledky této studie zároveň přispívají k diskusi ohledně existence takzvaného „unskilled-but-unaware“ fenoménu.

Zatímco předešlé studie poukazují na propojenost studijních výsledků a sebedůvěry studentů (Camerer a Lovallo, 1999), někteří jedinci systematicky nadhodnocují vlastní schopnosti (např. Ehrlinger a spol., 2008). Podávání zpětné vazby subjektům ohledně jejich studijních výsledků se ukázalo jako účinné ve snaze snížit přehnaná očekávání (např. Ryvkin a spol., 2012). Nicméně sebedůvěra studentů zůstává nadhodnocená (např. Lipko a spol., 2009). Možným vysvětlením je právě to, že subjektům chybí detailní informace ohledně výsledků ostatních subjektů.

Studenti tohoto experimentu, kteří navštěvovali základní a střední školy v jižní části Ugandy (na rozdíl od v literatuře převažujícího vzorku studentů vysokých škol z rovzinutých zemí), byli testovaní a odměňování ve skupinkách opakovaně po dobu

jednoho školního roku. Studenti v motivačním schématu se zpětnou vazbou získávali informace ohledně vlastních studijních výsledků a výsledků členů jejich skupiny.

Výsledky poukazují na to, že studenti v kontrolní skupině (kteří nedostali v průběhu školního roku žádnou zpětnou vazbu) postupně zvyšovali svojí sebedůvěru s každým kolem testování, zatímco studenti, kteří dostávali opakovaně zpětnou vazbu snížili svá přehnaná očekávání ohledně vlastních výsledků. Studenti reagovali hned na podání první zpětné vazby tím, že snížili přehnaná očekávání. Nicméně zůstali přehnaně sebejistí. Studenti postupně zlepšovali přesnost hodnocení vlastních výsledků, nicméně hlavní zlepšení se dostavilo po obdržení první a druhé zpětné vazby. Dívky zlepšily sebehodnocení lépe než kluci.

V souladu s „unskilled-and-unaware" fenoménem, studenti ze spodního kvartilu statistické distribuce studijních výsledků vykazovali signifikantně vyšší sebedůvěru ve srovnání se studenty z horního kvartilu. Studenti z horního kvartilu nadhodnotili vlastní výsledky sice také ale v signifikantně nižší míře. Výsledky zároveň poukazují na to, že „unskilled-and-unaware" fenomén je spíš behaviorální zákonitost než statistický artefakt.

Preface: Essential Background

Although substantial progress has been made in improving access to schooling in developing countries, higher enrollment needs to be accompanied by advances in education quality in order to achieve sustainable improvement (Hanushek, 2005). Among the approaches to improving quality, considerable attention has been paid recently to provision of controlled information and different types of incentives. Little attention has been paid, however, to the consequences of incentives on agents' well-being, despite the fact that well-being is related to health, awareness, memory, and performance.

Improvements in performance may be connected to students' expectations regarding their performance. People – especially the unskilled at the bottom end of the performance distribution - are typically overconfident about their performance, i.e., they expect that they will score higher than they do in reality. Inaccurate predictions of one's own ability may have economic consequences (e.g., enterpreneur failures as in Camerer and Lovallo, 1999). The design of this experiment allows me to compare the effects of various incentive schemes on calibration of student self-assesment.

To the best of my knowledge, the current study is the first large scale experiment in a developing country studying the effects of feedback, incentives, and their interactions on student performance, well-being and confidence levels. The uniqueness of the experiment lies in its complexity as well as in the fact that more than 5,000 students in 52 schools in Southern Ugandan villages were tested and interviewed repeatedly during the 2012 academic year.

The dissertation is organized as follows. First, in the Preface: in Essential Background, I describe the education system in Uganda and explain the experimental design in detail. In Chapter 1, "Dark Side of Incentives: A Randomized Field Eexperiment in Uganda," I first provide a literature review using relevant studies from psychology and economics before discussing the effects of two types of feedback (within- and across-class feedback), two types of rewards (monetary and non-monetary rewards), and their combinations (each feedback type interacted with each reward type), on students' performance and their well-being measured in terms of students' perceived stress and happiness levels. In Chapter 2, "Persistent Overconfidence: A Randomized Field Experiment in Uganda," I analyze the depth of overconfidence present among students and whether their self-assessment is affected by repeated feedback. Moreover, I contribute to the debate regarding the existence of the unskilled-and-unaware phenomenon.

The Education System in Uganda and the Experiment

Access to schooling has substantially increased in developing countries since the "Education For All" movement was launched in 1990. Uganda was one of the first African countries to introduce Universal Primary Education (UPE) in January 1997, and the initiative was expanded to secondary schools in 2007 (Universal Secondary Education, USE). As a consequence of elimination of tuition fees, student access to primary education increased by 27.7%, enrollment into secondary schools increased by 136% and the literacy rate improved to 74.6% (UNESCO, 2015).

The flip side of the success story is that many indicators show that improvements in quantity were not equaled by improvements in quality. In 2013 (according to World Bank

Development Indicators), only 56% of students completed primary school, giving Uganda the 8th lowest completion rate in the world. Only 29.4% of students completed lower secondary school in 2013 (the 7th lowest worldwide completion rate). More than 180,000 female and 297,000 male children of official school age were not enrolled in primary or secondary school. The pupil-teacher ratio (the average number of pupils per teacher) in primary school class is 46 – the 6th highest ratio of all 125 countries reported. The quality of Ugandan education remains poor.

Ugandan Education System

The academic year in Uganda starts in the 3rd or 4th week of January and finishes in late November/early of December. It consists of three trimesters separated by short holidays and a long holiday in December and January. Students in Uganda have free access to public primary and secondary schooling (due to UPE and USE). Public schools receive government funding based on the total number of students in each class. According to a 2015 UNESCO report, each primary school was supposed to receive 5,000 Ugandan Shillings (UGX) per year for each child in P1 – P3 and 8,100UGX for each child in P4 – P7. Government contributions to secondary schools was up to 141,000UGX per student. In both cases, parents still pay for uniforms, meals, and supplies. During the time the experiment was implemented, 1,000UGX was approximately 0.80USD. It represented approximately 0.4% of the monthly salary of a public primary-school teacher. For this sum, a student could buy one bottle of soft drink, three to four exercise books, one quarter of grilled chicken, three chapattis (a local salty pancake) or one “rolex” (rolled-eggs in chapatti), or approximately 0.25 liters of gasoline.

My data show that the average fee per term that public school students in the sample were asked to pay was 6,400UGX (excluding lunch) for P1–P3 students, 8,400UGX for P4-P6 students and 14,400UGX for P7. Lunch fees ranged from 4,000 to 5,000UGX. Most schools charged an admission fee which ranged from 1,000UGX to 5,000UGX[1]. While it is definitely not free education, this is significantly lower than the fees charged by private schools (the average fee per term was 29,400UGX for P1 – P3, 47,250UGX for P4 – P6 and 53,000UGX for P7). Lunch fees ranged from 10,000UGX to 35,000UGX. In both private and public schools, students in P7 had an option to attend remedial classes for fees from 500UGX to 35,000UGX. Sometimes students were asked to make additional payments, such as (re)construction fees, development fees, and contributions to the teachers' salary or rent. The tuition fees for secondary schools are approximately double the primary school fees.

Students are admitted to primary schools at the age of 6 or 7 (or exceptionally at 5). Very often students attend pre-school education (nursery section) starting from the age of 3 (86.3% of students indicated that they attended nursery). The official language in primary and secondary schools is English; however, especially in lower primary schools, children are often taught in the local language[2].

Students are supposed to pass each grade to qualify to enter the next higher grade. Only slightly more than half of the students in my sample (51.2%) had not repeated at least one grade. Successful completion of the Primary Leaving Examination (PLE) is considered successful completion of primary education. Since not all schools have the rights to conduct PLE

[1] I discussed the detailed scheme of fees during my personal meetings with headmasters and directors. Information was often publicly available.
[2] In Uganda there are 41 local languages. The common language in Mukono and Buikwe districts was Luganda.

examinations, students may register at a different primary school to sit the exam (sometimes students switch to a new school in the second or third term of P7). In 2011, students paid 11,000UGX to attend a PLE examination administered in English and consisting of four mandatory subjects – Math, English, Science and Social studies. From each subject, students receive a score from 1 to 9 (1 being the best). The scores are summed up and each student is placed into a category/division (the higher the sum, the worse the aggregated score). The best grade is therefore 4 (1 from each subject). Students pass the exam if they are placed in Divisions 1 to 4 (1 being the best[3]). Students who received higher scores are placed in Division U and recorded as failing. Absent students who paid the fee but did not participate fall into Division X. It is very important for each school to have at least one student in Division 1. Only students who passed PLE exams can be admitted to a secondary school and only students who scored below 28 in aggregated scores can be admitted to the Universal Secondary Education program.

Secondary schools have the right to set their own selection criteria when admitting new students to their first year (very often they set the minimum aggregate grade from the PLE examination in order to be admitted, which is higher than the PLE passing grade). Secondary education is divided into "O-level" (or lower secondary, from grades S1 to S4) and "A-level" (or higher secondary, grades S5 and S6). Only students who successfully pass the national examination in their S4 (Ugandan Certificate of Education, UCE) can continue to the higher A-level. In 2011, students paid registration fees of 68,000UGX to participate in the UCE exam, which includes eight compulsory subjects – English, Math, Biology, Chemistry, Physics, Geography, History and Literature. Grading follows a similar structure as the PLE exam. The best

[3] Students are placed in Division 1 if they scored between 4 and 12 aggregated points, in Division 2 if between 13 and 23, in Division 3 if between 24 and 29, and Division 4 between 30 and 34. If a student received more than 34 aggregated points, she is placed in Division U. Absent students are placed in Division X.

score in the UCE exam is therefore 8, while the worst is 72. After successful completion of the O-level, students choose a specialization – art or science – and proceed to the A-level. A-level studies are finalized by passing the Ugandan Advanced Certificate of Education (UACE), which consists of four taught subjects according to the specialization and a general paper. In 2011, the registration fee was 70,000UGX. Successful completion of secondary school is a necessary requirement to apply to university. Students can alternatively apply to a vocational school (even directly after primary school) or for alternative diplomas.

Students in all levels can repeat national examinations if they pay the registration fee. During the national examinations an external committee – consisting of teachers selected by the Ugandan National Examination Board from all participating schools - visits the school, conducts the exam and collects examinations for external evaluations. Precautions are taken to minimize opportunities to cheat, teachers helping their students, and teachers influencing the evaluation of exams. The exam questions are equal for all schools and the results are therefore comparable across all schools in Uganda.

The education system has many drawbacks. Students are not regularly informed about their performance. Only approximately 40% of students in my sample could describe their performance in their class. Headmasters often indicated that they lack resources to buy examinations for students. However, providing feedback to students may motivate them (especially girls) to improve their performance. Further, student absence rates are very high. The average absence rate of students interviewed and examined during the 2012 academic year was 29.2% (37.9% of students who were interviewed in 2011 changed schools or completely dropped out of school altogether). Reasons for absences and their lengths vary. Students were mostly absent for less than a term (77.4%), 17.7% missed 1 to 2 terms, 3.4% 1 to 2 years, and

1.5% more than 2 years. The main reasons for long-term absences were lack of money for school fees, help required by family members and sickness (their own or of family members).

Experimental Design

In this experiment, I study whether the provision of comparative feedback about their own performance and the performance of other group members can influence students' performance and psychological well-being measured by self-reported stress and happiness levels. To evaluate the effect of the intervention, I designed a Randomized Control Trial (RCT) experiment. Two types of feedback were offered – within-class and across-class feedback. Students randomized into the within-class feedback group were randomly divided into groups of three to four classmates within each class and evaluated as groups within the class. Group averages constituted the basis for performance comparisons. The students in the across-class feedback group were evaluated as a whole class (using the class average) and compared to other classes of the same grade in different schools. Comparisons were based on the average of the Math and English scores in the group.

Feedback differed in content across the treatment groups. Each student in the within-class feedback group received information about his Math and English scores, his/her group-mates' scores, the group average and the ranking of his/her group within his class. Furthermore, starting in testing round 3, each student received information about his/her (and his/her group-mates') improvement or decline from the two preceding testing rounds. Students in the across-class feedback group received information about how they scored in Math and English personally (they were not given information about their classmates), the class average and the

ranking of their class compared to other classes. The positions in both treatments were presented on a rank-order graph (see Appendix B1.4 and B1.5). Students in the control group received no information; they only took exams. Students were tested repeatedly during the 2012 academic year and received informational feedback three to four times depending on the feedback group (across-class/within-class feedback, respectively). Note that students in the across-class group (T2) first received feedback in testing round 3 (one-round delay compared to the within-class group students) due to logistical reasons. As shown in section 1.4, the effects of within- and across-class feedback are comparable.

Students were not offered rewards until testing round 4 was finished. In order to study the effects of monetary and non-monetary rewards, I orthogonally re-randomized the sample at school level[4] before the final school visit (three to four weeks in advance[5]) and introduced financial and reputational rewards (see also Figure 1). The randomization divided the sample into 9 groups – one control group, four sole treatment groups (i.e., one type of treatment only) and four combined treatment groups (two types of feedback interacted with two types of rewards). The scheme with all treatments offered is shown in Figure 3. Students were informed about the exact rules of the competition during our personal visit and also via posters we left in each class. Note that I can only study short-term effects of rewards since they were offered only once at the end of the academic year.

The aim of this cross-cutting design was to observe whether rewards could enhance student performance, especially if combined with within- and across-class feedback treatments

[4] The randomization was done at the school level in order to avoid spillover effects and possible confusion.
[5] Therefore, compared to other studies, students in this experiment had some time to adjust to the treatment (e.g., to prepare for the test).

(see also Figure 2) and whether student well-being would be affected. Students in the financial treatment groups could win 2,000UGX per person (which is approximately 0.80USD according to that day's exchange rate[6]). Students in the reputational reward scheme were promised that if they qualified for the reward, their names would be announced in the most popular local newspaper in the region, Bukedde. The qualification criteria differed based on original randomization into treatments (see Table 1) but the general rule was to reward the 15% best performing students/groups/classes, and the 15% most improved students/groups/classes[7]. In order to avoid confusion, students were given exact information regarding the number of

Table 1: Qualification criteria for rewards

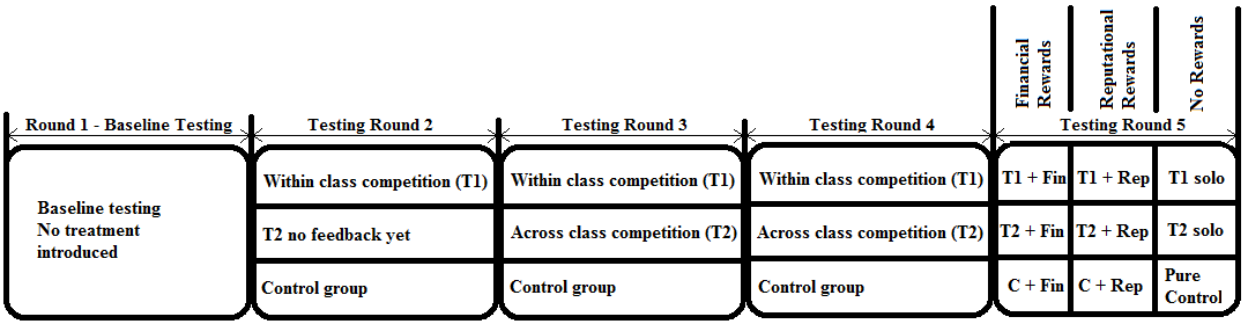
	Financial rewards (2000UGX)	**Reputational Rewards (Winners' names published in a local newspaper)**	**No rewards**
Within-class social comparison (Treatment 1)	15% of best performing and 15% most improved **groups** (524 students)	15% of best performing and 15% most improved **groups** (666 students)	Sole within-class social comparison group, no rewards (1205 students)
Across-class social comparison (Treatment 2)	15% of best performing and 15% most improved **classes** (409 students)	15% of best performing and 15% most improved **classes** (543 students)	Sole across-class comparison group, no rewards (1460 students)
Control group	15% of best performing and 15% most improved **students** (498 students)	15% of best performing and 15% most improved **students** (585 students)	Sole Control Group, no rewards (1260 students)

[6] For 2,000UGX, a student could buy, for example, two bottles of soft drink, a decent lunch in a canteen, three to four pens, two to three avocados, etc.

[7] In other words, if students were part of a within-class feedback group and competed for rewards, they would win the reward if their group scored in the top 15% of all groups or if they ranked among the top 15% of the most improved groups between testing rounds 4 and 5. If students were part of the across-class feedback group, the whole class would win the reward if the class was among the 15% top performing or 15% most improved classes. Finally, if students received no feedback, they would win the reward if they ranked among the 15% best performing or 15% most improved students in their class.

winning groups (if in a within-class feedback group), the number of winning classes (if in an across-class feedback group), or the number of winning students (if originally in a control group). I used percentages in order to guarantee a comparable number of winners across all treatment groups.

Figure 1: Orthogonal randomization of the sample into reward treatments



Timing and Logistics

The experiment was conducted between August 2011 and August 2013. The baseline survey was conducted between September and December 2011. In total, 8158 students answered questionnaires containing basic demographic questions, questions regarding family background and family composition, parental status, education and job, family wealth and additional questions regarding the students' interests, opinions, self-esteem and aspirations.

The intervention and the core data collection took place from January 2012 to December 2012. Students were tested twice per term, i.e., approximately every one and half months. In total, five testing rounds were conducted. Testing dates and times were arranged in advance by phone with the headmaster or the director of the school, and confirmed one day before testing. In general, three to four schools were visited every day, 5 times per week.

The agenda of each visit was similar. After we entered the class, students in feedback-treatment groups received their feedback while control students immediately started answering the questionnaires and exam questions. The order was as follows: "Before Math questionnaire", followed by a 30-minute Math examination, then "After Math Before English questionnaire", the English exam in the subsequent 20 minutes, and finally the "After English questionnaire". The core questions of these short questionnaires were related to students' expectations regarding how many points they thought they would earn on the Math and English examinations, how much effort they planned to put/they had put into answering the questions and the level of their current happiness. All questions were asked before and after each exam. No before-Math or before-English questionnaires were collected during the baseline survey since students saw the examinations for the first time.

During the academic year, students in the feedback groups received feedback in the form of a report card, which was glued into a small progress report book that each child in the treatment group received. My team members explained the content of the report card repeatedly to minimize the risk that students would not understand feedback content (also, the score cards were designed by students during our interviews in 2011). The books were stored at the schools, and headmasters promised to allow children to check their books at any point. The books contained necessary information to keep a child's attention and motivation levels active. After the experiment, students were allowed to keep their books.

Students were tested in Math and English. In order to ensure transparency, I used self-constructed tests based on questions students must answer on the Primary and Secondary Leaving examinations, which are developed and published by the Ugandan National Examination Board (available in bookstores). The selection of questions was tested in pilot sessions in schools

in Wakiso District which were not part of the final sample in the 2012 testing (for further details, see the next section). The level of difficulty was adjusted to grade curriculums and student proficiency. All tests were evaluated by myself and my team.

Table 2: Project timeline

Reward scheme introduced

2011 Baseline Survey		**2012**						**2013 Follow-up Session**
		Testing 1	**Testing 2**	**Testing 3**	**Testing 4**	**Testing 5**		
Students, teachers and headmasters interviewed	BREAK	Baseline testing from Math and English and question-naires; No treatment	**Within-class feedback group (T1)** received first treatment; **Across-class feedback group (T2)** no treatment	**Within-class feedback group (T1)** received treatment including improvement status **Across-class feedback group (T2)** received first treatment	**Within-class feedback group (T1)** received treatment including improvement status **Across-class feedback group (T2)** received treatment including improvement status	**Within-class feedback group (T1)** received treatment including improvement status **Across-class feedback group (T2)** received treatment including improvement status Chosen students competed to win prizes	BREAK	No treatment provided, students examined from Math and English;

Rewards disseminated

Final Sample

The project was designed in close cooperation with the Uganda Czech Development Trust (UCDT), an affiliation of the non-governmental organization Archdiocese Caritas Prague, Czech Republic, which has been running a sponsorship program "Adopce na dalku" in Uganda since

1993. According to UCDT representatives, students were enrolled in primary and secondary schools based on their own choices, therefore supported students should not differ from non-supported students in terms of their school choice. In 2011, UCDT sponsored students studying at 46 primary and 30 secondary schools located in 5 districts in Central Uganda – Mpigi (4 schools), Wakiso (9 schools), Mukono (14 schools), Buikwe (45 schools) and Buvuma (4 schools). Mpigi and Buvuma districts were excluded from my experiment from the beginning because in each district there were only 4 primary schools and no secondary schools[8].

During the baseline survey, my team and I visited 60 schools, including 34 primary and 26 secondary schools in Wakiso, Mukono and Buikwe districts. The baseline survey, however, showed that Wakiso district is different from Mukono and Buikwe in terms of the demographic characteristics of its students, as it encircles the capital city, Kampala. Time and budget constraints were other reasons to exclude Wakiso from the sample.

The final sample consisted of 52 schools (31 primary and 21 secondary) of which 19 are public, 23 are private and 10 are community schools (community schools are similar to private schools but are founded by a community as opposed to by an individual entity). All schools were located in rural areas. Initially there were 53 schools in my sample; one decided not to participate after I conducted the baseline survey. This school was initially randomized into the control group and its exclusion did not lead to significant differences in terms of the baseline observables. The headmasters of the remaining 52 schools agreed to participate in the experiment. The headmasters had an option to withdraw from participation at any time during the experiment, nonetheless no school opted to do so. I asked the headmasters to communicate

[8] It is also worthwhile to note that Mpigi is the only district located South-west of Kampala and Buvuma is an island district.

the content of the project to parents during regular parental meetings. In addition to the headmasters' consent, I also had the full support of UCDT (the letter of accordance appears in Appendix B1.3[9]). In order to minimize possible costs from our presence at schools, the duration of meetings was set to a maximum of 120 minutes. All meetings were organized with the headmasters one week in advance to find the most suitable and least harmful time in terms of the curriculum delivered. Administering exams in Math and English was supposed to serve students as additional training for the leaving examinations they face during the final years of their studies in primary (grade 7) and lower secondary (grade 4) schools.

In total, 146 classes[10] (P6 and P7 in primary schools, S1 up to S4 in secondary schools) ammounting to more than 5,000 students were repeatedly tested. Based on the power calculations using Optimal Design software (Raudenbush, S. W., et al., 2011) such a number of classes is sufficient to detected effect size of 0.15 standard deviations. Treatment effects that are lower than 0.15 standard deviations may or may not be detected, depending on the standard errors. The calculation accounts for stratification and clustering at the higher level. A figure plotting effect size with respect to the total number of clusters can be found in Appendix A1.4.

In addition to Math and English scores, I collected information about students' reported immediate effort, their strategic effort in preparation for the exams and their happiness level, measured immediately before and after each exam. I also repeatedly inquired about student expectations of their own scores from the Math and English tests in order to measure their confidence. To study students' well-being, I collected data on their happiness based on the

[9] There is no Institutional Review Board (IRB) for social sciences in the Czech Republic which could issue an IRB approval for my experiment. The experiment was designed in line with the conventions of IRB standards.
[10] If a school had more than one class of P6 – P7 or S1 – S4, all classes were included in the testing. Students in P1 – P5 were not included because they repeatedly failed to understand the instructions in the pilot testing.

Subjective Happiness Scale (Lyubomirsky and Lepper, 1997) and subjective stress based on the Perceived Stress Scale (Cohen et al., 1983). The happiness score is calculated as a sum across four questions using a 7-point Likert scale (with 1 being maximum and 7 minimum). Similarly, stress scores are based on the answers from four questions from a 5-point Likert scale in which 1 equals no stress and 5 is maximum stress. The questionnaires can be found in Appendices B1.1 and B1.2. In addition to student-level data, I also collected information regarding school (school type, school area, school fee structure and school equipment), headmasters and teachers (demographic information, years of experience, salary and their opinions on education).

Due to large attrition between 2011 and 2012 and the admission of new students throughout the 2012 academic year, detailed information collected in 2011 is available for only about 52% of students who participated in the 2012 experiment. In every testing round during the academic year 2012 it happened that some students got sick during the testing (mainly malaria) or stole the examinations, which resulted in an unequal number of Math and English exams available. The total number of such cases is between 0.1 and 0.3%. Excluding these students does not change the results. Some students failed to correctly answer questions in the questionnaires and either marked more than one option (if only one was possible) or forgot to answer all questions. This results in an unequal number of observations, e.g., in the effort exerted into Math or English exam, subjective happiness or the expected number of points. The total number of such cases does not exceed 1%. The crucial difference in the number of observations is between the number of students who completed baseline Math and English exams and those who completed baseline happiness and stress questionnaires. Due to logistical issues, happiness and stress questionnaires were collected at the very beginning of the second testing round before any feebdack had been distributed. Therefore, 19% students who were

present in testing round 1 were not present in round 2. In order to see to what extent the treatment effects differ, I compared the estimations of the treatment effects from regressions conditioned on students' presence in testing round 2 to regressions conditioned on their absence[11]. The results are similar in size with lower standard errors. A Kolmogorov-Smirnov test resulted in students present in the first two testing rounds and those present in the first but not the second testing round coming from the same distribution.

Stratification and Randomization

In order to increase the balance between control and treatment groups, the sample was stratified along three dimensions – school location (the sample was divided into four areas differing by level of remoteness), average school performance in national testing (above average or below average) and student level (grade 6 and 7 of primary education and grades 1 to 4 of secondary education). Within each strata, I randomized the sample into treatment and control groups. The randomization was done in two stages (as shown in Figure 3). First, after the stratification of the sample by school performance and area, I randomized the whole sample of 53 schools into treatment and control groups in a 2:1 ratio. The randomization was performed at the school level and resulted in 36 treatment and 17 control schools. School-level randomization in the first stage was chosen in order to minimize control group contamination due to information spillovers. In the second stage, I divided classes of the treatment schools randomly into within- class feedback (T1) and across-class feedback groups (T2) in a 1:1 ratio (class-level randomization). In this scenario, no student in a control-group school received any treatments,

[11] Note that the dependent variable in the regression is endline performance of students and I control for the baseline performance.

and students in the treatment-group schools received either within- or across-class feedback depending on the type of intervention their class was randomized into. Overall, 1/3 of the sample is the control group, 1/3 is treatment group 1 (T1) and 1/3 is treatment group 2 (T2). Exposure to the treatment is the only difference in the outcomes between the control and treatment groups.

Figure 2: Map with coordinates of schools participating in the study

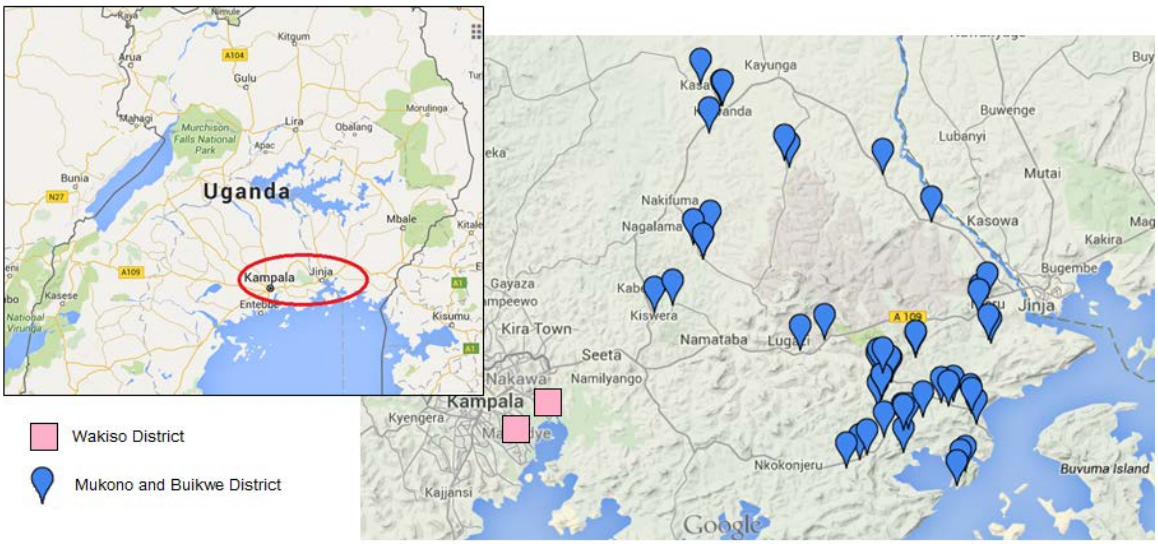
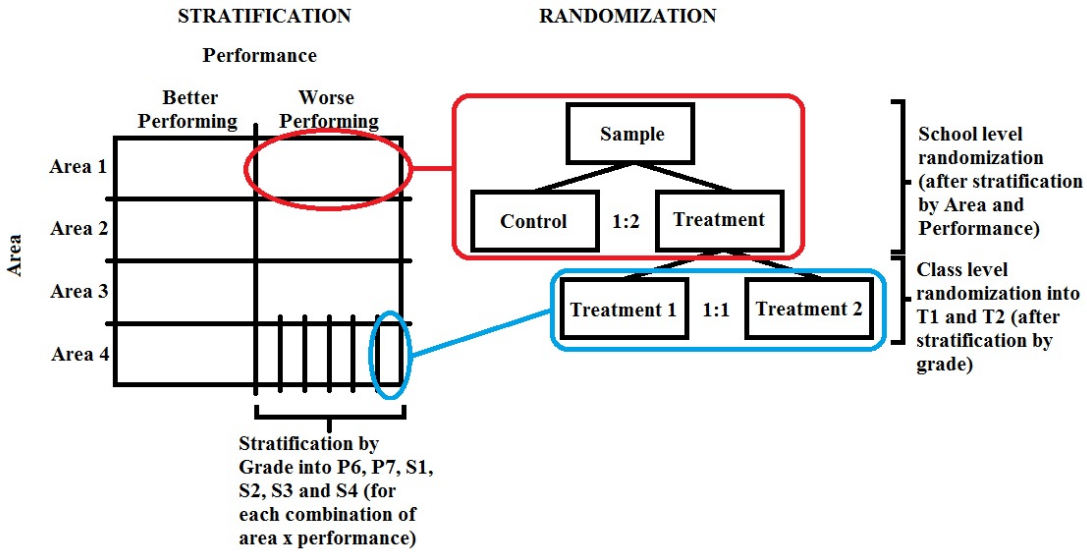


Figure 3: Stratification and randomization



1 Dark Side of Incentives in Schools: Evidence from a Randomized Field Experiment in Uganda

1.1 Introduction

A trophy for the best student in a class, a certificate for the most improved learner, or a bonus payment for the employee of the month; we are routinely faced with incentives of different types (symbolic, reputation, or financial rewards) throughout our lives. Rewards are often believed to motivate subjects and subsequently improve their performance, and are therefore implemented in many different environments (Lazear, 2000; Fryer, 2010; etc). We are also routinely compared to classmates, colleagues, or other competitors by receiving relative feedback about our performance, which can also improve performance. Feedback may motivate subjects to improve their performance (Andrabi et al., 2014; Azmat and Iriberri, 2010) though the evidence of such positive effects is more scattered.[12] Feedback and incentives may also influence our well-being (Azmat and Iriberri, 2016) and changes in well-being may further influence people's decision-making and economic outcomes (e.g., Juster et al., 2010; for more details see section 1.2 Literature Review).

The current work is a unique study implemented in the field that analyzes the effects of various types of motivation schemes on performance and well-being, measured by perceived stress and happiness of students evaluated in groups. Its main contribution comes from explicitly accounting for the performance-versus-well-being tradeoff introduced by incentives.

[12] According to psychologists, positive feedback is believed to increase intrinsic motivation and foster long-term motivation, whereas negative feedback decreases intrinsic motivation (Burgers et al., 2015; Arnold, 1976; Deci, 1972). A short description of the extrinsic and intrinsic motivation can be found in section 1.4.7.

The novelty of the experiment comes from the wide scope of outcome measures observed, its rich experimental design and its unique data set. The sample size consists of more than 5000 primary and secondary school students from 146 classes in Southern Uganda who were repeatedly tested and interviewed over a full academic year in 2012. In total, five testing rounds were administered. The design offers a direct comparison of the effects of two feedback types and two reward types as well as their combinations (each feedback type interacted with each reward type).

Feedback differed across feedback-treatment groups with respect to its content. Each student in the *within-class feedback group* (class randomly divided into groups of three to four students) received information about how he/she scored in Math and in English, how his group-mates scored and the position of the group within his/her class. Students in the *across-class comparative feedback group* (comparisons of entire classes) received information about how they scored in Math and in English personally but were not given information about their classmates and the position of their class compared to other classes.

Students were not offered rewards until testing round 4 was finished. They were then orthogonally randomized into financial, reputation and no-reward groups. Students in the financial reward group could win 2,000UGX per person (approximately 0.80USD according to that day's exchange rate). Students in the reputational reward group were promised that if they qualified for the reward, their names would be announced in Bukedde, the most popular regional newspaper, and they would receive a certificate. The general criterion used was to reward 15% of the top performing and 15% of the most improving students/groups/classes.

The results show that students improved their performance in response to feedback or reward provision. The improvements are mild in terms of size (0.08 to 0.13 SD) but comparable

to existing literature. The improvements are, however, significantly higher when students received a combination of feedback and rewards (up to 0.28 SD). While feedback and reputational rewards motivated students to improve only in Math (no improvement in English), financial rewards led to comparable improvements in both subjects.

The results for outcomes other than learning, i.e., happiness and stress, counterweight the benefits of providing rewards. Students who were offered only rewards (without any feedback) reported elevated stress levels and decreased happiness, whereas the well-being of students who received only feedback remained unchanged. Moreover, most of the treatment combinations led to a decrease in students' stress and an increase in or no effect on happiness. Thus, we can speak of an important trade-off: the introduction of rewards increases performance more than feedback alone, but at the same time they lowered students' well-being. The effects persist when I control for multiple comparison testing by adjusting the p-values using the Simes step-up method (Simes, 1986).

In some experiments, boys and girls responded very differently to certain incentives. The second major contribution of this paper is to shed light on the underlying reasons for these gender differences. I find that if girls were given rewards but no group feedback, they significantly underperformed boys. If girls were repeatedly informed about their performance and performance of their groups, however, no matter what type of feedback they received, their performance improved and became comparable to that of boys. In other words, comparative feedback in a tournament environment played a crucial role in motivating girls to improve their performance. Boys, by contrast, reacted only to rewards.

The current design of the experiment does not allow me to distinguish whether gender differences were caused by the fact that students were evaluated in groups (group identity

effect) or were repeatedly informed about their standing. Nevertheless, since both within- and across-class feedback groups delivered similar effects, it seems more likely that the effect is driven by social comparison rather than by group identity. Such a result would be in line with "reference group neglect", i.e., students neglect information about others and focus solely on feedback regarding their own performance (Camerer and Lovallo, 1999).

1.2 Literature Review

According to social comparison theory[13], informing a child about his/her performance without comparing it to other children causes inappropriate evaluations of the child's ability and can influence effort negatively (Festinger, 1954;[14] founder of the social comparison theory). On the other hand, comparison enables a child to find his/her relative position within a particular group, which, via enhanced competitiveness, can lead to an increase in effort and subsequent improved performance.

Feedback provision, as a way to inform subjects about their absolute or relative standing, has been analyzed in different environments and has delivered opposing results. Andrabi et al., (2014), for example, provided parents, teachers and headmasters with report cards informing them how children were doing in a particular school. The intervention resulted in 0.1 SD improvement in student test scores. Azmat and Iriberri (2010) informed high school students about their relative standing, resulting in a 5% improvement in grades. Additionally, university

[13] Social comparison theory is about "our quest to know ourselves, about the search for self-relevant information and how people gain self-knowledge and discover reality about themselves" (Mettee and Smith 1977, p. 69–70).

[14] Festinger in his original paper focused on the social comparison of abilities and opinions only. Since then, however, many different dimensions of social comparison have been studied (e.g., Buunk and Gibbons, 1997 and 2000; Suls and Wheeler, 2000). See for example Locke and Latham (1990); Suls and Wheeler (2000), for an overview of papers in psychology and management science. See Buunk and Gibbons (2007) for an overview of work in social comparison and the expansions of research on social comparison.

students in the United Kingdom responded positively and improved their performance by 13% in response to feedback regarding their own absolute performance (Bandiera et al., 2015).[15] On the other hand, not all studies find positive responses to feedback provision. Azmat et al., (2015) do not find any effect of relative feedback on performance among students at University Carlos III in Madrid, Spain. In the short period after feedback was provided they find a slight downturn in student performance. More evidence of negative effects of incentives on performance can be found in experiments implemented in the workplaces. Workers in a crowd-sourcing experiment (using Amazon's Mechanical Turk crowd-sourcing webpage) lowered their performance after they received information about their rank position (Barankay, 2011). Health workers also decreased their performance during a training program in Zambia when exposed to social comparison (Ashraf et al., 2014).[16]

The effect of feedback depends on to whom the subjects are compared, how they are compared and whether they are rewarded for their performance. Students face social comparison in their classrooms on a daily basis, and it can strongly influence their self-esteem and their performance (Dijskstra et al., 2008) as well as their well-being (Azmat and Iriberri, 2016). It is therefore important to understand with whom to optimally compare students. If students are compared to those who are slightly better, their effort and performance tend to

[15] Tran and Zeckhauser (2012), Blanes-i-Vidal and Nossol (2010) and Fryer (2010) are examples of other studies with positive effects of feedback provision.

[16] There are also controlled lab environments studying the effects of feedback provision, e.g., Falk and Ichino (2006) and Mas and Moretti (2009) which have found that if one lets people observe the behavior of their peers, their performance improves. Kuhnen and Tymula (2012) and Duffy and Kornienko (2010) find a positive effect to the provision of private feedback. Eriksson et al. (2009), on the contrary, find that rank feedback does not improve performance (even if pay schemes were used). Hannan et al. (2008) find a negative effect of feedback on relative performance under a tournament incentive scheme (if feedback is sufficiently precise).

increase[17]. Students can be compared individually or in groups where a group's outcome depends on each member's contribution, which may foster mutual help (Slavin, 1984), in addition to positive peer effects (Hoxby, 2000; Sacerdote, 2001). Groups can be formed endogenously, e.g., by students themselves based on friendship, or exogenously (Blimpo, 2014). In some studies, the effects of interventions are more pronounced if students are involved in tournaments (Eriksson et al., 2009; Bigoni et al., 2010; VanDijk et al., 2001).[18]

Subjects often improve their performance if they are rewarded financially. Bettinger (2012), Angrist et al. (2002, 2006, 2009, 2010), Kremer (2004), Bandiera (2010), and others studied the effects of providing cash payments, vouchers or merit scholarships to students who successfully completed a pre-defined task. In such experiments, knowing their relative position is not crucial since success does not depend on the performance of others.

In order to induce stronger competitive pressure, subjects need to be put into a tournament with a limited number of winners. VanDijk et al. (2001), based on an experiment comparing different payment schemes, conclude that it is better for a firm to introduce a tournament-based scheme over a piece-rate or group payment scheme. In the case of Blimpo (2014), groups involved in a tournament improved by approximately the same amount as groups rewarded for reaching a performance target. All treatments (with or without competition) resulted in positive improvement in student performance, which increased between 0.27 to 0.34 SD. Not only positive treatment effects have been found. Fryer (2010) and Eisenkopf (2011) studied the impact of different financial rewards on student performance and

[17] Ray (2002), using a theoretical model, shows that performance and effort decrease if the comparison target is too far from a student's ability.
[18] See Hattie and Timperley (2007) for a review of the literature on the provision of feedback.

did not find any significant improvements (although Fryer (2010) claims that the effect might not have been detected because of lack of statistical power).

Even when financial rewards result in performance improvements, they may not necessarily be cost-effective (e.g., Bandiera et al., 2015)[19]. Alternative rewards[20] that could possibly be more cost-effective have drawn researchers' attention. For example, Kosfeld and Neckerman (2011) designed a field experiment where students in the treatment group were offered symbolic rewards (a congratulatory card) for the best performance while students in the control group were offered nothing. Their results provide strong evidence that status and social recognition rewards have motivational power and lead to an increase in work performance (by 12% on average). Subjects in a real-effort experiment conducted by Charness et al. (2010) increased their effort in response to the relative performance and expressed their "taste for status". Jalava et al. (2015) offered sixth grade students in primary schools different types of non-monetary rewards (criterion-based grading, a certificate or a prize in the form of a pen if they scored among the top 3 students). The effects were heterogeneous with respect to original ability (students from the two middle quartiles responded the most to the incentives) and with respect to gender (boys improved their performance in response to rank-based incentives only, girls also improved when given symbolic rewards). Rank-based grading and symbolic rewards, however, crowded out intrinsic motivations of students.

If non-monetary rewards have the power to motivate subjects to improve their performance, then naturally, questions arise: what can we learn from direct comparison of

[19] Bandiera et al. (2012) find the financial rewards cost-ineffective since only a fraction of the students from the second quartile of initial ability distribution react positively to financial rewards.

[20] See also theoretical models studying the effects of reputation and symbolic rewards on subjects' performance in Weiss and Fershtman (1998), Ellingsen and Johannesson (2007), Besley and Ghatak (2008), Moldovanu et al. (2007) and Auriol et al. (2008).

monetary and non-monetary rewards? Would financial rewards prevail? Levitt et al. (2012) present the results of a set of field experiments in primary and secondary schools, in which they provided students with financial and non-financial rewards, with and without delay and with incentives framed as gains and losses. In terms of performance change the experiment showed that for younger students both monetary and non-monetary rewards brought similar results and therefore non-monetary rewards were more cost-effective[21].

Feedback and incentives may also influence psychological well-being (Azmat and Iriberri, 2016). Change in well-being has been found to influence people's decision making and economic outcomes. An increase in happiness[22] is associated with better health, sharper awareness, and higher activity in addition to better social functioning (Veenhoven, 1998). Education is one determinant of happiness with higher education associated with greater well-being (Helliwell et al., 2012; Dolan et al., 2008).

Subjects under stress make suboptimal decisions, which, in the case of students, could lead to incorrect answers during examinations, or suboptimal choices in their activities (e.g., to be absent from school, to drop out of school or to exert low levels of effort). Both stress and happiness influence subjects' health (Juster et al., 2010; McEwen, 2008; Schneiderman et al., 2005). Stress can influence learning and memory creating learning problems (Lubin et al., 2007; Wolf, 2009). In the extreme, stress hormones may even influence brain structure (Lupien et al., 2009).

[21] They also found that rewards provided with delay lose their motivational power, and that it depends whether the rewards are framed as gains or losses (the second alternative being more robust).
[22] See Fordyce (1988) for a review of happiness measures and MacKerron (2012) for a review of the economics of happiness; Dolan et al. (2008) review well-being.

The current experiment differs from existing studies in the complexity of incentive schemes implemented and its broader scope of outcomes. In adition to performance commonly used as a dependent variable, I study students' confidence, stress, happiness and their academic aspirations. The results of the existing literature suggest a possible trade-off between performance and change in well-being. Evaluation of students in groups should enhance cooperation within groups and lead to group average improvements. If the group is big enough, however, free-riding behavior may prevail and result in heterogeneity within the group outcomes. Informing students about the position of their group could either lead to improvements in performances via enhanced competition or demotivate students with a negative attitude toward competition. Alternatively, students could neglect information about their group members and focus solely on their own performance (Camerer and Lovallo, 2002). The effect potentially depends on group gender and/or ability composition (Apesteguia et al., 2012) and group position in the group ability distribution. Students included in both financial and reputational reward treatments are expected to improve their scores, at least those in the second quartile of ability distribution. Students involved in a competition may experience increased stress levels and it is a question whether "short term pain" can bring "long term gain" and what the consequences of decreased well-being may be.

1.3 Baseline Summary Statistics

Data on student performance, demographics and student responses to questions suggests that randomization divides the sample into groups that are similar in expectations (see Tables 1.1 and 1.2 below, and Appendices A1.1 to A1.3 for the treatment-control group comparisons).

Table 1.1: Randomization balance: by feedback treatment status

	Means			Mean Differences		Joint P-value
	Within-class Feedback (T1)	Across-class Feedback (T2)	Control (C)	(T1 – C)	(T2 – C)	
PERFORMANCE (Baseline)						
Math	8.063	8.838	8.655	-0.564	0.197	0.183
				(0.435)	(0.414)	
English	14.072	14.630	14.432	-0.359	0.198	0.699
				(0.584)	(0.528)	
Sum Math + English	22.134	23.468	23.088	-0.923	0.395	0.426
				(0.957)	(0.886)	
OTHER THAN PERFORMANCE						
Gender	0.539	0.516	0.517	0.022	-0.001	0.239
				(0.015)	(0.014)	
Age	17.058	17.049	16.999	0.059	0.049	0.737
				(0.079)	(0.078)	
Average class size	43.912	47.245	43.337	0.575	3.908	0.546
				(3.208)	(3.776)	
Expected number of points from Math	4.331	4.536	4.552	-0.221	-0.015	0.299
				(0.150)	(0.145)	
Expected number of points from English	5.715	5.757	5.796	-0.081	-0.039	0.879
				(0.161)	(0.144)	
Perceived difficulty of Math exam	3.341	3.495	3.423	-0.082	0.072	0.030
				(0.053)	(0.052)	
Perceived difficulty of English exam	3.644	3.644	3.677	-0.033	-0.033	0.752
				(0.052)	(0.049)	
Immediate happiness after Math exam	3.287	3.226	3.132	**0.155***	0.094	0.230
				(0.092)	(0.092)	
Immediate happiness after English exam	2.909	2.869	2.782	0.127	0.087	0.303
				(0.085)	(0.085)	
Effort put into Math exam	3.447	3.535	3.504	-0.057	0.021	0.298
				(0.053)	(0.052)	
Effort put into English exam	3.547	3.627	3.553	-0.006	**0.074***	0.141
				(0.046)	**(0.044)**	
Subjective stress	1.504	1.588	1.439	0.065	**0.149*****	0.001
				(0.041)	**(0.036)**	
Subjective happiness	2.869	2.913	2.806	0.064	**0.107***	0.155
				(0.058)	**(0.055)**	
Education over work	3.538	3.496	3.477	0.060	0.019	0.526
				(0.057)	(0.059)	
Education over relax	3.834	3.756	3.778	0.056	-0.021	0.269
				(0.049)	(0.049)	
Work over relax	2.766	2.701	2.803	-0.037	-0.102	0.524
				(0.094)	(0.090)	

Note: comparison of mean characteristics of students in treatment and control groups. T1 (T2) stands for within-(across-) class social comparison groups and C for control group. Robust standard errors clustered at class level are in parentheses, adjusted for stratification. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 1.2: Randomization balance: by reward treatment status

	Means			Mean Differences		Joint P-value
	Financial Reward **(Fin)**	Reputation Reward **(Rep)**	No Rewards **(No)**	**(Fin – No)**	**(Rep – No)**	
PERFORMANCE (Baseline)						
Math	9.92	10.72	10.49	-0.507	0.292	0.788
				(1.215)	(1.258)	
English	10.796	10.394	10.853	-0.096	-0.497	0.937
				(1.211)	(1.418)	
Sum Math + English	20.718	21.116	21.353	-0.603	-0.206	0.955
				(2.199)	(2.660)	
OTHER THAN PERFORMANCE						
Gender	0.553	0.545	0.514	**0.039****	**0.031***	**0.089***
				(0.019)	**(0.017)**	
Age	14.376	14.196	14.437	-0.088	-0.267	0.798
				(0.350)	(0.359)	
Average class size	45.434	55.137	45.987	0.388	**10.087****	0.067
				(4.173)	**(4.332)**	
Expected number of points from Math	4.839	4.964	4.917	-0.028	0.096	0.882
				(0.257)	(0.236)	
Expected number of points from English	5.152	5.132	5.162	0.018	-0.002	0.994
				(0.255)	(0.276)	
Perceived difficulty of Math exam	3.283	3.361	3.256	0.043	0.121	0.499
				(0.088)	(0.093)	
Perceived difficulty of English exam	3.407	3.417	3.398	0.017	0.027	0.975
				(0.080)	(0.085)	
Immediate happiness after Math exam	2.616	2.633	2.713	-0.108	-0.091	0.793
				(0.159)	(0.145)	
Immediate happiness after English exam	2.504	2.529	2.548	-0.053	-0.029	0.911
				(0.099)	(0.112)	
Effort put into Math exam	3.617	3.631	3.684	-0.053	-0.040	0.722
				(0.083)	(0.086)	
Effort put into English exam	3.526	3.489	3.554	-0.027	-0.064	0.603
				(0.055)	(0.064)	
Subjective stress	6.849	6.799	6.883	-0.034	-0.084	0.912
				(0.230)	(0.192)	
Subjective happiness	10.376	10.933	10.671	-0.324	0.234	0.226
				(0.343)	(0.278)	
Education over work	3.822	3.735	3.910	-0.088	-0.176*	0.209
				(0.111)	(0.098)	
Education over relax	4.234	4.266	4.296	-0.053	-0.020	0.748
				(0.081)	(0.077)	
Work over relax	2.767	3.250	3.141	**-0.369*****	0.115	**0.001*****
				(0.125)	(0.102)	

Note: comparison of mean characteristics of students in treatment and control groups. Fin(Rep) stands for financially (reputationally) rewarded groups, and No represents the control group with no rewards. Robust standard errors clustered at class level are in parentheses; adjusted for stratification. * significant at 10%; ** at 5%; *** at 1%.

Few significant differences can be observed between the across-class feedback and control groups, although students in the across-class feedback group were slightly more stressed, slightly less happy and exerted slightly more effort compared to the control group. If the covariates are correlated with student performance, such an imbalance could bias the estimation of the treatment effect of the intervention (Firpo et al., 2014). One can expect some imbalances between treatment and control groups to occur purely by chance - as the number of balance tests rises, the probability of rejection of the null hypothesis of no difference between treatment and control group also increases. In my case, treatment and control groups differ significantly in less than 5% of all cases.

The average student scored 8.06 points out of 50 in the Math exam and 14.07 points out of 50 in English. In most of the cases, the real scores are below students' expectations. A student's miscalibration of his/her own performance is approximately 100%. The average student put "a lot of effort" into answering the exam questions (intensity level 4 in the 5-point Likert scale) and seems to be "very happy" according to the immediate happiness scale (intensity level 2 in the 7-point Likert scale where 1 is the maximum). The average student finds the Math exam of comparable difficulty to the regular exams at school and the English exam easier. Overall the average student is quite happy (based on the Perceived Happiness Scale) and has a low level of stress (Perceived Stress Scale).

1.4 The Effects of Incentives on Students' Performance and Their Well-Being

The core question of the experiment is how different incentive schemes (social comparison, financial and non-financial rewards) influence student performance and well-being.

I first analyze the aggregated treatment effects (i.e., the overall treatment effect of the within- or across-class feedback in each testing round and the effects of financial and reputational rewards in the final testing round (see Appendix C1.1 for details regarding calculation of the treatment effect). Later, I disentangle the sole treatment effects (only feedback and only rewards) and the interaction effects (each type of feedback interacted with each type of reward). I discuss the role of group gender and ability composition and I study whether the type of feedback students receive matters for improvement. Finally, I look at the distributional analysis.

1.4.1 Average Treatment Effects on Students' Performance

Repeated provision of feedback (pooled or separately by feedback type) increased students' overall performance by 0.07 SD, which is typically considered a small effect in the education literature. In other words, the average student who received within-class or across-class feedback scored higher than 52.8% of students in the control group. The type of feedback does not play a significant role. The results are similar once I separate sole effects of feedback provisions from those interacted with rewards (as shown in equation 3 in Appendix C1.1). Table 1.3 summarizes the aggregated average treatment effects of feedback and rewards on students' overall performance in columns 1 and 4, and on their subjective well-being, i.e., happiness and stress in columns 5 and 6. For aggregated average treatment effects on Math and English separately see Appendix C1.2. The effects are expressed in SD.

The results are very similar to the results of Jalava et al. (2015), who tested the effects of different grading schemes on primary school students in Sweden and found 0.077 SD for criterion-based grading, and 0.080 SD for tournament grading. Similarly, in a study in Pakistan,

parents and teachers received report cards regarding the performance of their children/students, which led to a 0.1 SD increase in student performance (Andrabi et al., 2014).

In the current study, students also improve in response to the provision of rewards. In aggregated terms, students who compete for financial (reputational) rewards score 0.176 (0.102) SD higher than control-group students. Again, Jalava et al. (2015) find similar results. In their study, students who competed for a certificate improved their performance by 0.083 SD. Students who competed for (non-monetary) prizes improved by 0.125 SD. Blimpo (2014) studied the effects of financial rewards provided to students in Benin on an individual or

Table 1.3: Aggregated average treatment effects of the provision of feedback and rewards on the overall performance and students' subjective well-being

Dependent variable:	STUDENTS' OVERALL PERFORMANCE				STRESS	HAPPINESS
	(1)	(2)	(3)	(4)	(5)	(6)
Aggregated feedback treatment pooled	**0.073***	0.064				
	(0.041)	(0.039)				
Aggregated reward treatment pooled		**0.133****				
		(0.052)				
Within-class social comparison, aggregated			0.074	0.061	-0.001	**-0.111***
			(0.045)	(0.043)	(0.090)	**(0.058)**
Across-class social comparison, aggregated			0.071	0.069	-0.104	-0.058
			(0.047)	(0.046)	(0.082)	(0.057)
Financial Rewards, aggregated				**0.176*****	**0.226****	-0.108
				(0.062)	**(0.107)**	(0.070)
Reputational Rewards, aggregated				**0.102****	0.177	-0.112
				(0.051)	(0.119)	(0.070)
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.714	0.715	0.634	0.645	0.058	0.078
F-statistics	424.45***	430.01***	389.52***	389.69***	4.56***	16.66***
N	5108	5108	5102	5102	4105	4056

Note: OLS. Pooled aggregated feedback treatment effect consists of the aggregated treatment effects of within- and across-class feedbacks. Pooled aggregated reward treatment effect consists of the aggregated treatment effect of financial and reputational rewards. Columns (1) – (4) show the aggregated average treatment effects (ATE) of differently aggregated treatments on students' overall performance. Columns (5) and (6) represent the average treatment effects on students' well-being (stress and happiness respectively). Controlled for stratum fixed effects (area, level and school performance in national examinations).
Full table with coefficients for stratification variables can be found in Appendix C1.10. * significant at 10%; ** at 5%; *** at 1%.

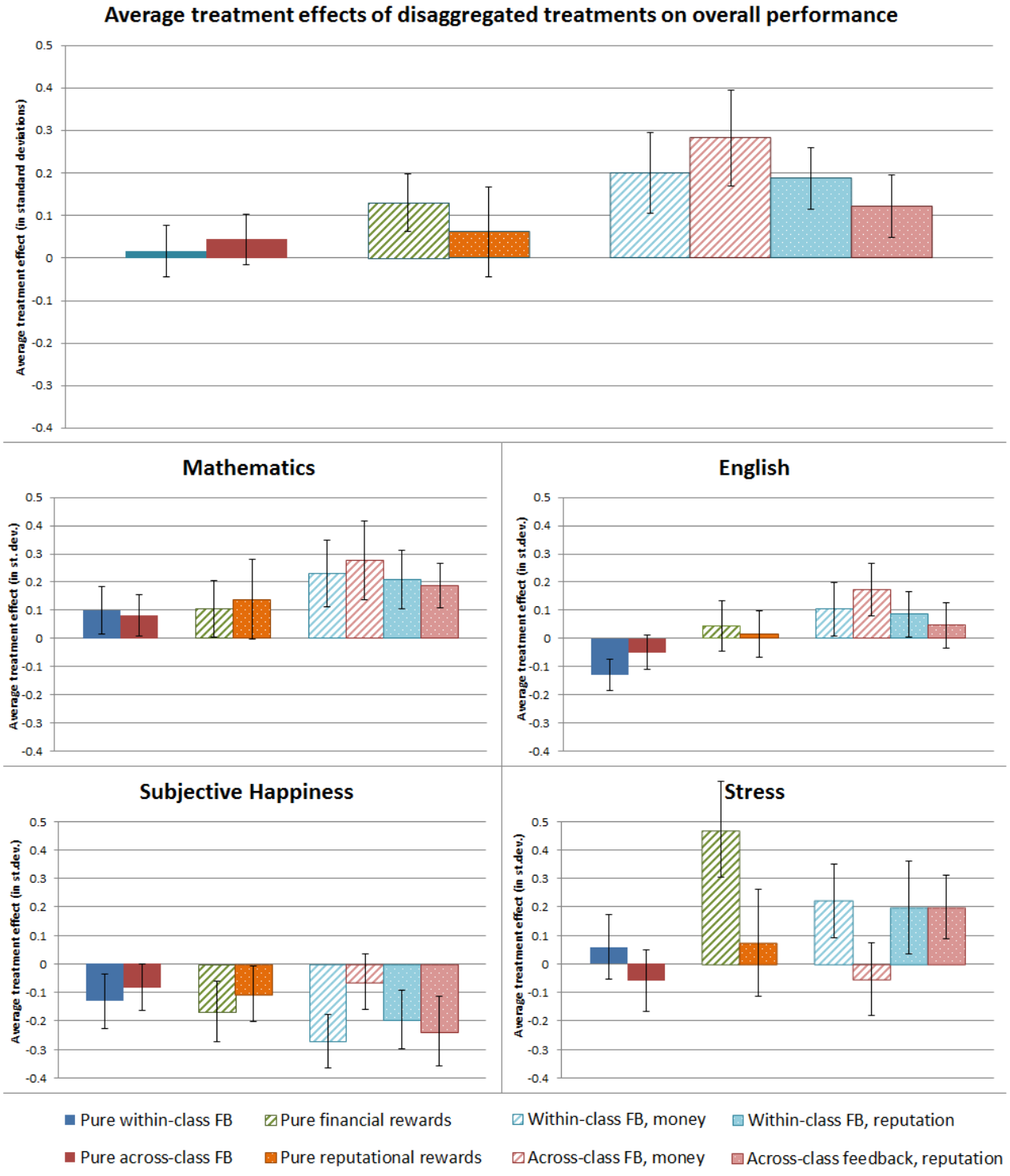
group basis. Students improved their performance by 0.27 to 0.34 SD. Blimpo's (2014) treatment effects are comparable in size to the results of the current study once I decompose the treatment effects into only feedback, only rewards and feedback-reward interaction treatments.

Being informed (in the form of repeated feedback) seems to play a role in the effect size of the rewards. While rewarded students without feedback improved by 0.06-0.13 SD, students exposed to repeated feedback improved their performance by 0.12 to 0.28 SD (see Figure 1.1). In other words, students who received repeated feedback and competed for rewards outperformed the control group by 2.4 to 6.2 percentage points more compared to students who competed only for rewards.

One caveat must be considered when I compare the effects of rewards to the effects of feedback. While feedback was provided to students repeatedly during the whole academic year, students competed for rewards only once[23]. In order to compare the immediate average treatment effects of feedback to average treatment effects of rewards, I compare the estimates from Round 2 (Round 3 for across-class feedback) controlling for the baseline performance to estimates from Round 5 controlling for the performance of students in Round 4. These comparisons should capture the differences in immediate responses of students to the feedback and rewards. For further details regarding alternative comparisons see Appendix C1.3, 1.4 and 1.8. Students who competed for financial rewards significantly outperformed students who received feedback in all scenarios. Students who competed for reputational rewards scored similarly to students in feedback treatment groups in most scenarios compared.

[23] It would be interesting in future research to observe how students' performance and their well-being would change if they faced competition for rewards repeatedly.

Figure 1.1:Dis-aggregated average treatment effects of incentives on students' performance and their subjective well-being



The effects on students' performance differ in Math and English. While the effects of feedback are driven solely by improvements in Math, rewards lead to similar improvements in both subjects as can be seen in Table 1.4. One explanation is that Math is more elastic (Bettinger, 2012), in the sense that it may be easier to detect the areas of Math in which the student is failing, e.g. multiplications, and focus on improvements in such area, while in English it may be hard to detect problematic areas and prepare for the test. If this is the case in the research presented here, the patterns should be, however, similar across all treatment groups, which is not the case. An alternative explanation comes from the overall motivation. Students may have low motivation to study science, which is often perceived as more difficult[24] and students may not see its usefulness in real life; but once they are incentivized (students see real rewards instead of abstract future benefits), they improve. Finally, rewards may have stronger motivatory power for subjects to remain in a competition compared to the feedback, which, in combination with order effect (the Math examination always preceded the English examination), could explain why subjects with rewards improve in both subjects while students with feedback potentially lost their motivation and improved only in Math. Current data show that students in the control group, whose performance is mimicking student evolution in absence of the treatments, have stagnated in Math across the academic year (their absolute performance decreased by 0.33%) but their absolute score in English increased by 50.25%. Based on such progress, it may be easier to improve in Math compared to English. A significant improvement in Math but not in English can be also found in other studies, e.g., Bettinger (2012) or Reardon et al. (2009). The evolution of the treatment effect can be found in Figure 1.2.

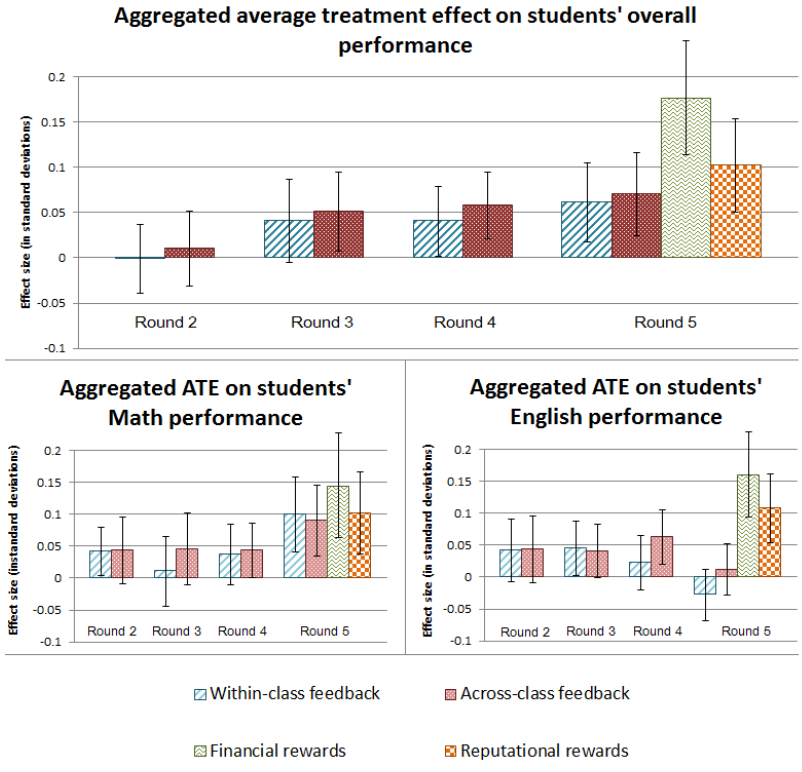
[24] Judging also by a consistently lower number of applicants for Science subjects as opposed to Arts subjects in the National examinations held by the Ugandan National Board Examination Committee.

Table 1.4: Aggregated average treatment effects of the provision of feedback and rewards performance in Math and English

Dependent variable:	MATH		ENGLISH	
	(1)	(2)	(3)	(4)
Aggregated feedback treatment	**0.094***		-0.009	
	(0.051)		(0.036)	
Aggregated reward treatment	**0.128****		**0.126****	
	(0.063)		**(0.055)**	
Within-class social comparison (T1)		**0.099***		-0.028
		(0.059)		(0.039)
Across-class social comparison (T2)		0.089		0.012
		(0.056)		(0.040)
Financial Rewards		**0.142***		**0.158****
		(0.078)		**(0.065)**
Reputational Rewards		**0.115***		**0.108****
		(0.064)		**(0.053)**
Controlled for stratas	Yes	Yes	Yes	Yes
R-squared	0.645	0.645	0.687	0.688
F-statistics	283.97***	267.81***	320.84***	295.53***
N	5102	5102	5093	5093

Note: OLS. Columns (1) – (2) show the average treatment effects (ATE) of differently aggregated treatments on students' performance in Math, columns (3) and (4) in English. Controlled for stratum fixed effects (area, level and school performance in national examinations).* significant at 10%; ** at 5%; *** at 1%.

Figure 1.2: The evolution of the aggregated average treatments on students' overall performance



1.4.2 Average Treatment Effects on Students' Well-Being (Stress and Happiness)

Both types of feedback left students' stress level unchanged, but the within-class feedback slightly decreased their happiness by 0.111 SD. The stress was induced when the feedback group students competed for rewards. Rewards, on the contrary, significantly increased the stress level of students and decreased students' happiness[25]. Decomposition of the aggregate treatment effects sheds more light on this performance-well-being-trade-off. Students without feedback who were competing for money reported their stress level as 0.466 SD higher compared to the control-group students. In other words, these students reported higher stress levels compared to 67.9% of the control-group students who received no feedback and no rewards.

Students who received repeated feedback reported significantly lower stress compared to the uninformed ones in the competition for money. While students in the within-class feedback group reported their stress level as 47.6% lower compared to the group without feedback, students in the across-class feedback group reported similar stress levels to students who were not in the competition. On the other hand, students who competed for reputational rewards without feedback reported comparable levels of stress to the control-group students but significantly higher stress if they received feedback[26].

In terms of happiness, students who competed for money reported lower happiness compared to control-group students and whether students received feedback or not did not play a significant role. Provision of reputational rewards only influenced the happiness level of

[25] In aggregated terms, financial rewards increased stress by 0.226 SD and reputational rewards by 0.177 SD; students' happiness decreased by 0.1 SD.

[26] Students who received within-class (across-class) feedback reported their stress as 0.196 SD (0.237 SD) higher compared to the control group students. In terms of percentages, students who received repeated within-class (across-class) feedback reported higher stress level in response to reputational rewards compared to 57.7% (59.4%) of students who received no feedback but competed for the rewards (the differences are only significant at 15%).

students enrolled in the across-class feedback group. Controlling for the baseline stress and happiness levels does not change the results significantly (see Appendix C1.5, C1.6 and C1.7).

A policy maximizer who would want to minimize the effects of interventions on students' well-being should therefore consider a class-level competition for financial rewards with regular feedback on each student's own performance and the performance of her class – students' performance should thus increase with no significant effect on stress or happiness.

1.4.3 Endogeneity between Performance and Stress

Provisions of feedback and/or rewards influenced performance, stress and happiness. Stress can increase because students are compared and/or they compete for rewards. Stress can influence performance (positively, if they focus better and increase their effort, or negatively, if they make more mistakes). Performance can, however, induce an increased stress level (either because students did not meet their expectations or because they internalize that their individual outcomes influenced the results of the whole group). Therefore the important question is whether students whose performance improved in response to the treatments are also those whose stress level increased. Changes in performance and stress are endogenous and I lack a proper instrumental variable to separate the treatment effects. To shed light on the problem I proceed with three exercises shown in Tables 1.5 and 1.6 (see also Appendix C1.9). The results suggest that the treatment effects of incentives are heterogeneous, i.e., students who improved their performance in response to incentives are not necessarily those whose stress increased/happiness decreased.

First, I regress students' performance in the final testing on all treatment dummies (column 1 in Table 1.5), on changes in stress (column 2) and all treatment dummies with

changes in stress (column 3). The effect sizes and their significance levels are comparable in all three cases. R-squared values and standard errors are also similar. Second, assuming that the effect of incentives on performance is driven through a change in stress (a strong but necessary assumption for this exercise), I use feedback and reward treatments as instruments for changes in stress, which are then used as a predictor of overall performance in the second stage regression, to estimate the average treatment effect on the treated (Table 1.6). First stage regressions suggest that excepting pure within-class treatment and reputational rewards interacted with both types of feedback, all other treatments predict a change in stress level. The resulting coefficient in the second stage is insignificant. The results of the Kleibergen-Paap LM test reports that the model is identified. Stock and Yogo test results in F-statistics equal to 4.121, which compared to Stock-Yogo weak ID test critical values suggest that the instruments are

Table 1.5: Sensitivity analysis and regression by the level of stress

Dependent variable:	Performance in the final testing				
Sample:	Overall sample (1)	Overall sample (2)	Overall sample (3)	Decreased or equal stress (4)	Increased stress (5)
Within-class feedback	0.061		0.065	0.036	0.068
	(0.043)		(0.044)	(0.051)	(0.046)
Across-class feedback	0.069		0.062	0.071	0.057
	(0.046)		(0.045)	(0.052)	(0.047)
Financial rewards	**0.176*** **		**0.194*** **	**0.213*** **	**0.159** **
	(0.062)		**(0.063)**	**(0.066)**	**(0.065)**
Reputation rewards	**0.102***		**0.124** **	**0.172*** **	0.069
	(0.051)		**(0.056)**	**(0.060)**	(0.055)
Change in stress		**-0.007*** **	**-0.008*** **		
		(0.003)	**(0.003)**		
Controlled for strata	Yes	Yes	Yes	Yes	Yes
R-squared	0.715	0.719	0.722	0.728	0.712
F-statistics	387.69***	367.77***	335.13***	311.37***	326.25***
Number of observation	5108	4096	4096	2002	3106

Note: OLS regressions on full sample (columns 1,2, and 3) or separately for students whose stress level increased/decreased/remained unchanged (columns 4 and 5). I control for stratification variables in the regressions – area, school level and results in national examinations (dummy variable which equals 1 if above the average). * significant at 10%; ** at 5%; *** at 1%.

Table 1.6: Performance and change in stress: Average treatment effect on the treated

Panel A: OLS and IV regressions		
Dependent variable:	Performance	
Estimated method:	OLS regression	IV regression
Change in stress	**-0.007*** **	-0.004
	(0.003)	(0.034)
Initial performance level	**0.855*** **	**0.855*** **
	(0.016)	**(0.016)**
Kleibergen-Paap statistic		[p=0.041]
Controlled for strata	Yes	Yes
Adjusted R-squared	0.7195	0.7194
F-statistics/Wald chi2(df)	367.77***	4329.32***
Number of observation	4096	4096
Durbin-Wu-Hausman test	p=0.979	
Endogeneity test%	p=0.842	
Panel B: First-stage regressions		
Dependent variable:	Change in stress level	
Estimated method:	OLS regression	
Within-class feedback,	0.025	
no feedback (T1_solo)	(0.359)	
Across-class feedback,	**-0.568****	
no feedback (T2_solo)	**(0.276)**	
Financial Rewards,	**1.515*** **	
no feedback (Fin_solo)	**(0.319)**	
Reputational Rewards,	**0.884***	
no feedback (Rep_solo)	**(0.487)**	
Within-class feedback,	**0.651***	
monetary reward (T1_fin)	**(0.386)**	
Across-class feedback,	**-0.741****	
monetary reward (T2_fin)	**(0.367)**	
Within-class feedback,	0.429	
reputat.reward (T1_rep)	(0.381)	
Across-class feedback,	-0017	
reputat.reward (T2_rep)	(0.452)	
Initial performance level	0.023	
	(0.081)	
F statistics	8.17***	
Controlled for strata	Yes	
Adjusted R-squared	0.043	
Number of observation	4105	

Note: OLS and 2SLS regressions with robust standard errors clustered at class level. Feedback and reward treatments are used as instruments for change in stress. I control for stratification variables in the regressions – area, school level and results in national examinations (dummy variable which equals 1 if above the average); % endogeneity test based on the difference of two Sargan-Hansen statistics. * significant at 10%; ** at 5%; *** at 1%

weak. In order to test for the presence of endogeneity I conducted the Durbin-Wu-Hausman test and the p-value implied that the suspected variable is not endogenous (p-value equaled 0.979). I also used an endogeneity test based on the difference of two Sargan-Hansen statistics which resulted again in the OLS estimates being consistent; i.e., students whose performance level changed in response to the proposed treatments are not the ones whose stress levels changed. Lastly, the decomposition of the average treatment effects by change in stress level also shows heterogeneity in results (columns 4 and 5 in Table 1.5). The effects of incentives on performance are similar for students whose stress level decreased or increased, which is in line with the heterogeneity of the treatment effects. The proposed treatments had impact on performance and well-being. The current experimental design, however, cannot distinguish the channels.

1.4.4 Group Composition

If students are allowed to choose whether they want to compete in groups or as individuals, the studies have shown that average-ability subjects have a higher tendency to choose group-work compared to high- or poor-ability subjects (Amann and Gall, 2006, Breton et al., 1998, 2003). In this study, I am interested in the behavior of students exogenously grouped with others from the whole ability-spectrum. Students were assigned to groups of three and received feedback about their own as well as group performance during the whole academic year. In some cases the number of students in the class was not divisible by three, in which case there were one or two groups of four students (in total 18 of 717 groups). In the following analysis I take only three-person groups into account.

Three types of ability groups (poor, mixed and good performers) and four types of gender groups (all boys, two boys and one girl, one boy and two girls, or all girls) were formed. The

analysis helps us to understand how well-informed groups who differ in terms of ability or gender composition perform in response to financial and reputational rewards.

The applicability of the results goes beyond the educational framework. Groups are increasingly used in decision-making processes in organizations (Hamilton et al., 2003; Woolley et al., 2010). Companies spend large amounts of money on incentivizing employees. With the aim of maximizing efficiency or group performance they often carefully select high-ability performers to work on a project or represent the firm, etc. In such an environment the results of this research offer comparison of responses of different ability groups with or without further incentives.

1.4.4.1 Ability Composition

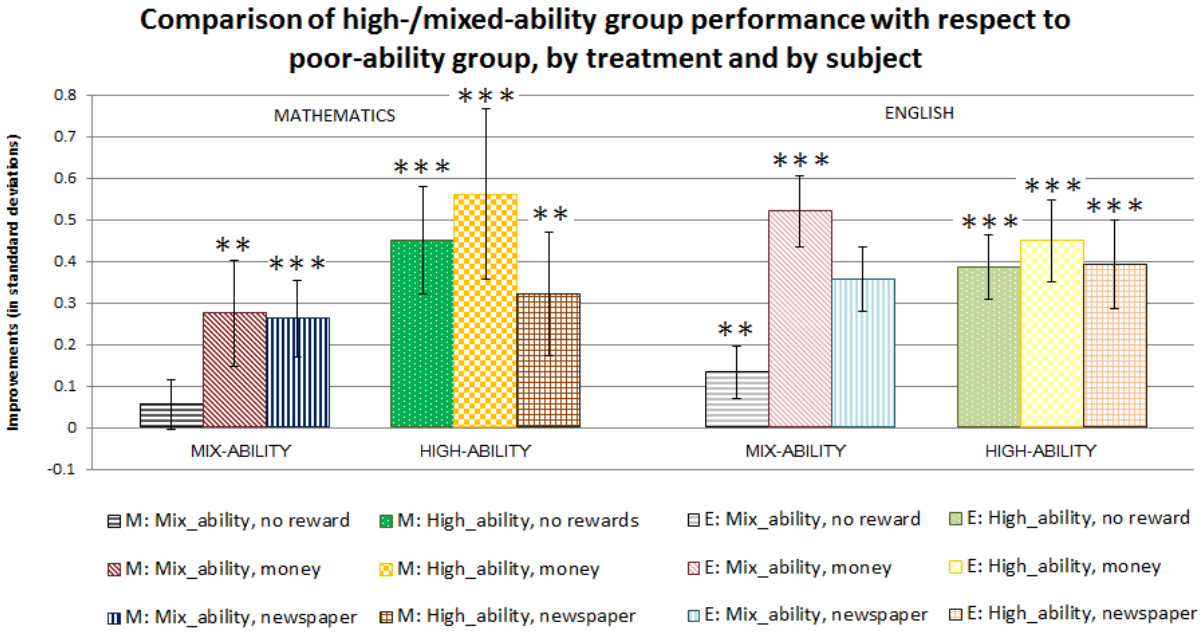
The aim is to observe whether and to what extent students randomized into within-class feedback groups differ in responses to rewards based on their initial group achievements. Group achievement is measured in terms of group baseline performance in Math and in English. I differentiate three types of groups: (1) high-achievers if all group-members scored above the median; (2) poor-achievers if all group-members scored below the median; or (3) mixed-achievers if the performance of group-members varied across the whole performance distribution. Ability is not directly observable, but I use student performance as a proxy for ability and in later discussion I denote the groups as high-ability, poor-ability, and mixed-ability.

The main result is that groups of mixed-ability students can perform similarly to high-ability groups if they are provided with incentives. The type of rewards matter for improvements in English: students competing for money improved by 0.164 SD more compared to those competing for reputational rewards, significant at α=5%.

Students in mixed-ability groups who were not enrolled in a competition for rewards performed similarly to students in poor-ability groups in Math, but they outperformed them in English by 0.133 SD. Once mixed-ability groups are incentivized, they significantly outperformed poor-ability students both in Math (0.21-0.22 SD) and English (0.38 SD and 0.224 SD in response to the financial and reputational rewards, respectively. The treatment effects were driven by those students in the mixed-ability groups who scored below median in the baseline testing.

Students in the high-ability groups who did not compete for rewards significantly outperformed students in the poor-ability groups by 0.451 SD in Math and 0.387 SD in English and therefore outperformed poor-ability students by 65 - 67.5%. There is no statistically significant value added to provision of rewards for high-ability group students as seen also in Figure 1.3 below. See also Appendix E1.1.

Figure 1.3: Average treatment effects among within-class feedback groups, by baseline group ability



Note: OLS. All groups received within-class feedback during the whole academic year. The bars show the average treatment effects of different incentive schemes of mixed- and high-ability groups in comparison to poor-ability groups. Vertical error bars show robust standard errors. Controlled for stratum fixed effects (area, level and school performance in national examinations). Stars indicate significance of the difference in means. * significant at 10%; ** at 5%; *** at 1%

Except in two cases, students do not differ across different ability groups in perceived stress and happiness. The average student from the mixed-ability group incurred higher stress compared to 62.2% of the poor-ability group students when offered financial rewards, and the average high-ability group student incurred higher stress compared to 62.6% of the poor-ability students when offered reputational rewards. Ability-composition of the groups does not seem to determine the level of effort exerted to answering the exam questions.

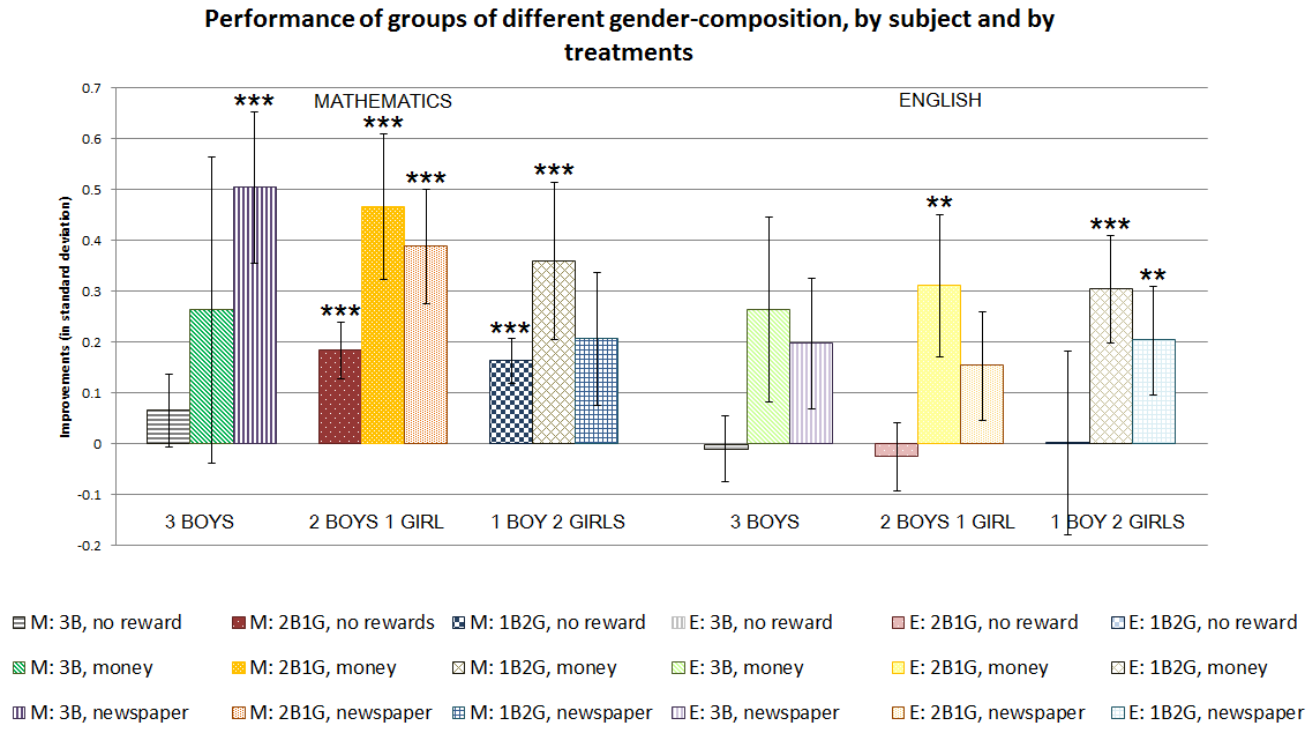
1.4.4.2 Gender Composition

Group performance may also be influenced by its gender composition. Apesteguia et al. (2012) studied the effects of gender compositions on the economic performance of undergraduate and MBA students involved in a business game competition. All-male and mixed-gender groups outperformed all-female groups. (In other words groups consisting of all women served as the comparison group.) The group composition of two men and one woman seemed to be optimal since they performed the best. In the current study the groups also consisted of three students and due to random assignment, four different gender-compositions were formed (all-girls, majority girls, majority boys and all-boys). The findings are similar. Mixed groups outperform single-gender groups, although I do not find strong support in the current study for the groups of two boys and one girl to be dominant.

Figure 1.4 compares the performance of mixed groups and groups of boys compared to groups of girls (see also Appendix E1.2). The results suggest that in the absence of rewards, mixed gender groups outperformed single-gender groups by 0.16 - 0.18 SD. All-boys groups performed comparably to all-girls groups. When rewards were offered, all-female groups were outperformed by all other types of groups from 0.19 to 0.50 SD depending on the group gender

composition and type of the rewards. Group composition does not seem to play a significant role in terms of students' perceived stress in reaction to different treatments. All-girls groups seem to be on average happier than other groups[27].

Figure 1.4: Average treatment effects among within-class feedback groups, by baseline gender composition of groups



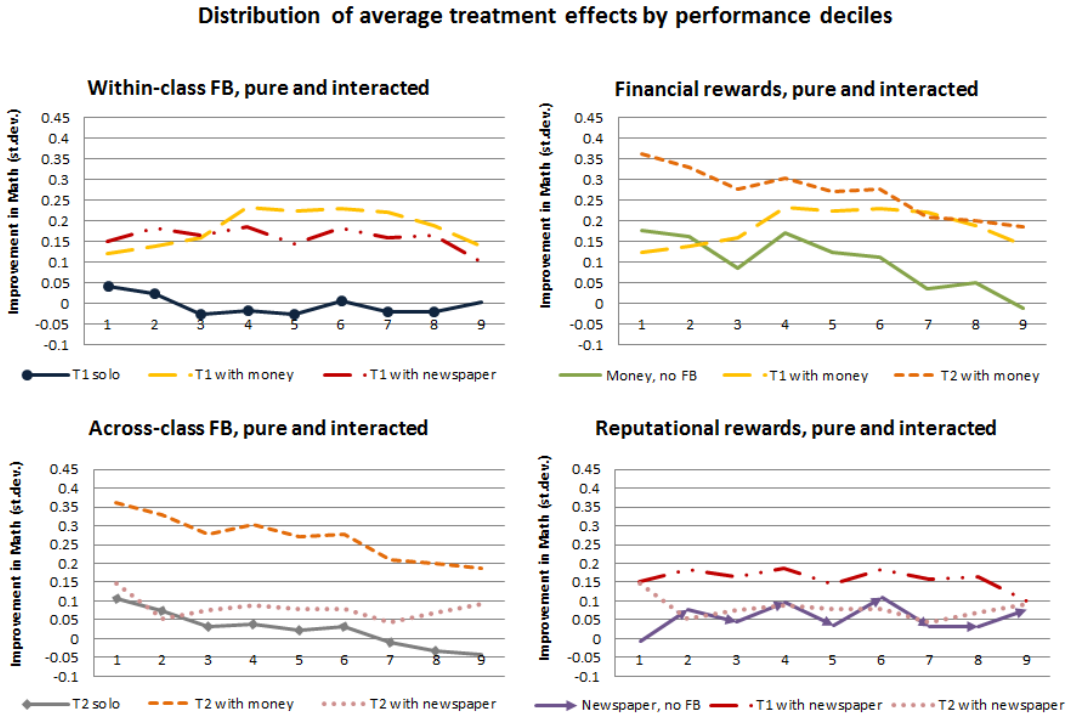
Note: OLS. All groups received within-class feedback during the whole academic year. 3 Boys denotes groups consisting of boys only, 2B1G groups of two boys and one girl, 1B2G one boy and two girls. Groups consisting of all-girls served as comparison groups. The bars show the average treatment effects of different incentive schemes of 3-boy-groups and mixed-groups in comparison to 3-girl-groups. Vertical error bars show robust standard errors. Controlled for stratum fixed effects (area, level and school performance in national examinations). Stars indicate significance of the difference in means. * significant at 10%; ** at 5%; *** at 1%

[27] Testing the differences in happiness levels in the baseline survey shows that groups of 3 females did not significantly differ in terms of their happiness compared to groups of 3 boys and groups of 1-boy-2-girls. The difference between 3-girls and 2-boy-1-girl groups is significant at 10% level (p=0.088). All other groups are comparably happy.

1.4.5 Distributional Analysis

It is also important to learn whether the treatment effect differs at different points of the performance distribution. To learn to what extent high performing students (those in the upper tail of the performance distribution) differ in their reactions to the treatment effects from low performing students, I estimate quantile regressions. Figure 1.5 shows the average treatment effect of the incentives on student performance of students by their rank in the performance distribution. The figure consists of four graphs that differ in the combinations of the treatments. Graphs in the first column compare the average treatment effects of the sole feedback with feedback-reward interacted treatments. Graphs in the second column compare average treatment effects of sole-reward with feedback-reward-interacted treatments. The main

Figure 1.5: Distribution of the average treatment effects of different incentives on overall performance (Math and English pooled), by deciles



Note: T1_solo stands for within-class feedback without further rewards, T1 with money/newspaper stands for within-class feedback and further monetary/non-monetary rewards.

observation is that the bottom performers responded more strongly than the top performers to sole financial rewards, sole across-class feedback or their combination. In all other cases the bottom performers responded comparably to the top performers.

1.4.6 Positiveness and Negativeness of Feedback and Well-Being

The nature of the feedback students receive may influence their well-being. Azmat and Iriberri (2016) show that positive feedback increases students' happiness. In the current study, students received feedback regarding their own performance together with information about the absolute and relative performance of their group. They were also informed whether they improved or worsened in the two subsequent testing rounds. Students are considered as receiving mostly positive (negative) feedback if in at least two out of three cases they improved (worsened).

I find similar results to Azmat and Iriberri (2016) but mainly for students outside of the competition for rewards.[28] Students enrolled in a competition reported higher happiness only in the case of competition for money combined with across-class feedback (see also Table 1.7). The nature of the feedback does not influence students' stress level.

[28] The within-class (across-class) feedback group who received mostly positive feedback reported higher happiness than 55.2% (58.9%) of the students in the within-class (across-class) feedback group who received mostly negative feedback.

Table 1.7: Average treatment effects on well-being, by feedback positiveness

Dependent variable:	Stress		Happiness	
	(1)	(2)	(3)	(4)
Within-class feedback aggregated	-0.007		**0.103***	
	(0.056)		**(0.056)**	
Within-class feedback, no rewards		-0.117		**0.132****
		(0.080)		**(0.063)**
Within-class feedback with monetary rewards		0.060		0.014
		(0.115)		(0.092)
Within-class feedback with reputational rewards		0.149		0.120
		(0.123)		(0.102)
Number of observations	1453	1453	1454	1454
Controlled for strata	Yes	Yes	Yes	Yes
R-squared	0.047	0.052	0.070	0.072
Across-class feedback aggregated	-0.012		**0.218****	
	(0.154)		**(0.088)**	
Across-class feedback, no rewards		-0.082		**0.259*****
		(0.160)		**(0.094)**
Across-class feedback with monetary rewards		0.037		**0.221***
		(0.209)		**(0.110)**
Across-class feedback with reputational rewards		0.282		0.013
		(0.199)		(0.146)
Number of observations	1451	1451	1416	1416
Controlled for strata	Yes	Yes	Yes	Yes
R-squared	0.059	0.070	0.103	0.109

Note: OLS. Students who received mostly positive feedback are compared to students who received mostly negative feedback. Columns (1) and (3) show aggregated treatment effects, columns (2) and (4) disaggregated. Controlled for stratum fixed effects (area, level and school performance in national examinations).* significant at 10%; ** at 5%; *** at 1%

1.4.7 Intrinsic and Extrinsic Motivation

Psychologists found that monetary and non-monetary incentives, despite their positive effects in the short run, may in the long run worsen students' performance due to decreases in subjects' intrinsic motivation,[29] and lowered interest in the task (Deci, 1971; Deci et al., 1999; Benabour and Tirole, 2003; and Frey and Jegen, 2000). In this study, I do not have any direct

[29] Definition of intrinsic and extrinsic motivation can be found in Ryan, and Deci, 2000.

measure of intrinsic and extrinsic motivation. However, immediately after the last exam from Math and English, I asked students how much time they spent on preparing for the exam. This self-reported measure of effort may serve as a proxy for motivation. The results show that students involved in a competition for rewards reported that they exerted significantly less effort on exam preparation than the control group students. The results are similar by gender. For further details see Appendix F1.6.

1.4.8 Treatment Effects on Attrition

High drop-out and absence rates are common issues with students in developing countries and there is no exception in my data. There are several reasons[30]. Some students did not have the money to pay the school fees and decided to change schools to avoid repaying their debt, others changed their school for family reasons (the family moved to a different area, they were sent to live with other family members, etc.). Some completely dropped out of school. Some just registered as new students. Some passed away. Due to the constraints of the experiment, all participation data are based on our scheduled visits only (no random visits were organized).

Estimates of treatment effects can be biased if the attrition from control versus treatment groups differs systematically and the difference is caused by the presence of the treatment. Students in treatment groups attrite less often in absolute values and are more often present in all five testing rounds compared to their control-group counterparts. In order to see whether and to what extent social comparison and reward treatments influence the probability of

[30] All answers are based on the questionnaires filled out by headmasters, directors and teachers.

dropping out. I run a probit model on attrition and full attendance on all treatment dummies controlling for strata variables (Table 1.8).

The attrition rate includes students who missed our last testing round but attended the baseline testing at the beginning of the project. Non-rewarded students exposed to both within and across-class social comparison feedback had from 6.5 to 6.9% lower probability of missing the final testing round. Among rewarded students who did not receive any feedback only

Table 1.8: Probabilities of students' absence rates and being present in all testing rounds, by gender

Dependent variable:	Absence rates			Students present in all testing rounds		
Sample:	Overall	Girls	Boys	Overall	Girls	Boys
NON-INTERACTED TREATMENT EFFECTS						
Within-class feedback,	**-0.066***	**-0.064***	-0.058	**0.105****	**0.091***	**0.108****
no rewards (T1_solo)	**(0.039)**	**(0.037)**	(0.049)	**(0.049)**	**(0.049)**	**(0.055)**
Across-class feedback,	**-0.071****	-0.046	**-0.097****	**0.124*****	**0.110****	**0.137*****
no rewards (T2_solo)	**(0.032)**	(0.035)	**(0.038)**	**(0.042)**	**(0.049)**	**(0.046)**
Financial Rewards,	**-0.130*****	**-0.128*****	**-0.124****	**0.127****	**0.168****	**0.093***
no feedback (Fin_solo)	**(0.038)**	**(0.033)**	**(0.055)**	**(0.056)**	**(0.067)**	**(0.057)**
Reputational Rewards,	-0.056	-0.021	**-0.100****	0.030	0.047	0.021
no feedback (Rep_solo)	(0.046)	(0.052)	**(0.050)**	(0.077)	(0.087)	(0.074)
TREATMENT INTERACTIONS						
Within-class feedback with	**-0.158*****	**-0.127*****	**-0.196*****	**0.233*****	**0.228*****	**0.247*****
financial rewards (T1_fin)	**(0.033)**	**(0.033)**	**(0.037)**	**(0.062)**	**(0.071)**	**(0.064)**
Across-class feedback with	**-0.147*****	**-0.128*****	**-0.157*****	**0.263*****	**0.257*****	**0.252*****
financial rewards (T2_fin)	**(0.032)**	**(0.031)**	**(0.041)**	**(0.060)**	**(0.068)**	**(0.063)**
Within-class feedback with	**-0.157*****	**-0.146*****	**-0.171*****	**0.208*****	**0.209*****	**0.217*****
reputation rewards (T1_rep)	**(0.038)**	**(0.036)**	**(0.043)**	**(0.067)**	**(0.073)**	**(0.064)**
Across-class feedback with	**-0.212*****	**-0.192*****	**-0.226*****	**0.099***	0.079	**0.126****
reputation rewards (T2_rep)	**(0.026)**	**(0.026)**	**(0.031)**	**(0.051)**	(0.057)	**(0.056)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes
Wald chi2 (df)	141.16***	115.21***	133.06***	157.22***	116.26***	145.21***
N	7050	3818	3139	7050	3818	3139

Note: Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level - P6,P7, S1 up to S4). N stands for the number of observations. * significant at 10%; ** significant at 5%; *** significant at 1%

students rewarded financially lowered their attrition by 7.9%. Reputation rewards without provided feedback do not affect attrition rate. All treatment interactions lowered the attrition rate (from 9.3 to 17.2%). There are no gender differences in terms of decrease in absence rates or increase in attendance rates. Similarly, no differences can be attributed to group gender composition. On the other hand, groups consisting of higher (mixed) achieving students as determined by baseline performance have an 11.9% (7.79%) lower probability to attrite compared to the poor-achievers group.

In sections 1.6.2. and 1.6.3 I discuss whether lower absence rates among treated students may potentially reintroduce selection bias. I also estimate the treatment effects using alternative specifications such as inverse probability weighting, median regression and two imputed methods to correct for potential selection bias. The results are similar to OLS results. Provision of feedback and rewards, therefore, potentially lower absence rates, which is an interesting result on its own and could be potentially tested in future work.

1.5 Gender Differences and the Channels of the Average Treatment Effects

Girls have performed differently than boys in various studies. Angrist and Lavy (2009) studied the effects of cash incentives on matriculation rates among Israeli students. Girls, contrary to boys, substantially increased their performance. A greater effect among girls was also found in the analysis of voucher provision within the PACES program in Colombia (Angrist et al., 2002). Stronger responsiveness to incentives among girls can be also found in studies of tuition provision by Dynarski (2008), early childhood interventions by Anderson (2008),

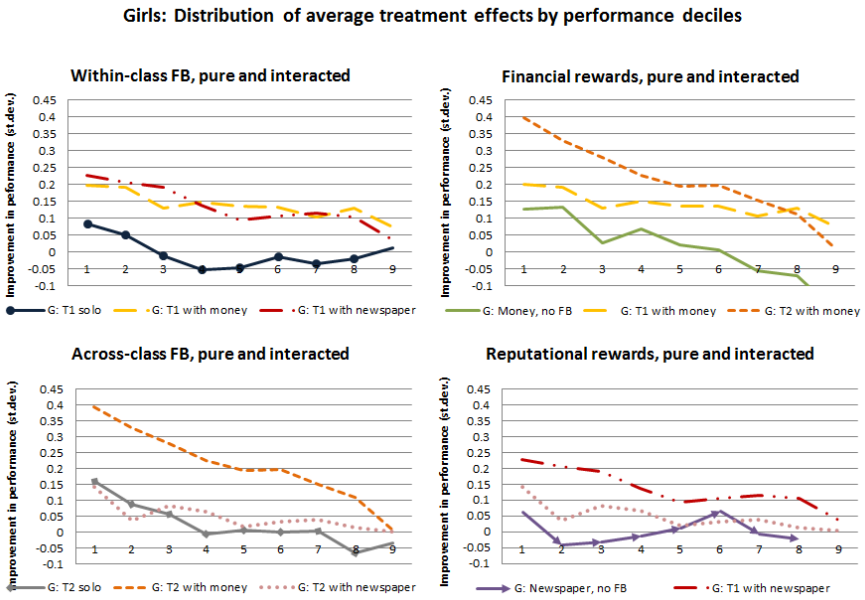
housing vouchers by Kling et al. (2007) and public sector programs by Lalonde (1995) and others[31].

The results of the current experiment show that girls reacted positively to feedback (0.12 – 0.14 SD) even if they were not offered rewards. Once involved in a competitive environment, girls improved by 0.2 to 0.28 SD (see Appendix D1). Therefore, girls can improve similarly to boys if they receive feedback about their performance, the performance of their group and their group's relative standing. In the absence of feedback, girls did not improve at all. Boys improved if they were offered rewards (with or without feedback) by 0.18 to 0.28 SD but did not react to feedback alone.

Figures 1.6 and 1.7 compare the treatment effects across the whole performance distribution by gender. While girls from the bottom of the performance distribution seem to be the most responsive to incentives, the most responsive boys are from the middle of the performance distribution. The results from quantile regressions can be found in Appendices F1.4 and F1.5. There are no gender differences from different incentive schemes on students' psychological well-being. Similar results can be found in Azmat and Iriberri (2016). The results are also in line with the literature on stereotype thread that boys are typically considered to be stronger in Mathematics (Muzzatti and Agnoli, 2007). Stereotype thread could explain why girls underperform boys in competition without feedback but perform similarly if feedback is provided. "[P]ublicly revealing the social identity of an individual can change his behavior even when that information is irrelevant to payoffs" (Hoff and Pandey, 2006).

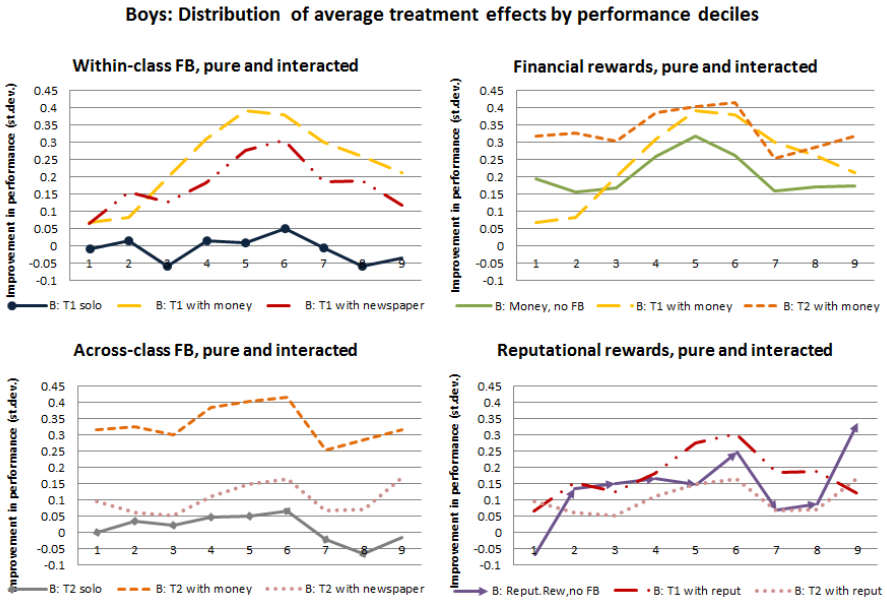
[31] For a review of gender differences in risk preferences, other-regarding preferences and competitive preferences, see Croson and Gneezy (2009).

Figure 1.6: Comparison of the sizes of the average treatment effects on girls' performance in Math and English, by deciles



Note: T1_solo stands for within-class feedback without further rewards, T1 with money/newspaper stands for within-class feedback and further monetary/non-monetary rewards.

Figure 1.7: Comparison of the sizes of the average treatment effects on boys' performance in Math and English, by deciles



Note: T1_solo stands for within-class feedback without further rewards, T1 with money/newspaper stands for within-class feedback and further monetary/non-monetary rewards.

1.6 Robustness Checks

1.6.1 Multiple Comparisons

The probability that the coefficients are significant purely by chance increases with the number of hypotheses tested. Multiple-test procedures take p-values from multiple comparisons testing and uncorrected critical p-values interpreted either as familywise error rate (FWER) or as false discovery rate (FDR) and result in adjusted critical p-values.[32]

In order to address these concerns about multiple inference I control for FWER using one-step methods (Bonferroni, in Dunn 1961; and Sidak, 1967 corrections), and step-down methods (Holm, 1979; and Holland-Copenhaver, 1987 corrections) and for FDR, using step-up methods (Simes, 1986; Hochberge, 1988; and Yakutieli-Benjamini, 2001 procedure). A detailed description of the procedures can be found in Newson (2010). The corrected p-values are summarized in Table 1.9.

Table 1.9: Adjusted p-values for aggregated treatment effects and disaggregated treatment effects

Type of correction (corresponding to uncorrected alpha = 0.1)			Correlation assumed	Aggregated treatment effect	Disaggregated treatment effects
FWER	One-step method	Bonferroni correction	Arbitrary	0.0077	0.0053
		Sidak correction	Nonnegative	0.0081	0.0055
	Step-down method	Holm correction	Arbitrary	0.0200	0.0083
		Holland correction	Nonnegative	0.0210	0.0087
FDR	Step-up method	Hochberg correction	Independence	0.0170	0.0077
		Simes correction	Nonnegative	0.0620	0.0680
		Yekutieli correction	Arbitrary	0.0190	0.0150

[32] "If the input uncorrected critical p-value $\alpha \in (0,1)$ is an FWER, then we can be $100(1-\alpha)\%$ confident that all the null hypotheses in the discovery set are false. If the input uncorrected critical p-value $\alpha = \beta * \gamma$ is an FDR, then we can be $100(1-\beta)\%$ confident that over $100(1-\gamma)\%$ of the null hypotheses in the discovery set are false" (Newson, 2010, p.569).

The disadvantage of FWER is that it can result in low power for testing single hypotheses in large experiments with high numbers of multiple comparisons. In such cases FDR is preferred since it controls for the proportion of Type I errors to true positives and therefore results in greater power. In the case of this experiment, one-step and step-down methods rule out any of the initially presented average treatment effects of interventions on students' performance. FWER procedures seem to be too conservative due to the number of multiple comparisons I test.

Among FDR procedures I rule out the Hochberg corrections because I do not meet its restriction of independence (Hochberge, 1988; Newson, 2010) since I compare all groups with treatments to the same control group. The significance of the average treatment effects of different incentive schemes on students' performance was confirmed when I used the Simes correction and, with some exception, when I used the Yekutieli correction. For instance, the effect of sole financial rewards and financial rewards combined with within-class feedback is insignificant. Similarly, regarding the average treatment effects on subjective well-being, with one exception, the negative impact of sole financial rewards is significant using all types of FWER and FDR corrections. A summary of corrected p-values for all disaggregated treatment effects can be found in Appendices F1.7a, b, and c.

1.6.2 Who are the Attrited Students? Random versus Non-Random Attrition

The treatments influenced the probability of students' presence or absence for our visits. In absolute numbers, fewer students dropped out from treated classes and more students from the treatment groups attended all five testing rounds compared to the control group students. Further, those who attrited from the within-class feedback group are worse in terms of their initial performance compared to students from the across-class feedback group or the control

group. That might re-introduce a bias if the treated students who are present during the final testing round are systematically different compared to the control-group students. As shown in Table 1.10, this is not the case in this project. The distribution of students who remained throughout the year in either treatment group (based on their initial performance) is not statistically different from the distribution of the initial abilities of students from the control group. Therefore, the OLS estimate should provide unbiased estimates of the treatment effects. Nevertheless, I use inverse probability weights and imputation methods to check the stability of the results (see section 1.6.3).

Table 1.10: Ksmirnov test on equality of distributions; by students' absence/presence rates

	Baseline differences		Students who attrited		Students who stayed		Always-present students	
	(T1 – C)	**(T2 – C)**	**(T1 – C)**	**(T2 – C)**	**(T1 – C)**	**(T2 – C)**	**(T1 – C)**	**(T2 – C)**
Math	0.123	0.274	0.000	0.158	0.752	0.192	0.677	0.958
English	0.952	0.168	0.003	0.546	0.230	0.282	0.211	0.840

Note: P-values reported. T1 stands for within-class social comparison group, T2 for across-class comparison group and C represents control group with no feedback provided. "Students who attrited" stands for students who were present in the baseline but not in endline testing. "Students who stayed" were present in both baseline and endline but may have been absent in Round 2,3, and/or 4. "Always-present students" participated in all testing rounds.

1.6.3 Stability of the Results

In order to adjust the results for non-random attrition, I proceed with imputation methods and inverse probability-weighted regressions (Imbens, 2004; Woolridge, 2007; Kwak (2010); Hirano et al., 2000, etc.). Inverse probability weighting (IPW) can adjust for confounding factors and selection bias. As the title suggests, IPW assigns a weight to every student that is equal to the student's inverse probability to being absent and adjusts for that in the estimation

of the treatment effects. An imputation method is used to fill in the missing observations of students who were absent or dropped out in the last testing round based on a predefined rule.

Tables 1.11 and 1.12[33] provide a comparison of ordinary least squares estimations (column 1) of the treatment effects to the weighted least squares using inverse probability

Table 1.11: Average treatment effects of different motivation schemes - alternative specifications

Dependent variable:	Math			
Estimation method:	OLS	IPW	Imputation (median ratio)	Imputation (class percentiles)
NON-INTERACTED TREATMENT EFFECTS				
Within-class feedback, no rewards (T1_solo)	0.100 (0.085)	0.046 (0.092)	**0.133*** **(0.079)**	0.123 (0.085)
Across-class feedback, no rewards (T2_solo)	0.082 (0.073)	0.067 (0.079)	**0.129*** **(0.068)**	0.087 (0.078)
Financial Rewards, no feedback (Fin_solo)	0.106 (0.101)	0.151 (0.102)	**0.169*** **(0.096)**	0.143 (0.106)
Reputational Rewards, no feedback (Rep_solo)	0.138 (0.141)	0.188 (0.149)	**0.206*** **(0.124)**	0.177 (0.128)
TREATMENT INTERACTIONS				
Within-class feedback, with monetary reward (T1_fin)	**0.231*** **(0.118)**	**0.338**** **(0.135)**	**0.281**** **(0.129)**	**0.273**** **(0.124)**
Across-class feedback, with monetary reward (T2_fin)	**0.277**** **(0.139)**	**0.456***** **(0.132)**	**0.331**** **(0.128)**	**0.305**** **(0.139)**
Within-class feedback, with reputation reward (T1_rep)	**0.209**** **(0.103)**	**0.212*** **(0.108)**	**0.266**** **(0.073)**	**0.258**** **(0.112)**
Across-class feedback, with reputation reward (T2_rep)	**0.188**** **(0.080)**	**0.208**** **(0.087)**	**0.186**** **(0.073)**	**0.250***** **(0.090)**
Controlled for stratas	Yes	Yes	Yes	Yes
F-statistics	247.06***	281.06***	107.19***	293.34***
N of observation	5102	5102	6736	7107

Note: Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). N stands for the number of observations. * significant at 10%; ** significant at 5%; *** significant at 1%

[33] See also Appendices F1.1, F1.2 and F1.3.

weights (column 2), separately for Math and English. Correcting for the probability of dropping out, treatment effects are similar or slightly higher in absolute terms but not significantly different. The results of the imputation methods (columns 3 and 4) bring similar conclusions.

I use two different measures to impute missing observations – median ratio and the class percentile ranks (inspired by Krueger, 1999). All of the measures take the advantage of repeated school visits and follow the same logic – if the observation from the last school visit is missing, I

Table 1.12: Average treatment effects of different motivation schemes - alternative specifications

Dependent variable:	English			
Estimation method:	OLS	IPW	Imputation (median ratio)	Imputation (class percentiles)
NON-INTERACTED TREATMENT EFFECTS				
Within-class feedback, no rewards (T1_solo)	**-0.128**** **(0.056)**	**-0.133*** **(0.070)**	**-0.133**** **(0.060)**	**-0.135***** **(0.045)**
Across-class feedback, no rewards (T2_solo)	-0.049 (0.059)	-0.079 (0.072)	-0.052 (0.063)	-0.046 (0.048)
Financial Rewards, no feedback (Fin_solo)	0.045 (0.088)	0.032 (0.085)	-0.006 (0.096)	0.041 (0.069)
Reputational Rewards, no feedback (Rep_solo)	0.016 (0.082)	0.004 (0.084)	-0.089 (0.123)	0.036 (0.059)
TREATMENT INTERACTIONS				
Within-class feedback, with monetary reward (T1_fin)	0.103 (0.094)	**0.145*** **(0.086)**	0.096 (0.108)	0.072 (0.080)
Across-class feedback, with monetary reward (T2_fin)	0.173 (0.094)	**0.258**** **(0.102)**	0.113 (0.099)	**0.137*** **(0.075)**
Within-class feedback, with reputation reward (T1_rep)	0.087 (0.080)	0.041 (0.078)	0.069 (0.082)	0.069 (0.058)
Across-class feedback, with reputation reward (T2_rep)	0.047 (0.080)	0.071 (0.077)	0.024 (0.082)	0.059 (0.064)
Controlled for stratas	Yes	Yes	Yes	Yes
F-statistics	265.22***	191.43***	125.65***	293.34***
N of observations	5093	5093	6736	7107

Note: Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). N stands for the number of observations. * significant at 10%; ** significant at 5%; *** significant at 1%

look at the last score available and adjust for the differences in test difficulty. The same procedure is done to impute Math and English scores separately. The median ratio measure imputes the last available observation and the class percentile ranks take into consideration the rank of the student in the last available distribution and impute the score corresponding to the student of the same rank in the final visit distribution. The imputation method artificially fills missing observations and the results serve only as bounds. Both imputation measures deliver similar or stronger results compared to ordinary least squares. Ordinary least squares results are also comparable to the weighted regression estimates.

1.7 Conclusion

Various interventions have been conducted in educational literature with the aim of lowering absenteeism and increasing student performance. Researchers usually focus on the main outcomes of their interventions, such as subjects' performance, absence or drop-out rates, leaving outcomes other than learning aside. Evidence from psychology indicates that well-being, measured in terms of stress and happiness, serves as an important precursor of future performance. This paper contributes to the current literature by studying the effects of various types of incentives on student performance and their well-being (measured by reported happiness and stress levels). I bring new evidence of the performance-versus-well-being tradeoff by implementing two types of social comparative feedback regimes (within- and across-class group feedback), two types of incentive regimes - financial and reputation rewards, and their combinations.

The results of this study show that students who received repeated feedback without further incentivization mildly improved their performance compared to the control group

students, but the difference was insignificant. The results differ in Math and in English. While in Math, students improved significantly by 0.08 to 0.13 SD, there was no improvement in English. The results are driven by improvements in girls' performance. Provision of rewards (students did not receive feedback), on the contrary led to an improvement of 0.1 to 0.18 SD in both Math and English. These results are driven by boys. The design of the experiment allows me to study only immediate effects of the provision of rewards. Future studies could therefore focus on the effects of repeated competition for rewards. Students improved significantly more in response to rewards compared to the immediate feedback (measured by their performance right after first feedback provision). The difference is not significant when I compare the average treatment effects of rewards to repeated feedback (measured by the performance in the final testing round).

Students exposed to the combined incentive scheme of feedback with rewards increased their performance by 0.20 and 0.29 SD if rewarded financially and 0.12 to 0.18 SD if rewarded reputationally. There is, however, a trade-off between improvements in performance and changes in students' well-being in response to different incentive schemes. While students exposed to feedback and reputational rewards improved their performance mildly compared to the control group students, neither their happiness nor stress changed. Financial rewards led to stronger improvements in performance but were associated with higher stress and lower happiness. Being informed seems to play a role in terms of stress. Students involved in a competition for monetary rewards reported significantly lower stress levels compared to those who competed for money without feedback. Stressed students exerted less effort, performed worse on average and attrited by 29% more compared to more relaxed students.

Furthermore, this paper sheds light on gender differences in responsiveness to different incentive types. According to the results, girls did not improve when they received no feedback but they competed for rewards of any type and significantly underperformed boys. If the girls were repeatedly given feedback (and the type of feedback does not matter), they performed comparably to boys. Moreover, girls also responded positively to sole feedback (without rewards). Comparative feedback played a crucial role for girls in enhancing their performance in a tournament environment. Boys reacted only with respect to rewards. Feedback did not play any role in their performance improvements. There were no gender-differences in the effects of incentives on well-being.

The results of the current experiment may be of importance especially for policy makers developing strategies to improve the performance and well-being of primary and secondary school students. The results call attention to the impact of incentives on students' stress and happiness. The current study cannot distinguish whether "short-term pain is for longer-term gain", in other words it is not clear whether an increase in stress could serve as a motivator for better performance in the longer run. Stress has been found to be related to higher absence rates, and high absenteeism is a prevalent problem of developing countries, and could potentially influence students' career choices, too. Therefore policy makers should exercise a great deal of caution in designing educational rewards and consider the impact on student well-being.. The goal of the current study is solely to point out another dark side of incentives and future research is needed to shed light on its long-term consequences.

Appendix 1

APPENDIX A1: Summary statistics and randomization balance

Appendix A1.1: Balance between control and treatment groups

Variable	Control	Within-class feedback	Across-class feedback
School Level:			
The number of primary schools	10	11	10
The number of secondary schools	7	7	8
School Type:			
Public Schools	8	5	6
Private Schools	7	9	8
Community Schools	2	4	4
By Population	2345 (48 groups)	2415 (51 groups)	2371 (51 groups)
By PLE/UCE results	3.175	3.039	3.102
By testing results	21.140	21.363	21.648

Note: min(PLE/UCE)= 1.7397, max(PLE/UCE)= 4.2857, mean(PLE/UCE)=3.1040
Note: min(TR)=8.3125, max(TR)=39.7765, mean(TR)=21.3192, where TR=Testing Results

Appendix A1.2: Randomization balance. aggregated treatments

	Means			Mean Differences		Joint P-value
	Within-class feedback (T1)	Across-class feedback (T2)	Control (C)	(T1 – C)	(T2 – C)	
A. QUESTIONNAIRES						
A.1 After Math questionnaire						
Q1: Expected number of points [min 1, max 10]	4.331	4.537	4.551	-0.221 (0.150)	-0.151 (0.145)	0.299
Q2: Subjective effort level [min 1, max 5]	3.447	3.525	3.504	-0.057 (0.053)	0.021 (0.052)	0.298
Q3: Perceived difficulty [min 1, max 5]	3.341	3.494	3.423	-0.082 (0.053)	0.072 (0.052)	**0.030**
Q4: Subjective level of happiness [min 1, max 7]	3.319	3.253	3.184	0.135 (0.092)	0.069 (0.094)	0.343
A.2 After English questionnaire						
Q1: Expected number of points [min 1, max 10]	5.715	5.757	5.796	-0.081 (0.161)	-0.039 (0.144)	0.879
Q2: Subjective effort level [min 1, max 5]	3.547	3.627	3.553	-0.006 (0.046)	**0.074*** **(0.044)**	0.141
Q3: Perceived difficulty [min 1, max 5]	3.644	3.644	3.677	-0.033 (0.052)	-0.033 (0.049)	0.752
Q4: Subjective level of happiness [min 1, max 7]	2.950	2.904	2.856	0.094 (0.084)	0.048 (0.086)	0.534
A.3 Aspiration questionnaire						
Education over Relax [min 1, max 5]	3.833	3.756	3.778	0.056 (0.049)	-0.021 (0.049)	0.269
Education over Work [min 1, max 5]	3.538	3.496	3.477	0.060 (0.057)	0.019 (0.059)	0.526
Work over Relax [min 1, max 5]	2.766	2.701	2.803	-0.037 (0.094)	-0.102 (0.090)	0.524
Perceived happiness scale [min 4, max 28]	11.479	11.653	11.223	0.256 (0.231)	**0.429**** **(0.222)**	0.155
Perceived stress [min 0, max 16]	6.018	6.352	5.756	0.262 (0.164)	**0.595***** **(0.142)**	**0.000**

Note: comparison of mean characteristics of students in treatment and control groups. T1/T2 stands for within-/across- class social comparison groups and C for control gropu. Robust clustered standard errors at class level are in parentheses, adjusted for stratification. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A1.2: Randomization balance, aggregated treatments (Continued)

	Means			Mean Differences		Joint P-value
	Within-class feedback (T1)	Across-class feedback (T2)	Control (C)	(T1 – C)	(T2 – C)	
B. OTHER						
B.1 Attrition rates						
All schools	0.359	0.346	0.454	**-0.095*** (0.034)**	**-0.108*** (0.033)**	**0.002**
Restricted sample#	0.358	0.348	0.417	**-0.059* (0.030)**	**-0.069** (0.029)**	**0.041**
B.2 Always-comers						
All schools	0.202	0.186	0.082	**0.121*** (0.033)**	**0.104*** (0.104)**	**0.000**
Restricted sample#	0.207	0.188	0.110	**0.097*** (0.033)**	**0.077** (0.031)**	**0.008**
B.3 Age	17.058	17.048	16.999	0.059 (0.079)	0.049 (0.078)	0.737
B.4 Gender						
All schools	0.534	0.512	0.508	**0.025* (0.015)**	0.004 (0.015)	0.192
Restricted sample#	0.548	0.524	0.533	0.015 (0.015)	-0.009 (0.015)	0.277
B.5 Class size						
All schools	52.26	56.42	60.00	**-7.741* (4.045)**	-3.581 (4.672)	0.146
Restricted sample#	52.15	56.56	55.14	-2.985 (3.988)	1.428 (4.651)	0.489

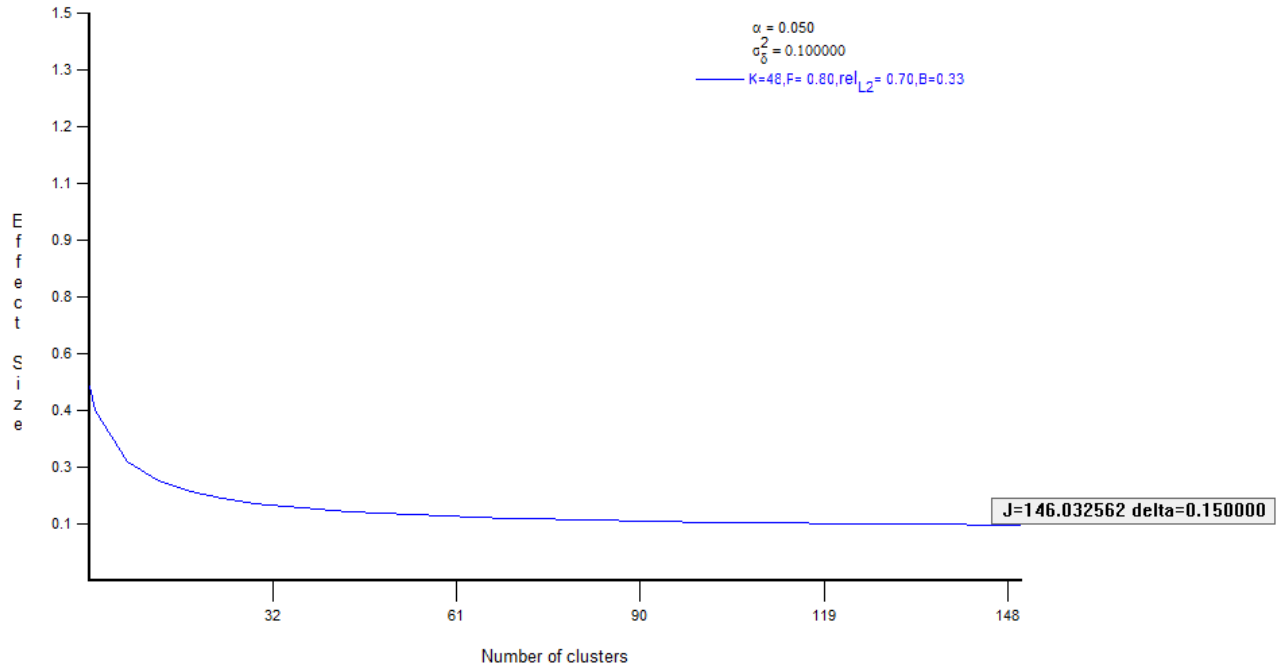
Note: comparison of mean characteristics of students in treatment and control groups. T1/T2 stands for within-/across- class social comparison groups and C for control group. Robust clustered standard errors at class level are in parentheses, adjusted for stratification. There was one school which experienced substantial reorganization (though exogenous to the intervention) resulting in a change in headmaster and a high number of student dropouts. Restricted sample (#) excludes that school from the analysis. Attrition rate is defined as the number of students missing in the last testing round conditional on student participation in the baseline testing. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A1.3: Randomization balance, by treatment decomposition

	Sum		Math		English	
	Mean	Difference from pure control	Mean	Difference from pure control	Mean	Difference from pure control
	(1)	(2)	(3)	(4)	(5)	(6)
NON-INTERACTED TREATMENT EFFECTS						
Within-class competition with no rewards (T1_solo)	19.551	**-3.136*** **(1.792)**	7.126	**-1.439*** **(0.771)**	12.425	-1.698 (1.096)
Across-class competition with no rewards (T2_solo)	21.575	-1.284 (1.814)	8.068	-0.559 (0.706)	13.507	-0.725 (1.178)
Financial rewards with no feedback (Fin_solo)	20.528	-1.891 (2.127)	7.719	-0.751 (1.035)	12.809	-1.139 (1.163)
Reputation rewards with no feedback (Rep_solo)	24.288	2.071 (1.898)	9.366	0.976 (0.815)	14.922	1.095 (1.191)
TREATMENT INTERACTIONS						
Within-class competition with financial rewards (T1_FIN)	24.111	1.728 (2.339)	8.485	0.029 (0.934)	15.625	1.698 (1.458)
Across-class competition with financial rewards (T2_FIN)	23.326	1.215 (1.700)	9.002	0.651 (0.719)	14.324	0.563 (1.117)
Within-class competition with reputation rewards (T1_REP)	22.734	0.453 (1.685)	8.834	0.418 (0.719)	13.899	0.035 (1.052)
Across-class competition with reputation rewards (T2_REP)	25.454	**3.304**** **(1.491)**	9.974	**1.606**** **(0.767)**	15.479	**1.698**** **(0.808)**
Pure control	22.697	0	8.583	0	14.115	0
Joint p-value	0.028	-	0.069	-	0.039	-

Note: Mean comparisons. Columns (1), (3) and (5) represent average scores from sum of Math and English, and separate scores by subject. Columns (2), (4), (6) represent differences between a particular treatment and the control group which received no feedback or reward). Robust standard errors adjusted for clustering at school level are in parentheses; adjusted for stratification. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A1.4: Randomization balance, by treatment decomposition



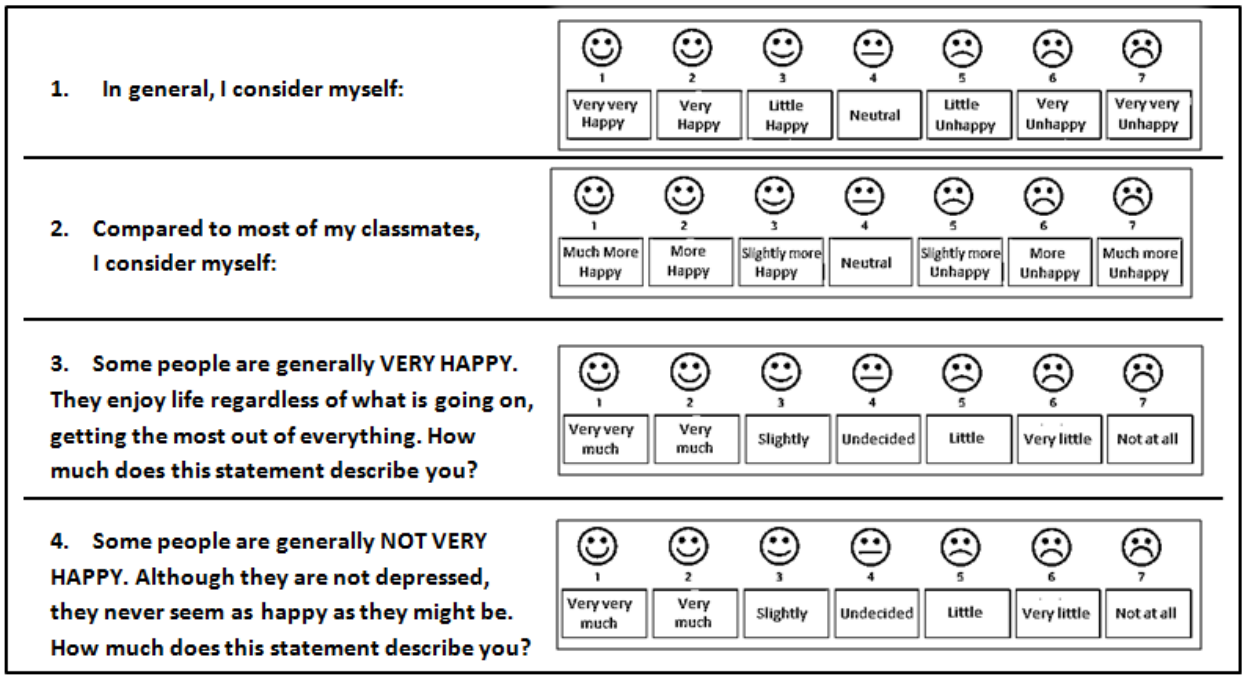
Note: y-axis represents minimal detectable effect size given parameters, x-axis shows the number of clusters required for a particular desired effect size; α is conventional significance level, K is the total number of sites, σ_δ^2 represents effect size variability, P stands for desired power (typically set to 0.800); relL2 stands for cluster level reliability, and B stands for the proportion of explained variance by the blocking variable(s). The result shows that for a given number of clusters (146 classes) the minimum detectable effect size equals 0.15.

APPENDIX B1: Questionnaires, score cards and supporting documents

Appendix B1.1: Short version of the Perceived Stress Scale

In the last month, how often did you feel or think a certain way? For each option please indicate the strength of your feelings. To indicate, make a tick ✓ .						
		Never	Rarely	Some-times	Often	Always
5.	How often have you felt unable to control the important things in your life?					
6.	How often have you felt confident about your ability to handle personal problems?					
7.	How often have you felt that things were going your way?					
8.	How often have you felt difficulties were piling up so high that you could not overcome them?					

Appendix B1.2: Short version of the Perceived Happiness Scale



Appendix B1.3: Letter of accordance from Archdiocese Caritas Prague

Arcidiecézní charita Praha
Rozvojové středisko, Adopce na dálku®
Londýnská 44, Praha 2, 120 00
Tel.: 224 246 573, 224 246 519
e-mail: adopce@charita-adopce.cz; www.charita-adopce.cz

Prague, November 8th 2010

Letter of accordance

I, below undersigned, hereby confirm the cooperation of Archdiocese Caritas Prague and its district office Uganda-Czech Development Trust with the researcher Dagmara Katrienaková (CERGE-EI) and her supervisor Michal Bauer (CERGE-EI and Charles University, Prague), for the purpose of the program evaluation titled Information and motives to learn.

Our organization has been providing the Child sponsorship programs in Uganda, India, Democratic Republic of Congo, Zambia, Thailand and Belarus since 1993. Currently, we sponsor over 16,000 children and the core of the program is their education. Thus, apart from improving chances for children to go to school, the quality-enhancement of education is in the best accord with our mission.

The research question, whether different information sets influence students' effort differently, will help us understand, how to motivate children from our Child sponsorship program and what their aspirations are. The results of the experiment can be presented further to our donors. Moreover, we consider the design of the experiment in the form of a game suitable for the conditions the researcher will face in the field.

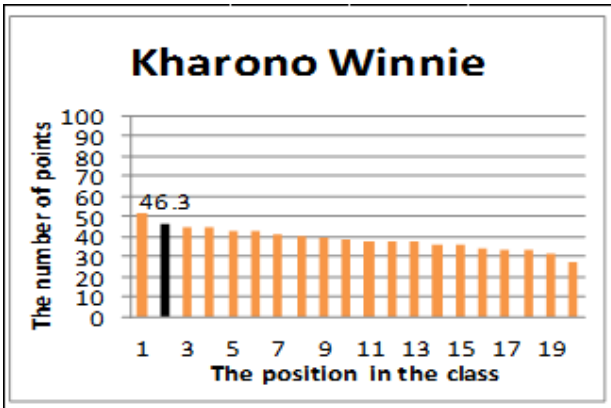
Archdiocese Caritas Prague fully supports the evaluation of the proposed intervention and is ready to provide local knowledge and cooperation with local field workers. For the purpose of the evaluation, two local workers will be hired and trained. These local workers will help Dagmara to organize meetings with sponsored children, to ensure fluent communication with them and to distribute cards after exams.

We are looking forward to a cooperating with Dagmara Katreniaková and sincerely hope that the proposed intervention will have the anticipated effect, which will be fully documented and reliably scientifically evaluated.

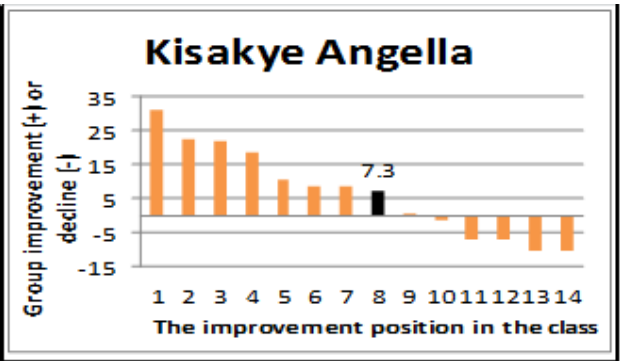
Jarmila Lomozova
Head of Development Centre
Archdiocese Caritas Prague

Arcidiecézní charita Praha
Rozvojové středisko
Londýnská 44, 120 00 Praha 2
Tel.: 224 246 519, 224 246 573
IČ: 43873499 Fax: 224 250 985

Appendix B1.4: Score cards for students in within-class comparison group

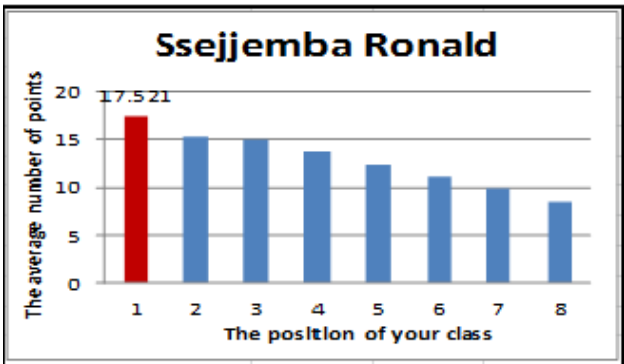


	Math	English	Together
Suubo Delia	26.0	26.0	52.0
Namugalula M.	24.0	22.0	46.0
Kharono Winnie	23.0	18.0	41.0
Your group achieved			**139.0**
Your group average is			**46.3**
Your group position is			**2**

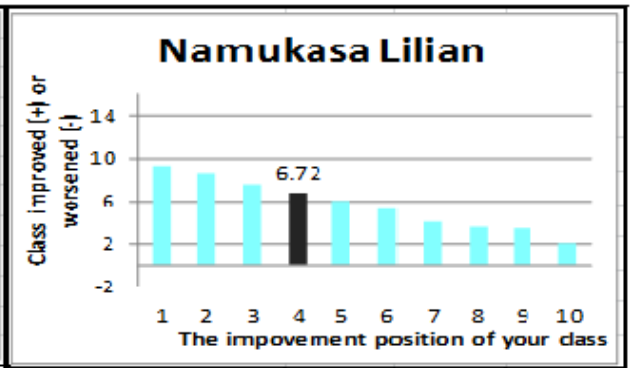


	Math	English	Together
Kisakye Angella	-12.0	19.5	7.5
Nakiwala Winnie	5.0	-18.0	-13.0
Busagwa Richard	31.0	-3.5	27.5
Your score	**INCREASED BY**		**7.5**
Group score	**INCREASED BY**		**22.0**
Group average improvement is			**7.3**
Group improvement position is			**8**
Your GROUP position in class			**WORSEN**

Appendix B1.5: Score cards for students in across-class comparison group



P7 - Maranatha Orphanage PS			
	Math	English	Together
Ssejjemba Ronald	11.0	6.5	17.5
Your class achieved	484.0	357.0	841.0
The number of students participating			48
MATH average of your class is			10.08
ENGLISH average of your class is			7.44
THE TOTAL AVERAGE of your class is			**17.52**
The position of your class is			**1**



	Math	English	Together
Namukasa Lilian	4.0	4.5	8.5
Did you improve?	YES		
The number of students			57
Improvement of your class in Math			3.23
Improvement of your class in English			3.48
THE TOTAL AVERAGE IMPROVEMENT			6.72
THE POSITION OF YOUR CLASS IS			4

Appendix C1: Average treatment effects

Appendix C1.1: Calculation of aggregated and dis-aggregated average treatment effects

	Financial rewards (FIN)	**Reputational rewards (REP)**	**No rewards (reward-control)**
Within-class feedback (T1)	T1_FIN [n1 number of observations]	T1_REP [n2 number of observations]	T1_solo [n3 number of observations]
Across-class feedback (T2)	T2_FIN [n4 number of observations]	T2_REP [n5 number of observations]	T2_solo [n6 number of observations]
No feedback (feedback-control)	C_FIN [n7 number of observations]	C_REP [n8 number of observations]	C_solo [n9 number of observations]

(1) The aggregated effect of the provision of within-class feedback:

$$ATE_{aggr}^{T1} = \left|\left(\frac{n1}{N}T1\_FIN + \frac{n2}{N}T1\_REP + \frac{n3}{N}T1\_solo\right) - \left(\frac{n7}{N}C\_FIN + \frac{n8}{N}C\_REP + \frac{n9}{N}C\_solo\right)\right|$$

(2) The aggregated effect of the provision of financial reward:

$$ATE_{aggr}^{FIN} = \left|\left(\frac{n1}{N}T1\_FIN + \frac{n4}{N}T2\_FIN + \frac{n7}{N}C\_FIN\right) - \left(\frac{n3}{N}T1_{SOLO} + \frac{n6}{N}T2\_SOLO + \frac{n9}{N}C\_solo\right)\right|$$

(3) The disaggregated effect of the provision of within-class feedback (without any rewards):

$$ATE_{dis-aggr}^{T1_{pure}} = \left|\left(\frac{n1}{N}T1\_solo\right) - \left(\frac{n9}{N}C\_solo\right)\right|$$

(4) The disaggregated effect of the provision of within-class feedback:

$$ATE_{dis-aggr}^{FIN_{pure}} = \left|\left(\frac{n7}{N}C\_FIN\right) - \left(\frac{n9}{N}C\_solo\right)\right|$$

Appendix C1.2: Aggregated average treatment effects, by subject

Dependent variable:	MATH						ENGLISH					
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Feedback provided (T)	**0.102****				**0.094***		-0.001				-0.009	
	(0.051)				**(0.051)**		(0.037)				(0.036)	
Rewards provided (Rew)		**0.138****			**0.128****			**0.125****			**0.126****	
		(0.066)			**(0.063)**			**(0.056)**			**(0.055)**	
Within-class feedback			**0.112***			**0.099***			-0.015			-0.028
(T1)			**(0.059)**			**(0.059)**			(0.042)			(0.039)
Across-class feedback			**0.093***			0.089			0.014			0.012
(T2)			**(0.055)**			(0.056)			(0.042)			(0.040)
Financial Rewards				**0.151***		**0.142***				**0.153****		**0.158****
(Finrew)				**(0.082)**		**(0.078)**				**(0.066)**		**(0.065)**
Reputational Rewards				**0.127***		**0.115***				**0.103***		**0.108****
(Reprew)				**(0.066)**		**(0.064)**				**(0.054)**		**(0.053)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5102	5102	5102	5102	5102	5102	5093	5093	5093	5093	5093	5093

Note: OLS. The treatment effects are calculated with respect to the control group. Students in group "T" are those who received any type of feedback (either within-class feedback, T1, or across-class feedback, T2). Students in the group "Rew" received any type of reward (either financial reward, Finrew, or reputational reward, Reprew). Columns (1) – (6) represent the treatment effects in Math, columns (7) – (12) in English. In all specifications I controlled for stratum fixed effects (area, level of study and school performance in national examination). N stands for the number of observations. Robust standard errors adjusted for clustering at school level are in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix C1.3: Aggregated average treatment effects of the provision of feedback or rewards on Math score at different points in time

Dependent variable:	MATH							
Specification:	Only FB (round 4)	Only FB (round 4)	Only FB (round 5)	Only FB (round 5)	Only Rewards (round 5)	Only Rewards (round 5)	Mix FB and Rewards (round 5)	Mix FB and Rewards (round 5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Within-class social comparison (T1)	0.024	0.037	0.084	**0.112***			0.086	**0.099***
	(0.062)	(0.048)	(0.081)	**(0.059)**			(0.079)	**(0.059)**
Across-class social comparison (T2)	0.005	0.043	0.024	**0.093***			0.046	0.089
	(0.058)	(0.043)	(0.084)	**(0.055)**			(0.081)	(0.056)
Financial Rewards					**0.231****	**0.151***	**0.233****	**0.142***
					(0.092)	**(0.082)**	**(0.093)**	**(0.078)**
Reputational Rewards					**0.185****	**0.127***	**0.184****	**0.115***
					(0.079)	**(0.066)**	**(0.078)**	**(0.064)**
Controlled for stratas	No	Yes	No	Yes	No	Yes	No	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5245	5245	5102	5102	5102	5102	5102	5102

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Columns (2), (4), (6) and (8) control for stratum fixed effects (areas (by distance from the capital city, Kampala), school performance at national examination and grade level; P6,P7, S1 up to S4). N stands for the number of observations. The first two columns analyze the effect in testing round 4 and the baseline testing in round 1. The remaining estimates are based on the differences between round 5 (the final testing round) and the baseline round 1. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.4: Aggregated average treatment effects of the provision of feedback or rewards on English score at different points in time

Dependent variable:	ENGLISH							
Specification:	Only FB (round 4)	Only FB (round 4)	Only FB (round 5)	Only FB (round 5)	Only Rewards (round 5)	Only Rewards (round 5)	Mix FB and Rewards (round 5)	Mix FB and Rewards (round 5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
OVERALL TREATMENT EFFECTS								
Within-class social comparison (T1)	-0.040	0.023	-0.102	-0.015			**-0.099***	-0.028
	(0.074)	(0.043)	(0.067)	(0.042)			**(0.058)**	(0.039)
Across-class social comparison (T2)	0.027	0.062	-0.039	0.014			-0.007	0.012
	(0.073)	(0.042)	(0.071)	(0.042)			(0.064)	(0.040)
Financial Rewards					**0.336*** **	**0.153** **	**0.340*** **	**0.158** **
					(0.055)	**(0.066)**	**(0.052)**	**(0.053)**
Reputational Rewards					**0.250** **	**0.103***	**0.254*** **	**0.108** **
					(0.066)	**(0.054)**	**(0.067)**	**(0.053)**
Controlled for stratas	No	Yes	No	Yes	No	Yes	No	Yes
Interactions	No	No	No	No	No	No	No	No
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5246	5246	5093	5093	5093	5093	5093	5093

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Columns (2), (4), (6) and (8) control for stratum fixed effects (areas (by distance from the capital city, Kampala), school performance at national examination and grade level; P6,P7, S1 up to S4). N stands for the number of observations. The first two columns analyze the effect in testing round 4 and the baseline testing in round 1. The remaining estimates are based on the differences between round 5 (the final testing round) and the baseline round 1. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.5: Estimation of the dis-aggregated treatment effects of different incentive schemes on performance, different specifications

Dependent variable:	Math					English				
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Within-class feedback,	0.100	0.104	0.101	0.100	0.095	**-0.128****	**-0.127****	**-0.134****	**-0.129****	**-0.134****
no feedback (T1_solo)	(0.085)	(0.085)	(0.085)	(0.084)	(0.087)	**(0.056)**	**(0.056)**	**(0.056)**	**(0.055)**	**(0.055)**
Across-class feedback,	0.082	0.081	0.077	0.081	0.070	-0.049	-0.049	-0.052	-0.045	-0.043
no feedback (T2_solo)	(0.073)	(0.073)	(0.074)	(0.075)	(0.075)	(0.059)	(0.059)	(0.059)	(0.061)	(0.060)
Financial Rewards,	0.106	0.097	0.101	0.104	0.085	0.045	0.043	0.042	0.053	0.063
no feedback (Fin_solo)	(0.101)	(0.099)	(0.101)	(0.105)	(0.106)	(0.088)	(0.089)	(0.088)	(0.087)	(0.089)
Reputational Rewards,	0.138	0.115	0.137	0.135	0.135	0.016	0.009	0.016	0.027	0.035
no feedback (Rep_solo)	(0.141)	(0.142)	(0.142)	(0.145)	(0.145)	(0.082)	(0.084)	(0.081)	(0.082)	(0.083)
Within-class feedback,	**0.231***	**0.215***	**0.229***	**0.227***	**0.215***	0.103	0.099	0.097	0.124	0.108
monetary reward (T1_fin)	**(0.118)**	**(0.115)**	**(0.119)**	**(0.127)**	**(0.120)**	(0.094)	(0.095)	(0.093)	(0.097)	(0.091)
Across-class feedback,	**0.277****	**0.281****	**0.279****	**0.272***	**0.251***	**0.173***	**0.174***	**0.166***	**0.193****	**0.189****
monetary reward (T2_fin)	**(0.139)**	**(0.136)**	**(0.141)**	**(0.146)**	**(0.144)**	**(0.094)**	**(0.093)**	**(0.093)**	**(0.097)**	**(0.090)**
Within-class feedback,	**0.209****	**0.196****	**0.203****	**0.205***	**0.212***	0.087	0.083	0.081	0.104	0.097
reputat.reward (T1_rep)	**(0.103)**	**(0.099)**	**(0.103)**	**(0.111)**	**(0.106)**	(0.080)	(0.081)	(0.079)	(0.086)	(0.082)
Across-class feedback,	**0.188****	**0.138***	**0.183****	**0.185****	**0.179****	0.047	0.034	0.043	0.059	0.053
reputat.reward (T2_rep)	**(0.080)**	**(0.079)**	**(0.081)**	**(0.087)**	**(0.086)**	(0.080)	(0.085)	(0.079)	(0.081)	(0.081)
Average class size		0.002					0.001			
		(0.002)					(0.001)			
Gender			**-0.068*****					**0.052*****		
			(0.022)					**(0.020)**		
Public				-0.009					0.041	
				(0.063)					(0.049)	
Food					**0.089*****					**0.058*****
					(0.029)					**(0.021)**
Baseline score	**0.725*****	**0.723*****	**0.719*****	**0.725*****	**0.723*****	**0.737*****	**0.737*****	**0.737*****	**0.736*****	**0.732*****
	(0.017)	**(0.017)**	**(0.017)**	**(0.017)**	**(0.017)**	**(0.016)**	**(0.016)**	**(0.016)**	**(0.016)**	**(0.016)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5102	5102	5065	5102	4906	5093	5093	5056	5093	4896

Note: Robust standard errors adjusted for clustering at class level are in parentheses. Public equals 1 if the school is public school, 0 otherwise. Food equals 1 if student received food a day before, 0 otherwise. Controlled for stratum fixed effects (area, level and school performance in national examinations) * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.6: Estimation of the aggregated treatment effects of different incentive schemes on performance, different specifications

Dependent variable:	Overall Performance (Math and English pooled)									
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Aggregated Feedback (T)	0.051	**0.069***	**0.067***	0.052	0.059					
	(0.039)	**(0.041)**	**(0.040)**	(0.039)	(0.039)					
Aggregated Rewards (REW)	**0.136****	**0.145****	**0.133****	**0.160*****	**0.137****					
	(0.052)	**(0.055)**	**(0.056)**	**(0.055)**	**(0.057)**					
Within-class feedback (T1)						0.061	**0.079***	0.068	0.059	0.058
						(0.043)	**(0.045)**	(0.045)	(0.044)	(0.045)
Across-class feedback (T2)						0.069	**0.088***	0.068	0.048	0.061
						(0.046)	**(0.048)**	(0.045)	(0.045)	(0.045)
Financial Rewards (Finrew)						**0.176*****	**0.185*****	**0.180*****	**0.201*****	**0.174****
						(0.062)	**(0.063)**	**(0.064)**	**(0.064)**	**(0.068)**
Reputational Rewards (Reprew)						**0.102****	**0.106***	0.090	**0.119****	**0.104***
						(0.051)	**(0.055)**	(0.055)	**(0.056)**	**(0.057)**
Stress		0.001					-0.001			
		(0.004)					(0.004)			
Happiness			0.005					0.005		
			(0.003)					(0.003)		
Change in Stress				**-0.008*****					**-0.008*****	
				(0.003)					**(0.002)**	
Change in Happiness					0.004					0.004
					(0.002)					(0.002)
Controlled for strata	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5108	4278	4226	4096	4047	5108	4278	4226	4096	4047

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. Students in group "T" are those who received any type of feedback (either within-class feedback, T1, or across-class feedback, T2). Students in the group "Rew" received any type of reward (financial reward, Finrew, or reputational reward, Reprew). Stress (Happiness) measures baseline level of students' stress (happiness). Change in stress (happiness) measures change in stress (happiness level) between endline and baseline testing. The treatment effects are calculated with respect to the control group. Controlled for stratum fixed effects (area, level and school performance in national examinations). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.7: Estimation of the dis-aggregated treatment effects of different incentive schemes on performance, different specifications

Dependent variable:	Overall Performance (Math and English pooled)				
Specification:	(1)	(2)	(3)	(4)	(5)
Within-class feedback, no feedback	0.017	0.029	0.027	0.028	0.027
(T1_solo)	(0.060)	(0.065)	(0.064)	(0.066)	(0.066)
Across-class feedback, no feedback	0.044	0.043	0.047	0.033	0.044
(T2_solo)	(0.059)	(0.062)	(0.060)	(0.063)	(0.062)
Financial Rewards, no feedback	**0.129***	**0.119***	0.103	**0.119***	0.094
(Fin_solo)	**(0.068)**	**(0.069)**	(0.066)	**(0.071)**	(0.067)
Reputational Rewards, no feedback	0.062	0.051	0.052	0.107	0.089
(Rep_solo)	(0.106)	(0.109)	(0.111)	(0.105)	(0.109)
Within-class feedback, monetary reward	**0.201****	**0.227****	**0.217****	**0.241****	**0.209****
(T1_fin)	**(0.094)**	**(0.094)**	**(0.096)**	**(0.092)**	**(0.098)**
Across-class feedback, monetary reward	**0.282****	**0.301*****	**0.287****	**0.293****	**0.282****
(T2_fin)	**(0.113)**	**(0.112)**	**(0.114)**	**(0.113)**	**(0.116)**
Within-class feedback, reputat.reward	**0.187****	**0.187****	**0.178****	**0.187****	**0.172****
(T1_rep)	**(0.073)**	**(0.077)**	**(0.074)**	**(0.077)**	**(0.078)**
Across-class feedback, reputat.reward	**0.122***	**0.153***	**0.135***	**0.139***	**0.139****
(T2_rep)	**(0.073)**	**(0.078)**	**(0.072)**	**(0.080)**	**(0.075)**
Stress		-0.001			
		(0.004)			
Happiness			0.005		
			(0.003)		
Change in stress				**-0.007*****	
				(0.003)	
Change in happiness					0.003
					(0.002)
Controlled for stratas	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	5108	4278	4226	4096	4047

Note: Robust standard errors adjusted for clustering at class level are in parentheses. Stress (Happiness) measures baseline level of students' stress (happiness). Change in stress (happiness) measures change in stress (happiness level) between endline and baseline testing. Controlled for stratum fixed effects (area, level and school performance in national examinations). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.8: Aggregated average treatment effects of the provision of feedback or rewards on overall score at different points in time

	Scenario 1 Round 2 versus baseline testing	**Scenario 2** Round 3 versus baseline testing	**Scenario 3** Round 5 versus baseline testing	**Scenario 4** Round 5 versus Round 4
Aggregated within-class social comparison (T1)	-0.001	**0.041***	0.061	0.019
	(0.038)	**(0.021)**	(0.043)	(0.034)
Aggregated across-class social comparison (T2)	0.010	**0.051****	0.069	0.019
	(0.042)	**(0.021)**	(0.046)	(0.039)
Aggregated financial rewards (Finrew)			**0.176*****	**0.159*****
			(0.062)	**(0.051)**
Aggregated Reputational rewards (Reprew)			**0.102****	**0.116*****
			(0.051)	**(0.032)**
Control for strata	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000
Number of observations	5821	5482	5108	4911

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Controlled for stratum fixed effects (area, level and school performance in national examinations) * significant at 10%; ** significant at 5%; *** significant at 1%

Differences of the immediate effects of feedback versus reward (p-values reported):		Scenario 1 (Round 2 versus baseline testing)	
Scenario 3 (Round 5 versus baseline testing)	p-values	Within-class comparison	Across-class comparison
	Financial rewards	**0.016**	**0.019**
	Reputational Rewards	0.101	0.144

Differences of the immediate effects of feedback versus reward (p-values reported):		Scenario 2 (Round 3 versus baseline testing)	
Scenario 3 (Round 5 versus baseline testing)	p-values	Within-class comparison	Across-class comparison
	Financial rewards	**0.054**	**0.064**
	Reputational Rewards	0.348	0.422

Differences of the immediate effects of feedback versus reward (p-values reported):		Scenario 1 (Round 2 versus baseline testing)	
Scenario 4 (Round 5 versus Round 4)	p-values	Within-class comparison	Across-class comparison
	Financial rewards	**0.009**	**0.010**
	Reputational Rewards	**0.039**	**0.056**

Differences of the immediate effects of feedback versus reward (p-values reported):		Scenario 2 (Round 3 versus baseline testing)	
Scenario 4 (Round 5 versus Round 4)	p-values	Within-class comparison	Across-class comparison
	Financial rewards	**0.061**	**0.067**
	Reputational Rewards	0.196	0.244

Note:

Scenario 1:

$$Performance_{ROUND\_2} = \alpha_0 + \alpha_1 T1 + \alpha_2 T2 + \alpha_3 Performance_{ROUND\_1} + \alpha_i X_i + \varepsilon$$

Scenario 2:

$$Performance_{ROUND\_3} = \beta_0 + \beta_1 T1 + \beta_2 T2 + \beta_3 Performance_{ROUND\_1} + \beta_i X_i + \tau$$

Scenario 3:

$$Performance_{ROUND\_5} = \gamma_0 + \gamma_1 T1 + \gamma_2 T2 + \gamma_3 FinRew + \gamma_4 RepRew + \gamma_5 Performance_{ROUND\_1} + \gamma_i X_i + \mu$$

Scenario 4:

$$Performance_{ROUND\_5} = \delta_0 + \delta_1 T1 + \delta_2 T2 + \delta_3 FinRew + \delta_4 RepRew + \delta_5 Performance_{ROUND\_4} + \delta_i X_i + \vartheta$$

where T1 stands for within-class feedback group (=1 if treated), T2 for across-class feedback group, FinRew for financially rewarded treatment group, RepRew for reputationally rewarded group, and X is a vector of stratification variables (school area, performance in the national examination and class level); p-values come from testing differences between estimated coefficients from regressions in various scenarios.

Appendix C1.9a: Sensitivity analysis

Dependent variable:	Performance (Math and English pooled)			
Sample:	Overall sample	Overall sample	Overall sample	Overall sample
Within-class feedback	0.061		**0.079***	0.065
	(0.043)		**(0.045)**	(0.044)
Across-class feedback	0.069		**0.088***	0.062
	(0.046)		**(0.048)**	(0.045)
Financial rewards	**0.176*****		**0.185*****	**0.194*****
	(0.062)		**(0.063)**	**(0.063)**
Reputation rewards	**0.102****		**0.106***	**0.124****
	(0.051)		**(0.055)**	**(0.056)**
Initial stress level			-0.001	
			(0.004)	
Change in stress		**-0.007*****		**-0.008*****
		(0.003)		**(0.003)**
Controlled for strata	Yes	Yes	Yes	Yes
Prob > F	0.716	0.719	0.716	0.723
R-squared	0.0000	0.0000	0.0000	0.0000
Number of observation	5108	4096	4278	4096

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Controlled for stratum fixed effects (area, level and school performance in national examinations). Initial stress level stands for the stress level in baseline testing. Change in stress stands for change in stress level between endline and baseline testing. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.9b: Sensitivity analysis

Dependent variable:	Stress					
Sample:	Overall sample	Overall sample	Overall sample	Overall sample	Decreased or equal performance	Increased performance
Within-class competition	0.003		-0.000	0.017	0.153	-0.003
	(0.233)		(0.232)	(0.234)	(0.332)	(0.239)
Across-class competition	-0.275		-0.267	-0.266	0.217	-0.330
	(0.213)		(0.211)	(0.212)	(0.361)	(0.215)
Financial rewards	**0.549****		**0.585****	**0.578****	0.216	**0.579****
	(0.277)		**(0.275)**	**(0.276)**	(0.549)	**(0.276)**
Reputation rewards	0.408		0.453	0.419	0.762	0.367
	(0.304)		(0.306)	(0.304)	(0.471)	(0.303)
Initial performance level			**-0.180*****			
			(0.002)			
Change in performance		**-0.016*****		**-0.018*****		
		(0.005)		**(0.002)**		
Controlled for strata	Yes	Yes	Yes	Yes	Yes	Yes
Number of observation	4105	4096	4105	4096	465	3640
R-squared	0.462	0.460	0.465	0.465	0.472	0.464
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Controlled for stratum fixed effects (area, level and school performance in national examinations). Initial performance stands for pooled score from Math and English in the baseline testing. Change in performance stands for difference in scores between endline and baseline testing. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.9c: Sensitivity analysis

Dependent variable:	Happiness					
Sample:	Overall sample	Overall sample	Overall sample	Overall sample	Decreased or equal performance	Increased performance
Within-class competition	**0.428***		**0.426***	**0.458****	-0.724	**0.587****
	(0.224)		**(0.224)**	**(0.222)**	(0.452)	**(0.231)**
Across-class competition	0.219		0.220	0.242	**-0.866***	**0.370***
	(0.219)		(0.218)	(0.219)	**(0.504)**	**(0.219)**
Financial rewards	0.398		0.411	**0.433***	**1.252****	0.370
	(0.267)		(0.258)	**(0.254)**	**(0.572)**	(0.259)
Reputation rewards	0.408		0.424	0.423	0.659	0.423
	(0.267)		(0.270)	(0.267)	(0.556)	(0.271)
Initial performance level			-0.054			
			(0.087)			
Change in performance		**-0.019****		**-0.023****		
		(0.009)		**(0.002)**		
Controlled for strata	Yes	Yes	Yes	Yes	Yes	Yes
Number of observation	4056	4047	4056	4047	464	3592
R-squared	0.394	0.392	0.393	0.395	0.334	0.405
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. The treatment effects are calculated with respect to the control group. Initial performance stands for pooled score from Math and English in the baseline testing. Controlled for stratum fixed effects (area, level and school performance in national examinations). Change in performance stands for difference in scores between endline and baseline testing. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C1.10: Aggregated average treatment effects of the provision of feedback and rewards on the overall performance and students' subjective well-being

Dependent variable:	STUDENTS' OVERALL PERFORMANCE				STRESS	HAPPINESS
	(1)	(2)	(3)	(4)	(5)	(6)
Aggregated feedback treatment pooled	**0.073***	0.064				
	(0.041)	(0.039)				
Aggregated reward treatment pooled		**0.133****				
		(0.052)				
Within-class social comparison, aggregated			0.074	0.061	-0.001	**-0.111***
			(0.045)	(0.043)	(0.090)	**(0.058)**
Across-class social comparison, aggregated			0.071	0.069	-0.104	-0.058
			(0.047)	(0.046)	(0.082)	(0.057)
Financial Rewards, aggregated				**0.176*****	**0.226****	-0.108
				(0.062)	**(0.107)**	(0.070)
Reputational Rewards, aggregated				**0.102****	0.177	-0.112
				(0.051)	(0.119)	(0.070)
Baseline testing	**0.841*****	**0.837*****	**0.841*****	**0.836*****	**0.079*****	**-0.219*****
	(0.015)	**(0.015)**	**(0.015)**	**(0.014)**	**(0.019)**	**(0.019)**
National performance	**0.109*****	**0.104*****	**0.109*****	**0.099*****	-0.116	0.016
	(0.039)	**(0.038)**	**(0.039)**	**(0.038)**	(0.079)	(0.044)
Area 2 (Kampala-Jinja road)	**0.254*****	**0.195*****	**0.254*****	**0.189*****	**-0.313*****	**0.135***
	(0.049)	**(0.062)**	**(0.049)**	**(0.054)**	**(0.101)**	**(0.074)**
Area 3 (Buikwe area)	**0.447*****	**0.368*****	**0.447*****	**0.325*****	**-0.237***	0.098
	(0.041)	**(0.059)**	**(0.042)**	**(0.059)**	**(0.136)**	(0.083)
Area 4 (the most remote area close to Victoria Lake)	**0.304*****	**0.231*****	**0.304*****	**0.201*****	**-0.339****	**0.228*****
	(0.042)	**(0.064)**	**(0.042)**	**(0.061)**	**(0.131)**	**(0.081)**
Primary 6	-0.049	-0.044	-0.049	-0.039	**-0.225****	**0.208*****
	(0.059)	(0.059)	(0.059)	(0.058)	**(0.113)**	**(0.071)**
Primary 7	**0.433*****	**0.439*****	**0.433*****	**0.441*****	-0.009	**0.248*****
	(0.069)	**(0.070)**	**(0.069)**	**(0.066)**	(0.089)	**(0.078)**
Secondary 1	**0.103***	**0.106***	**0.103***	**0.112***	**0.208****	0.121
	(0.058)	**(0.058)**	**(0.058)**	**(0.057)**	**(0.080)**	(0.077)
Secondary 2	-0.018	-0.017	-0.018	-0.015	**0.233****	-0.005
	(0.053)	(0.053)	(0.053)	(0.053)	**(0.095)**	(0.084)
Secondary 3	**0.197*****	**0.194*****	0.197***	0.197***	-0.065	0.011
	(0.051)	**(0.050)**	(0.051)	(0.050)	(0.093)	(0.087)
R-squared	0.714	0.715	0.634	0.645	0.058	0.078
F-statistics	424.45***	430.01***	389.52***	389.69***	4.56***	16.66***
N	5108	5108	5102	5102	4105	4056

Note: OLS. Pooled aggregated feedback treatment effect consists of the aggregated treatment effects of within- and across-class feedbacks. Pooled aggregated reward treatment effect consists of the aggregated treatment effect of financial and reputational rewards. Columns (1) – (4) show the aggregated average treatment effects (ATE) of differently aggregated treatments on students' overall performance. Columns (5) and (6) represent the average treatment effects on students' well-being (stress and happiness respectively). Controlled for stratum fixed effects (area, level and school performance in national examinations). * significant at 10%; ** at 5%; *** at 1%.

Appendix D1: Gender differences

Appendix D1.1: Average treatment effects on students' performance and well-being, by gender

Dependent variable:	Mathematics	English	Sum
Within-class feedback, no feedback (T1_solo)	0.089 (0.107)	**-0.121*** **(0.072)**	0.019 (0.077)
Across-class feedback, no feedback (T2_solo)	0.011 (0.085)	-0.026 (0.072)	0.026 (0.065)
Financial Rewards, no feedback (Fin_solo)	**0.199*** **(0.119)**	0.128 (0.109)	**0.237***** **(0.077)**
Reputational Rewards, no feedback (Rep_solo)	0.215 (0.153)	0.063 (0.104)	0.131 (0.120)
Within-class feedback, monetary reward (T1_fin)	0.218 (0.133)	0.179 (0.112)	**0.235**** **(0.104)**
Across-class feedback, monetary reward (T2_fin)	0.276 (0.173)	**0.237**** **(0.112)**	**0.322***** **(0.114)**
Within-class feedback, reputational reward (T1_rep)	0.194 (0.131)	0.069 (0.095)	**0.165*** **(0.085)**
Across-class feedback, reputational reward (T2_rep)	**0.169*** **(0.097)**	0.030 (0.109)	0.096 (0.079)
Gender	-0.079 (0.060)	0.098 (0.061)	0.030 (0.055)
Within-class feedback, no feedback x gender (T1_solo x gender)	0.019 (0.078)	-0.023 (0.073)	-0.007 (0.073)
Across-class feedback, no feedback x gender (T2_solo x gender)	**0.123*** **(0.071)**	-0.044 (0.069)	0.028 (0.064)
Financial Rewards, no feedback x gender (Fin_solo x gender)	**-0.172**** **(0.086)**	-0.152 (0.109)	**-0.197**** **(0.083)**
Reputational Rewards, no feedback x gender (Rep_solo x gender)	**-0.155*** **(0.080)**	-0.086 (0.091)	-0.134 (0.081)
Within-class feedback, monetary reward x gender (T1_fin x gender)	0.016 (0.071)	**-0.148*** **(0.088)**	-0.076 (0.071)
Across-class feedback, monetary reward x gender (T2_fin x gender)	0.005 (0.166)	-0.118 (0.085)	-0.073 (0.099)
Within-class feedback, reputational reward x gender (T1_rep x gender)	0.015 (0.097)	0.018 (0.083)	0.020 (0.073)
Across-class feedback, reputational reward x gender (T2_rep x gender)	0.023 (0.089)	0.022 (0.099)	0.033 (0.074)
Baseline score	**0.719***** **(0.017)**	**0.737***** **(0.016)**	**0.835***** **(0.014)**
Controlled for stratas	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000
Number of observation	5065	5056	5071

Note: OLS. Controlled for stratum fixed effects (four areas, school performance at national examination and grade level). Robust standard errors clustered at class level are in brackets. * significant at 10%; ** at 5%; *** at 1%

Appendix D1.2: Aggregated average treatment effects on the overall performance (Math and English pooled), gender differences

Dependent variable:	OVERALL PERFORMANCE BY GIRLS						OVERALL PERFORMANCE BY BOYS					
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Feedback provided (T)	**0.104****						0.026				0.013	
	(0.043)				**0.097****		(0.048)				(0.047)	
Rewards provided (Rew)		0.067			**(0.041)**			**0.134****			**0.133****	
		(0.065)			0.042			**(0.063)**			**(0.063)**	
					(0.061)							
Within-class feedback			**0.098****			**0.089****			0.038			0.019
(T1)			**(0.058)**			**(0.044)**			(0.056)			(0.052)
Across-class feedback			**0.109****			**0.106****			0.016			0.021
(T2)			**(0.060)**			**(0.050**)			(0.053)			(0.051)
Financial Rewards				0.114		0.105				**0.259*****		**0.258*****
(Finrew)				(0.076)		(0.072)				**(0.066)**		**(0.065)**
Reputational Rewards				0.083		0.070				**0.129****		**0.128****
(Reprew)				(0.059)		(0.057)				**(0.059)**		**(0.059)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	2862	2862	2862	2862	2862	2862	2209	2209	2209	2209	2209	2209

Note: OLS. Students in group "T" are those who received any type of feedback (either within-class feedback, T1, or across-class feedback, T2). Students in the group "Rew" received any type of reward (either financial reward, Finrew, or reputational reward, Reprew). Columns (1) – (6) represent the treatment effects on performance of girls, columns (7) – (12) on boys' performance. In all specifications I controlled for stratas (i.e., students' performance in the national examinations, area and the level of studies). Robust standard errors adjusted for clustering at school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix D1.3: Girls' aggregated average treatment effects on performance in Math and English

Dependent variable:	MATH						ENGLISH					
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Feedback provided (T)	**0.159*** **				**0.154*** **		0.001				-0.007	
	(0.053)				**(0.053)**		(0.041)				(0.039)	
Rewards provided (Rew)		0.093			0.073			0.093			**0.094***	
		(0.076)			(0.071)			(0.058)			**(0.057)**	
Within-class feedback			**0.157*** **			**0.149** **			-0.016			-0.027
(T1)			**(0.058)**			**(0.058)**			(0.044)			(0.042)
Across-class feedback			**0.163*** **			**0.159*** **			0.019			0.014
(T2)			**(0.060)**			**(0.061)**			(0.046)			(0.045)
Financial Rewards				0.103		0.088				0.089		0.094
(Finrew)				(0.095)		(0.088)				(0.069)		(0.068)
Reputational Rewards				0.085		0.062				0.095		**0.099***
(Reprew)				(0.075)		(0.071)				(0.058)		**(0.056)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	2858	2858	2858	2858	2858	2858	2854	2854	2854	2854	2854	2854

Note: OLS. Students in group "T" are those who received any type of feedback (either within-class feedback, T1, or across-class feedback, T2). Students in the group "Rew" received any type of reward (financial reward, Finrew, or reputational reward, Reprew). Columns (1) – (6) represent the treatment effects in Math, columns (7) – (12) in English. In all specifications I controlled for stratas (i.e., students' performance at the national examinations, area and the level of studies). Robust standard errors adjusted for clustering at school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix D1.4: Boys' aggregated average treatment effects on performance in Math and English

Dependent variable:	MATH						ENGLISH					
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Feedback provided (T)	0.027				0.019		-0.011				-0.017	
	(0.061)				(0.060)		(0.047)				(0.046)	
Rewards provided (Rew)		**0.192*** **			**0.191*** **			**0.157** **			**0.158** **	
		(0.022)			**(0.069)**			**(0.068)**			**(0.068)**	
Within-class feedback			0.055			0.038			-0.022			-0.038
(T1)			(0.073)			(0.071)			(0.054)			(0.051)
Across-class feedback			-0.001			0.003			0.001			0.005
(T2)			(0.064)			(0.065)			(0.053)			(0.051)
Financial Rewards				**0.213** **		**0.207** **				**0.226*** **		**0.234*** **
(Finrew)				**(0.089)**		**(0.089)**				**(0.078)**		**(0.078)**
Reputational Rewards				**0.175** **		**0.170** **				0.106		**0.111***
(Reprew)				**(0.074)**		**(0.073)**				(0.066)		**(0.067**)
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	2207	2207	2207	2207	2207	2207	2202	2202	2202	2202	2202	2202

Note: Rows represent aggregated treatment groups. Students in group "T" are those who received any type of feedback (either within-class feedback (T1) or across-class feedback (T2)). Students in the group "Rew" received any type of reward (financial reward (Finrew) or reputational reward (Reprew)). Columns (1) – (6) represent the treatment effects in Math, columns (7) – (12) in English. In all specifications I controlled for stratas (i.e., students' performance at the national examinations, area and the level of studies). Robust standard errors adjusted for clustering at school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix D1.5: Gender differences in aggregated average treatment effects on stress and happiness

Dependent variable:	PERCEIVED STRESS						PERCEIVED HAPPINESS					
Sample:	**OVERALL**		**GIRLS**		**BOYS**		**OVERALL**		**GIRLS**		**BOYS**	
Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Feedback provided (T)	-0.133		-0.089		-0.183		**-0.326***		-0.313		-0.348	
	(0.204)		(0.234)		(0.216)		**(0.195)**		(0.214)		(0.251)	
Rewards provided (Rew)	**0.516***		**0.508***		**0.533***		**-0.437***		-0.349		**-0.544***	
	(0.269)		**(0.284)**		**(0.295)**		**(0.228)**		(0.255)		**(0.304)**	
Within-class feedback		0.003		0.105		-0.119		**-0.428***		**-0.399***		-0.483
(T1)		(0.233)		(0.259)		(0.255)		**(0.224)**		**(0.235)**		(0.297)
Across-class feedback		-0.275		-0.295		-0.248		-0.219		-0.221		-0.213
(T2)		(0.213)		(0.239)		(0.239)		(0.219)		(0.253)		(0.268)
Financial Rewards		**0.549****		**0.599****		0.477		-0.398		-0.331		-0.504
(Finrew)		**(0.277)**		**(0.282)**		(0.334)		(0.258)		(0.277)		(0.375)
Reputational Rewards		0.408		0.307		**0.547***		-0.408		-0.314		-0.493
(Reprew)		(0.304)		(0.331)		**(0.318)**		(0.267)		(0.298)		(0.327)
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observations	4105	4105	2279	2279	1821	1821	4056	4056	2264	2264	1788	1788

Note: Rows represent aggregated treatment groups. Students in group "T" are those who received any type of feedback (either within-class feedback, T1, or across-class feedback, T2). Students in the group "Rew" received any type of reward (financial reward, Finrew, or reputational reward, Reprew). Columns (1) – (6) represent the treatment effects on perceived stress, columns (7) – (12) on perceived happiness. In all specifications I controlled for stratas (i.e., students' performance at the national examinations, area and the level of studies). Robust standard errors adjusted for clustering at school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix D1.6: Average treatment effects on students' performance and well-being, by gender

Treatment group:	Only within-class feedback		Within-class feedback rewarded financially		Within-class feedback rewarded reputationally	
	Girls	**Boys**	**Girls**	**Boys**	**Girls**	**Boys**
Math (st.dev)	0.121	0.076	**0.229***	**0.228***	**0.201****	0.204
	(0.081)	(0.107)	**(0.118)**	**(0.137)**	**(0.102)**	(0.129)
English (st.dev.)	**-0.141****	-0.116	0.016	**0.199***	0.069	0.092
	(0.059)	(0.072)	(0.092)	**(0.116)**	(0.088)	(0.094)
Stress	0.072	0.043	**0.258****	0.143	0.178	**0.313***
	(0.124)	(0.119)	**(0.116)**	(0.189)	(0.159)	**(0.179)**
Happiness	0.023	**0.213***	**0.304*****	**0.282****	0.073	**0.297*****
	(0.087)	**(0.123)**	**(0.101)**	**(0.112)**	(0.115)	**(0.111)**
Confidence (Math)	**-7.385*****	**-4.13*****	**-6.104*****	**-4.07*****	**-5.324*****	**-6.604*****
	(0.929)	**(0.954)**	**(1.214)**	**(1.249)**	**(1.144)**	**(1.069)**
Confidence (English)	**-5.023*****	**-2.79*****	**-5.528*****	**-4.604*****	**-5.722*****	**-5.129*****
	(0.994)	**(0.909)**	**(1.375)**	**(1.363)**	**(1.115)**	**(1.193)**
Aspirations						
Education over work	-0.035	0.098	**0.163****	**0.146***	0.052	0.042
	(0.079)	(0.082)	**(0.081)**	**(0.086)**	(0.094)	(0.101)
Education over rest	0.017	**0.219*****	**0.109****	0.098	0.061	0.046
	(0.047)	**(0.068)**	**(0.044)**	(0.074)	(0.061)	(0.099)
Work over rest	0.038	-0.009	-0.043	**-0.267****	-0.027	-0.057
	(0.069)	(0.113)	(0.091)	**(0.110)**	(0.093)	(0.117)
Treatment group:	Only across-class feedback		Across-class feedback rewarded financially		Across-class feedback rewarded reputationally	
	Girls	**Boys**	**Girls**	**Boys**	**Girls**	**Boys**
Math	**0.135***	0.009	**0.275***	0.284	**0.189****	**0.175***
	(0.077)	(0.088)	**(0.159)**	(0.173)	**(0.091)**	**(0.103)**
English	-0.076	-0.019	0.108	**0.249****	0.041	0.042
	(0.066)	(0.072)	(0.101)	**(0.112)**	(0.083)	(0.103)
Stress	-0.099	0.016	-0.016	-0.022	**0.229***	0.286
	(0.119)	(0.119)	(0.146)	(0.155)	**(0.121)**	(0.174)
Happiness	0.020	0.124	-0.022	0.193	0.153	**0.241***
	(0.089)	(0.098)	(0.109)	(0.130)	(0.143)	**(0.122)**
Confidence (Math)	**-8.148*****	**-4.74*****	**-6.948*****	**-4.597*****	**-6.957*****	**-6.125*****
	(0.841)	**(1.083)**	**(1.170)**	**(1.538)**	**(1.406)**	**(1.675)**
Confidence (English)	**-6.013*****	**-4.49*****	**-6.528*****	**-4..047*****	**-6.411*****	**-5.327*****
	(0.980)	**(1.058)**	**(1.154)**	**(1.363)**	**(1.580)**	**(1.579)**
Aspirations						
Education over work	0.101	**0.174***	0.099	**0.219****	0.101	-0.026
	(0.072)	**(0.089)**	(0.093)	**(0.105)**	(0.091)	(0.136)
Education over rest	0.023	**0.140****	-0.049	-0.091	-0.006	0.109
	(0.044)	**(0.067)**	(0.069)	(0.096)	(0.066)	(0.087)
Work over rest	0.038	-0.103	-0.043	-0.011	-0.027	-0.069
	(0.069)	(0.100)	(0.091)	(0.120)	(0.093)	(0.112)

Note: OLS. Controlled for stratum fixed effects (four areas, school performance at national examination and grade level). Robust standard errors clustered at class level are in brackets. Prob > F = 0.0000. * significant at 10%; ** at 5%; *** at 1%

Appendix D1.7: Average treatment effects on students' performance and well-being, by gender

Treatment group:	Only financial rewards		Within-class feedback rewarded financially		Across-class feedback rewarded financially	
	Girls	**Boys**	**Girls**	**Boys**	**Girls**	**Boys**
Math (st.dev)	0.018	**0.207***	**0.229***	**0.228***	**0.275***	0.284
	(0.102)	**(0.123)**	**(0.118)**	**(0.137)**	**(0.159)**	(0.173)
English (st.dev.)	-0.038	0.139	0.016	**0.199***	0.108	**0.249****
	(0.097)	(0.112)	(0.092)	**(0.116)**	(0.101)	**(0.112)**
Stress	**0.431****	**0.482*****	**0.258****	0.143	-0.016	-0.022
	(0.198)	**(0.162)**	**(0.116)**	(0.189)	(0.146)	(0.155)
Happiness	0.015	**0.322****	**0.304*****	**0.282****	-0.022	0.193
	(0.014)	**(0.132)**	**(0.101)**	**(0.112)**	(0.109)	(0.130)
Confidence (Math)	**1.869***	-1.322	**-6.104*****	**-4.07*****	**-6.948*****	**-4.597*****
	(1.074)	(1.429)	**(1.214)**	**(1.249)**	**(1.170)**	**(1.538)**
Confidence (English)	**2.239****	-0.387	**-5.528*****	**-4.604*****	**-6.528*****	**-4..047*****
	(1.108)	(1.099)	**(1.375)**	**(1.363)**	**(1.154)**	**(1.363)**
Aspirations						
Education over work	0.046	0.006	**0.163****	**0.146***	0.099	**0.219****
	(0.098)	(0.111)	**(0.081)**	**(0.086)**	(0.093)	**(0.105)**
Education over rest	0.009	0.016	**0.109****	0.098	-0.049	-0.091
	(0.078)	(0.083)	**(0.044)**	(0.074)	(0.069)	(0.096)
Work over rest	-0.017	0.137	-0.043	**-0.267****	-0.043	-0.011
	(0.092)	(0.112)	(0.091)	**(0.110)**	(0.091)	(0.120)
Treatment group:	Only reputation rewards		Within-class feedback rewarded reputationally		Across-class feedback rewarded reputationally	
	Girls	**Boys**	**Girls**	**Boys**	**Girls**	**Boys**
Math	0.059	0.218	**0.201****	0.204	**0.189****	**0.175***
	(0.147)	(0.154)	**(0.102)**	(0.129)	**(0.091)**	**(0.103)**
English	-0.039	0.079	0.069	0.092	0.041	0.042
	(0.087)	(0.106)	(0.088)	(0.094)	(0.083)	(0.103)
Stress	-0.008	0.158	0.178	**0.313***	**0.229***	0.286
	(0.203)	(0.195)	(0.158)	**(0.179)**	**(0.121)**	(0.174)
Happiness	-0.005	0.144	0.073	**0.297*****	0.153	**0.241***
	(0.116)	(0.103)	(0.115)	**(0.111)**	(0.143)	**(0.122)**
Confidence (Math)	**1.905***	-0.399	**-5.324*****	**-6.604*****	**-6.957*****	**-6.125*****
	(0.972)	(1.224)	**(1.144)**	**(1.069)**	**(1.406)**	**(1.675)**
Confidence (English)	0.989	-1.301	**-5.722*****	**-5.129*****	**-6.411*****	**-5.327*****
	(1.096)	(1.008)	**(1.115)**	**(1.193)**	**(1.580)**	**(1.579)**
Aspirations						
Education over work	0.021	0.165	0.052	0.042	0.101	-0.026
	(0.096)	(0.101)	(0.094)	(0.101)	(0.091)	(0.136)
Education over rest	-0.017	0.109	0.061	0.046	-0.006	0.109
	(0.061)	(0.078)	(0.061)	(0.099)	(0.066)	(0.087)
Work over rest	**0.164***	-0.004	-0.027	-0.057	-0.027	-0.069
	(0.088)	(0.131)	(0.093)	(0.117)	(0.093)	(0.112)

Note: OLS. Controlled for stratum fixed effects (four areas, school performance at national examination and grade level). Robust standard errors clustered at class level are in brackets. Prob > F = 0.0000. * significant at 10%; ** at 5%; *** at 1%

Appendix E1: Group composition

Appendix E1.1: Comparison of average treatment effects on performance, exerted effort and subjective well-being, by group ability composition

Dependent variable:	Math	English	Perceived Stress	Subjective Happiness	Effort Math	Effort English
Mixed ability: only feedback	0.055	**0.133****	-0.082	0.096	-0.069	0.004
	(0.059)	**(0.063)**	(0.078)	(0.094)	(0.094)	(0.096)
High-ability: only feedback	**0.451*****	**0.387*****	0.001	0.078	0.082	0.126
	(0.129)	**(0.078)**	(0.161)	(0.176)	(0.096)	(0.081)
Mixed ability: feedback and monetary reward	**0.275****	**0.521*****	**0.313****	0.121	0.109	0.102
	(0.127)	**(0.086)**	**(0.124)**	(0.155)	(0.112)	(0.128)
High-ability: feedback and monetary reward	**0.561*****	**0.449*****	0.051	0.225	0.053	0.042
	(0.206)	**(0.099)**	(0.254)	(0.182)	(0.109)	(0.128)
Mixed ability: feedback and reputational reward	**0.262*****	**0.357*****	0.181	0.219	**0.252****	**0.284****
	(0.092)	**(0.078)**	(0.143)	(0.166)	**(0.125)**	**(0.132)**
High-ability: feedback and reputational reward	**0.322****	**0.393*****	**0.327****	0.129	0.127	0.071
	(0.149)	**(0.107)**	**(0.148)**	(0.171)	(0.148)	(0.150)
Financial rewards	**0.271*****	**0.289***	0.208	0.094	0.050	0.089
	(0.079)	**(0.169)**	(0.156)	(0.172)	(0.130)	(0.128)
Reputational rewards	**0.253****	**0.280***	-0.028	-0.193	0.058	-0.036
	(0.123)	**(0.141)**	(0.162)	(0.172)	(0.121)	(0.166)
Initial value	**0.672*****	**0.703*****	**0.082*****	**0.229*****	**0.254*****	**0.219*****
	(0.029)	**(0.029)**	**(0.026)**	**(0.035)**	**(0.032)**	**(0.039)**
Stratification variables	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	1426	1425	1327	1323	1392	1365

Note: OLS. Mixed ability groups consist of students who performed below and above median in baseline testing. Well-performing (poor-performers) groups consist only of students who performed above (below) median. Only groups of three students are taken into account. Poor-performing groups serve as comparison group in this exercise. Robust standard errors adjusted for clustering at school level are in parentheses. Controlled for stratification variables (area, level and school performance in national examinations). * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix E1.2: Comparison of average treatment effects on performance, exerted effort and subjective well-being, by group gender composition

Dependent variable:	Math	English	Perceived Stress	Subjective Happiness	Effort Math	Effort English
Three boys: only feedback	0.065	-0.009	-0.056	**-0.288****	0.067	-0.143
	(0.071)	(0.065)	(0.165)	**(0.136)**	(0.148)	(0.139)
Two boys + One girl:	**0.184*****	-0.026	-0.104	**-0.160****	-0.121	-0.171
only feedback	**(0.055)**	(0.068)	(0.111)	**(0.079)**	(0.133)	(0.107)
One boy + Two girls:	**0.164*****	0.002	-0.022	**-0.204****	**-0.192***	**-0.258****
only feedback	**(0.044)**	(0.044)	(0.104)	**(0.099)**	**(0.106)**	**(0.108)**
Three boys: feedback and	0.263	0.264	0.295	-0.117	-0.096	-0.259
monetary reward	(0.301)	(0.181)	(0.366)	(0.234)	(0.187)	(0.216)
Two boys + 1 girl: feedback and	**0.465*****	**0.310****	0.142	**-0.303***	-0.019	-0.007
monetary reward	**(0.143)**	**(0.139)**	(0.236)	**(0.172)**	(0.148)	(0.128)
One boy + 2 girls: feedback and	**0.359****	**0.303*****	0.302	**-0.345****	0.077	-0.034
monetary reward	**(0.154)**	**(0.106)**	(0.197)	**(0.123)**	(0.139)	(0.139)
Three boys: feedback and	**0.504*****	0.197	0.046	-0.258	0.033	-0.256
reputational reward	**(0.149)**	(0.129)	(0.272)	(0.207)	(0.167)	(0.173)
Two boys + 1 girl: feedback and	**0.388*****	0.153	0.263	**-0.302***	0.071	-0.137
reputational reward	**(0.113)**	(0.107)	(0.205)	**(0.179)**	(0.139)	(0.134)
One boy + 2 girls: feedback and	0.206	**0.203****	0.174	-0.182	-0.046	-0.135
reputational reward	(0.131)	**(0.106)**	(0.197)	(0.172)	(0.156)	(0.146)
Financial rewards	**0.411*****	**0.309****	-0.029	-0.196	0.195	0.029
	(0.154)	**(0.137)**	(0.205)	(0.157)	(0.167)	(0.189)
Reputational rewards	**0.403*****	**0.382*****	-0.114	-0.182	-0.071	-0.167
	(0.125)	**(0.119)**	(0.215)	(0.178)	(0.138)	(0.146)
Initial value	**0.707*****	**0.734*****	**0.081*****	**0.233*****	**0.258*****	**0.224*****
	(0.029)	**(0.028)**	**(0.026)**	**(0.034)**	**(0.032)**	**(0.039)**
Stratification variables	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observation	1624	1623	1624	1313	1392	1365

Note: OLS. Three boys stands for groups consisting only of boys. Two boys + One girl stands (One boy + Two girls) for groups consisting of two (one) boys and one (two) girl. Only groups of three students are taken into account. Groups consisting of only girls serve as the comparison group in this exercise. Robust standard errors adjusted for clustering at school level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix F1: Other

Appendix F1.1: Aggregated average treatment effects, OLS versus other specifications, by subject

Specification:	OLS	IPW	Imputation (median ratio)	Imputation (class percentiles)	Median Regression
MATH					
Within-class feedback (T1)	**0.099***	0.076	**0.124***	**0.112***	**0.096****
	(0.059)	(0.065)	**(0.063)**	**(0.055)**	**(0.052)**
Across-class feedback (T2)	0.089	**0.107***	**0.116****	**0.096***	**0.069**
	(0.056)	**(0.066)**	**(0.054)**	**(0.053)**	**(0.051)**
Financial Rewards (Finrew)	**0.142***	**0.327*****	**0.198****	**0.169****	**0.175****
	(0.078)	**(0.100)**	**(0.081)**	**(0.079)**	**(0.083)**
Reputational Rewards (Reprew)	**0.115***	**0.152****	**0.164****	**0.157****	**0.124****
	(0.064)	**(0.073)**	**(0.073)**	**(0.067)**	**(0.062)**
Baseline Mathematic score	**0.725*****	**0.731*****	**0.757****	**0.668*****	**0.750*****
	(0.017)	**(0.021)**	**(0.049)**	**(0.019)**	**(0.019)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	NA
ENGLISH					
Within-class feedback (T1)	-0.028	-0.029	0.019	-0.011	-0.018
	(0.039)	(0.048)	(0.052)	(0.042)	(0.042)
Across-class feedback (T2)	0.012	0.028	0.056	0.025	0.017
	(0.040)	(0.048)	(0.051)	(0.043)	(0.042)
Financial Rewards (Finrew)	**0.158****	**0.290*****	**0.159***	**0.211*****	**0.176****
	(0.053)	**(0.083)**	**(0.075)**	**(0.066)**	**(0.075)**
Reputational Rewards (Reprew)	**0.108****	**0.155****	**0.103**	**0.158*****	**0.095**
	(0.053)	**(0.068)**	**(0.073)**	**(0.056)**	**(0.058)**
Baseline English score	**0.739*****	**0.696*****	**0.737*****	**0.691*****	**0.758*****
	(0.016)	**(0.024)**	**(0.026)**	**(0.016)**	**(0.016)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	NA

Note: Various specifications. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). IPW stands for the inverse probability weight regression adjusting for students' probability of dropping out. The imputations based on median ratio imputed the last available observation in Math or English adjusted for the difference in the test difficulties using median ratio. In the imputations based on class percentiles I first seek for the percentile rank of the student in his last observed score within his class and then assign the student grade from the testing round 5 of a student from the same percentile rank. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.2: Dis-aggregated average treatment effects on Math, OLS versus other specifications

Dependent variable	Math				
Specification:	OLS	IPW	Imputation (median correction)	Imputation (class percentiles)	Median Regression
NON-INTERACTED TREATMENT EFFECTS					
Within-class feedback, no feedback (T1_solo)	0.100 (0.085)	0.046 (0.092)	**0.133*** **(0.079)**	0.070 (0.082)	**0.121*** **(0.072)**
Across-class feedback, no feedback (T2_solo)	0.082 (0.073)	0.067 (0.079)	**0.129*** **(0.068)**	0.036 (0.627)	0.070 (0.069)
Financial Rewards, no feedback (Fin_solo)	0.106 (0.101)	0.151 (0.102)	**0.169*** **(0.096)**	0.070 (0.097)	**0.162*** **(0.093)**
Reputational Rewards, no feedback (Rep_solo)	0.138 (0.141)	0.188 (0.149)	**0.206*** **(0.124)**	0.092 (0.115)	0.159 (0.132)
TREATMENT INTERACTIONS					
Within-class feedback, monetary reward (T1_fin)	**0.231*** **(0.118)**	**0.338**** **(0.135)**	**0.281**** **(0.129)**	**0.202*** **(0.116)**	**0.267**** **(0.113)**
Across-class feedback, monetary reward (T2_fin)	**0.277**** **(0.139)**	**0.456***** **(0.132)**	**0.331**** **(0.128)**	0.209 (0.130)	**0.296*** **(0.160)**
Within-class feedback, reputational reward (T1_rep)	**0.209**** **(0.103)**	**0.212*** **(0.108)**	**0.266**** **(0.112)**	**0.171*** **(0.099)**	**0.214**** **(0.100)**
Across-class feedback, reputational reward (T2_rep)	**0.188**** **(0.080)**	**0.208**** **(0.087)**	**0.186**** **(0.073)**	**0.164**** **(0.076)**	**0.193***** **(0.075)**
Baseline Math score	**0.725***** **(0.017)**	**0.732***** **(0.021)**	**0.755***** **(0.048)**	**0.658***** **(0.019)**	**0.747***** **(0.019)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	NA

Note: Various specifications. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). IPW stands for the inverse probability weight regression adjusting for students' probability to dropout. The imputations based on median ratio imputed the last available observation in Math or English adjusted for the difference in the test difficulties using median ratio. In the imputations based on class percentiles I first seek for the percentile rank of the student in his last observed score within his class and then assign the student grade from the testing round 5 of a student from the same percentile rank. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.3: Dis-aggregated average treatment effects on Math, OLS versus other specifications

Dependent variable	English				
Specification:	OLS	IPW	Imputation (median correction)	Imputation (class percentiles)	Median Regression
NON-INTERACTED TREATMENT EFFECTS					
Within-class feedback, no feedback (T1_solo)	**-0.128**** **(0.056)**	**-0.133*** **(0.070)**	**-0.151***** **(0.043)**	**-0.207***** **(0.062)**	**-0.149***** **(0.053)**
Across-class feedback, no feedback (T2_solo)	-0.049 (0.059)	-0.079 (0.072)	-0.062 (0.046)	**-0.139**** **(0.065)**	-0.074 (0.055)
Financial Rewards, no feedback (Fin_solo)	0.045 (0.088)	0.032 (0.085)	0.036 (0.069)	-0.047 (0.093)	0.009 (0.084)
Reputational Rewards, no feedback (Rep_solo)	0.016 (0.082)	0.004 (0.084)	0.026 (0.059)	-0.099 (0.086)	-0.025 (0.078)
TREATMENT INTERACTIONS					
Within-class feedback, monetary reward (T1_fin)	0.103 (0.094)	**0.145*** **(0.086)**	0.065 (0.079)	0.043 (0.101)	0.129 (0.107)
Across-class feedback, monetary reward (T2_fin)	**0.173*** **(0.094)**	**0.248**** **(0.102)**	**0.128*** **(0.074)**	0.062 (0.104)	**0.175**** **(0.084)**
Within-class feedback, reputational reward (T1_rep)	0.087 (0.080)	0.041 (0.078)	0.062 (0.057)	-0.009 (0.081)	0.096 (0.077)
Across-class feedback, reputational reward (T2_rep)	0.047 (0.080)	0.071 (0.077)	0.052 (0.065)	-0.042 (0.087)	-0.039 (0.079)
Baseline Math score	**0.737***** **(0.016)**	**0.697***** **(0.023)**	**0.702***** **(0.014)**	**0.682***** **(0.017)**	**0.759***** **(0.017)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	NA

Note: Various specifications. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). IPW stands for the inverse probability weight regression adjusting for students' probability to dropout. The imputations based on median ratio imputed the last available observation in Math or English adjusted for the difference in the test difficulties using median ratio. In the imputations based on class percentiles I first seek for the percentile rank of the student in his last observed score within his class and then assign the student grade from the testing round 5 of a student from the same percentile rank. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.4: OLS versus quantile regressions, by subject

	OLS	Quantile Regression (q=0.25)	Quantile Regression (q=0.5)	Quantile Regression (q=0.75)
MATH				
Within-class feedback (T1)	**0.099***	0.069	**0.096****	0.101
	(0.059)	(0.047)	**(0.052)**	(0064)
Across-class feedback (T2)	0.089	0.069	0.069	0.061
	(0.056)	(0.048)	(0.051)	(0.063)
Financial Rewards (Finrew)	**0.142***	**0.145***	**0.175****	0.127
	(0.078)	**(0.080)**	**(0.083)**	(0.092)
Reputational Rewards (Reprew)	**0.115***	0.078	**0.124****	**0.125***
	(0.064)	(0.063)	**(0.062)**	**(0.075)**
Baseline Mathematic score	**0.725*****	**0.702*****	**0.750*****	**0.770*****
	(0.017)	**(0.024)**	**(0.019)**	**(0.027)**
Controlled for stratas	Yes	Yes	Yes	Yes
ENGLISH				
Within-class feedback (T1)	-0.028	-0.061	-0.018	-0.021
	(0.039)	(0.039)	(0.042)	(0.049)
Across-class feedback (T2)	0.012	0.014	0.017	-0.021
	(0.040)	(0.044)	(0.042)	(0.044)
Financial Rewards (Finrew)	**0.158****	**0.181****	**0.176****	**0.160****
	(0.053)	**(0.071)**	**(0.075)**	**(0.065)**
Reputational Rewards (Reprew)	**0.108****	**0.099***	**0.095**	**0.105***
	(0.053)	**(0.058)**	**(0.058)**	**(0.058)**
Baseline English score	**0.739*****	**0.746*****	**0.758*****	**0.764*****
	(0.016)	**(0.018)**	**(0.016)**	**(0.020)**
Controlled for stratas	Yes	Yes	Yes	Yes

Note: OLS and quartile regressions. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.5a: OLS versus quantile regressions, disaggregated average treatment effects on Math

Dependent variable	**Math**			
Specification:	OLS	Quantile Regression (q=0.25)	Quantile Regression (q=0.5)	Quantile Regression (q=0.75)
NON-INTERACTED TREATMENT EFFECTS				
Within-class feedback, no feedback (T1_solo)	0.100 (0.085)	0.097 (0.059)	**0.121*** **(0.072)**	0.087 (0.092)
Across-class feedback, no feedback (T2_solo)	0.082 (0.073)	0.084 (0.059)	0.070 (0.069)	0.050 (0.093)
Financial Rewards, no feedback (Fin_solo)	0.106 (0.101)	0.131 (0.089)	**0.162*** **(0.093)**	0.012 (0.123)
Reputational Rewards, no feedback (Rep_solo)	0.138 (0.141)	**0.181*** **(0.102)**	0.159 (0.132)	0.137 (0.156)
TREATMENT INTERACTIONS				
Within-class feedback, monetary reward (T1_fin)	**0.231*** **(0.118)**	**0.240**** **(0.123)**	**0.267**** **(0.113)**	0.239 (0.147)
Across-class feedback, monetary reward (T2_fin)	**0.277**** **(0.139)**	**0.253*** **(0.137)**	**0.296*** **(0.160)**	0.257 (0.192)
Within-class feedback, reputational reward (T1_rep)	**0.209**** **(0.103)**	0.148 (0.092)	**0.214**** **(0.100)**	**0.237*** **(0.128)**
Across-class feedback, reputational reward (T2_rep)	**0.188**** **(0.080)**	**0.163*** **(0.096)**	**0.193***** **(0.075)**	0.141 (0.099)
Baseline English score	**0.725***** **(0.017)**	**0.700***** **(0.024)**	**0.747***** **(0.019)**	**0.764***** **(0.028)**
Controlled for stratas	Yes	Yes	Yes	Yes

Note: OLS and quartile regressions. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects - four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.5b: OLS versus quantile regressions, disaggregated average treatment effects on English

Dependent variable	English			
Specification:	OLS	Quantile Regression (q=0.25)	Quantile Regression (q=0.5)	Quantile Regression (q=0.75)
NON-INTERACTED TREATMENT EFFECTS				
Within-class feedback, no feedback (T1_solo)	**-0.128**** **(0.056)**	**-0.137***** **(0.052)**	**-0.149***** **(0.053)**	**-0.127**** **(0.064)**
Across-class feedback, no feedback (T2_solo)	-0.049 (0.059)	-0.025 (0.055)	-0.074 (0.055)	-0.074 (0.064)
Financial Rewards, no feedback (Fin_solo)	0.045 (0.088)	0.078 (0.100)	0.009 (0.084)	-0.029 (0.099)
Reputational Rewards, no feedback (Rep_solo)	0.016 (0.082)	0.009 (0.084)	-0.025 (0.078)	-0.002 (0.098)
TREATMENT INTERACTIONS				
Within-class feedback, monetary reward (T1_fin)	0.103 (0.094)	0.049 (0.091)	0.129 (0.107)	**0.169*** **(0.089)**
Across-class feedback, monetary reward (T2_fin)	**0.173*** **(0.094)**	**0.209**** **(0.097)**	**0.175**** **(0.084)**	0.099 (0.078)
Within-class feedback, reputational reward (T1_rep)	0.087 (0.080)	0.093 (0.084)	0.096 (0.077)	0.060 (0.087)
Across-class feedback, reputational reward (T2_rep)	0.047 (0.080)	-0.024 (0.089)	-0.039 (0.079)	0.040 (0.077)
Baseline English score	**0.737***** **(0.016)**	**0.749***** **(0.017)**	**0.759***** **(0.017)**	**0.759***** **(0.018)**
Controlled for stratas	Yes	Yes	Yes	Yes

Note: OLS and quartile regressions. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects - four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.6: Differences in reported time for preparation in Math and English

	Preparation time in Math	Preparation time in Math	Preparation time in English	Preparation time in English
AGGREGATED ATE				
Within-class feedback	-0.012	0.004	0.059	0.063
	(0.062)	(0.062)	(0.068)	(0.070)
Across-class feedback	**-0.157****	-0.108	-0.066	-0.059
	(0.075)	(0.076)	(0.076)	(0.077)
Financial rewards	**-0.188****	**-0.171****	-0.101	-0.099
	(0.078)	**(0.079)**	(0.096)	(0.101)
Reputational rewards	**-0.189****	**-0.155***	**-0.159***	**-0.163***
	(0.082)	**(0.090)**	**(0.089)**	**(0.093)**
DISAGGREGATED ATE				
NON-INTERACTED TREATMENT EFFECTS				
Within-class feedback, no feedback (T1_solo)	-0.028	-0.040	-0.002	-0.022
	(0.092)	(0.091)	(0.108)	(0.109)
Across-class feedback, no feedback (T2_solo)	-0.057	-0.025	-0.022	-0.047
	(0.075)	(0.078)	(0.104)	(0.108)
Financial Rewards, no feedback (Fin_solo)	**-0.241****	**-0.230****	-0.158	-0.179
	(0.106)	**(0.105)**	(0.115)	(0.125)
Reputational Rewards, no feedback (Rep_solo)	-0.022	-0.023	-0.114	-0.163
	(0.116)	(0.126)	(0.121)	(0.127)
TREATMENT INTERACTIONS				
Within-class feedback, monetary reward (T1_fin)	-0.045	-0.013	0.089	0.084
	(0.100)	(0.103)	(0.114)	(0.125)
Across-class feedback, monetary reward (T2_fin)	**-0.372*****	**-0.351*****	**-0.277*****	**-0.291*****
	(0.096)	**(0.106)**	**(0.105)**	**(0.112)**
Within-class feedback, reputational reward (T1_rep)	**-0.178****	-0.135	-0.103	-0.126
	(0.086)	(0.096)	(0.092)	(0.095)
Across-class feedback, reputational reward (T2_rep)	**-0.428*****	**-0.336***	**-0.272***	**-0.255***
	(0.164)	**(0.178)**	**(0.141)**	**(0.142)**
Initial Math/English score		**-0.009*****		**-0.008*****
		(0.003)		**(0.003)**
Controlled for stratas	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000
Number of observations	5782	4822	5659	4714

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (four areas by distance from the capital city, Kampala, school performance at national examination and grade level (P6,P7, S1 up to S4). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix F1.7a: Comparison of uncorrected and corrected p-values using different multiple-comparison procedures: for aggregated and disaggregated average treatment effects on students' overall performance

P-VALUE CORRECTIONS	Uncorrected	Bonferroni	Sidak	Holm	Holland	Hochberg	Simes	Yekutieli
AGGREGATED ATE								
Within-class feedback	0.102	1.000	0.754	0.409	0.350	0.389	0.133	0.422
Across-class feedback	0.130	1.000	0.836	0.409	0.350	0.389	0.153	0.487
DIS-AGGREGATED ATE								
Within-class feedback, no rewards	0.774	1.000	1.000	1.000	0.961	0.776	0.776	1.000
Across-class feedback, no rewards	0.464	1.000	1.000	1.000	0.956	0.776	0.587	1.000
Financial Rewards, no feedback	0.059	1.000	0.685	0.413	0.347	0.413	0.086	0.306
Reputational Rewards, no feedback	0.555	1.000	1.000	1.000	0.961	0.776	0.659	1.000
Within-class feedback with financial rewards	0.035	0.667	0.493	0.316	0.275	0.293	0.058	0.205
Across-class feedback with financial rewards	0.014	0.257	0.228	0.135	0.127	0.135	0.026	0.091
Within-class feedback with reputation rewards	0.012	0.221	0.199	0.128	0.121	0.128	0.025	0.087
Across-class feedback with reputation rewards	0.097	1.000	0.856	0.581	0.457	0.581	0.131	0.047

Appendix F1.7b: Comparison of uncorrected and corrected p-values using different multiple-comparison procedures: for aggregated and disaggregated average treatment effects on students' stress

P-VALUE CORRECTIONS	Uncorrected	Bonferroni	Sidak	Holm	Holland	Hochberg	Simes	Yekutieli
AGGREGATED ATE								
Within-class feedback	0.827	1.000	1.000	1.000	0.970	0.838	0.838	1.000
Across-class feedback	0.199	1.000	0.944	1.000	0.703	0.838	0.287	0.913
DIS-AGGREGATED ATE								
Within-class feedback, no rewards	0.600	1.000	1.000	1.000	0.993	0.938	0.712	1.000
Across-class feedback, no rewards	0.582	1.000	1.000	1.000	0.993	0.938	0.712	1.000
Financial Rewards, no feedback	0.005	0.089	0.085	0.075	0.072	0.075	0.022	0.079
Reputational Rewards, no feedback	0.696	1.000	1.000	1.000	0.993	0.938	0.735	1.000
Within-class feedback with financial rewards	0.088	1.000	0.826	0.792	0.563	0.792	0.152	0.539
Across-class feedback with financial rewards	0.673	1.000	1.000	1.000	0.993	0.938	0.735	1.000
Within-class feedback with reputation rewards	0.226	1.000	0.992	1.000	0.833	0.938	0.330	1.000
Across-class feedback with reputation rewards	0.111	1.000	0.893	0.888	0.610	0.888	0.176	0.623

Appendix F1.7c: Comparison of uncorrected and corrected p-values using different multiple-comparison procedures: for aggregated and disaggregated average treatment effects on students' subjective happiness

P-VALUE CORRECTIONS	Uncorrected	Bonferroni	Sidak	Holm	Holland	Hochberg	Simes	Yekutieli
AGGREGATED ATE								
Within-class feedback	0.036	0.468	0.379	0.304	0.266	0.288	0.078	0.248
Across-class feedback	0.341	1.000	0.996	1.000	0.876	0.984	0.493	1.000
DIS-AGGREGATED ATE								
Within-class feedback, no rewards	0.169	1.000	0.970	1.000	0.773	0.979	0.268	0.949
Across-class feedback, no rewards	0.316	1.000	0.999	1.000	0.874	0.979	0.399	1.000
Financial Rewards, no feedback	0.115	1.000	0.902	1.000	0.667	0.979	0.199	0.705
Reputational Rewards, no feedback	0.281	1.000	0.998	1.000	0.874	0.979	0.381	1.000
Within-class feedback with financial rewards	0.005	0.089	0.086	0.071	0.069	0.071	0.018	0.064
Across-class feedback with financial rewards	0.517	1.000	0.999	1.000	0.946	0.979	0.614	1.000
Within-class feedback with reputation rewards	0.059	1.000	0.684	0.708	0.518	0.708	0.140	0.497
Across-class feedback with reputation rewards	0.054	1.000	0.654	0.706	0.516	0.706	0.140	0.497

2 Information Provision and Overconfidence: Evidence from a Randomized Control Trial in Schools

"No problem in judgment and decision making is more prevalent and more potentially catastrophic than overconfidence" (Plous, 1993, p. 217)

2.1 Introduction

People often believe they are better than average. Overconfidence[34] is an important behavioral bias observed in different domains[35] and may have economic consequences[36]. An important question, therefore, is how to lower student inaccuracy in self-assessment. Provision of feedback regarding one's performance lowers the confidence gap (e.g., Ryvkin et al., 2012) but people remain overconfident (e.g., Lipko et al., 2009). A possible explanation is that people have insufficient information about others.

Unlike other studies largely using university students in developed countries as their samples, I bring evidence from a large field experiment conducted with more than 5,000 primary and secondary school students repeatedly tested and interviewed in a

[34] What I mean by overconfidence is, according to the definition by Moore and Healy (2008), overestimation of one's own ability or performance.
[35] Ranging from university professors (Cross, 1977), students (Clayson, 2005), engineers (Zenger, 1992), to drivers (Marottoli and Richardson, 1998), etc.
[36] Such as in Levy (1983), Howard (1983), Odean (1998), Camerer and Lovallo (1999), Kennedy et al. (2002), Moore and Kim, (2003), Clayson (2005), etc. For details see the literature review section.

developing country, Uganda. Students were randomly assigned to a control group and two treatment groups. Students in the within-class feedback group were randomly divided into groups of three to four classmates and received feedback about their own performance and the performance of their group within their respective classes. Students in the across-class feedback group were evaluated as a class and received feedback about their own performance and the performance of their class compared to other classes in the district. Students were tested and interviewed five times over one academic year and received feedback repeatedly.

The first contribution of this research is that I test whether the provision of feedback about their own performance, the performance of group members and the relative position of the group leads to a decrease in excessive overconfidence.

Second, a randomized experiment with five repeated measures of student performance and their performance expectations represents a unique opportunity to bring new evidence to the debate regarding the existence of the Dunning-Kruger effect, the existence of which has been questioned. Kruger and Dunning (1999) first documented a systematic pattern in overconfidence: unskilled students tend to strongly overestimate their performance and the skilled students to weakly underestimate their performance. Krueger and Mueller (2002) attribute the pattern to the regression-to-the-mean combined with better-than-average effect. The main argument is that a fraction of students appear in the bottom/top performance quartile by chance and their performance will regress towards the mean in subsequent testing. Random allocation of students into comparison and treatment groups (a method used in this experiment) results in both groups being equally influenced by regression-to-the-mean (Barnett, van der Pols and Dobson, 2005). In

the current experiment repeated measures of student performance help to distinguish between students who remain in the bottom performance quartile during the entire duration of the experiment ("non-switchers") and students who improve and as a result depart from the original bottom performance quartile ("switchers").

The results confirm that students are overconfident and the overconfidence persists over time. The pattern is similar in Math and English separately. One may expect that being exposed to a task repeatedly should help students adjust their expectations. The evidence from this experiment, however, suggests the opposite. Based on the expected performance of the control group, in the final testing round students miscalibration increased by 69.1% (86.6% in Math and 52.9% in English). The provision of feedback lowers the degree of miscalibration but the confidence gap remains persistent.

Students react to feedback provision immediately, with the student confidence gap decreasing by 12.5% right after they received feedback compared to the students in the control group. Students significantly improved in their estimations in response to the first two subsequent feedbacks. Afterwards their confidence level did not improve significantly. The type of feedback does not play a significant role. Girls improve in their estimations more compared to boys. Students in secondary schools improve more compared to primary school students, which is most likely caused by greater selectivity into secondary schools (age does not play a significant role in calibration). The miscalibration of results in English is lower than the prediction of inaccuracy in Math. A possible explanation is that control-group students did not improve in Math during the entire academic year whereas they improved in English by more than 50%. This finding suggests that as students' skills in English become stronger they recognize the extent of their mistakes and therefore become

more precise in their evaluations, which is in line with the results of Kruger and Dunning (1999).

The results confirm the presence of the Dunning-Kruger effect on the tasks that students perceive as more difficult than tasks to which they are usually exposed. On easier tasks or tasks of comparable difficulty, all students overestimated their performance, with bottom performers being even more inaccurate compared to the top performers. Regression-to-the-mean emphasized mainly by Krueger and Mueller (2002) does not seem to play a significant role. Other predictions by Kruger and Dunning (1999) are discussed in the Results Section.

The chapter is organized as follows. In section 2.2 I first review the existing literature. In section 2.3 I describe the final sample. In section 2.4, I discuss overconfidence among students and I present the evidence for and against the predictions of Kruger and Dunning (1999). In section 2.5, I discuss the average treatment effects of different incentives on student overconfidence. In section 2.6 I discuss the heterogeneity of the results. Finally, in section 2.7 I conclude.

2.2 Literature Review

People have been found overconfident in many domains. Cross (1977) reports that 94% of college professors considered their performances to be above average and more than two-thirds thought they performed in the top quartile. Bottom quartile students (in terms of performance) evaluated their score to be in the 60th percentile (Kruger and Dunning, 1999). 42% of engineers estimated their performance to be within the top 5%

among all colleagues (Zenger, 1992). All active drivers older than 77 years who participated in a study by Marottoli and Richardson (1998) ranked themselves as average or above average in their driving skills compared to other drivers of the same age. Similarly, 90% of drivers in Svenson's (1981) study believed they possessed above average driving ability; etc.

2.2.1 The Dunning-Kruger Effect

Subjects who are particularly overconfident typically come from the bottom performance quartile, whereas subjects from the top performance quartile have a tendency to underestimate their performance. Kruger and Dunning (1999) were the first to describe asymmetries in evaluations of self-assessment known as the Dunning-Kruger effect or the "unskilled-and-unaware" phenomenon. Since then the topic has been vividly discussed in the literature[37]. The authors argue that incompetent students face a "dual burden: not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realize it" (p.1121). Skilled subjects, on the other hand have a tendency to underestimate their performance. The authors suggest that the skilled face the "false-consensus effect:" since they perform well, they expect others to perform in a similar way. The skilled fail to identify their strengths. The study was replicated several times in different domains showing similar results (e.g., Dunning et al., 2003; Ehrlinger et al., 2008, etc.).

[37] The unskilled-and-unaware phenomenon has been confirmed to arise among medical laboratory technicians (Haun et al., 2000), medical students (Hodges et al., 2001), undergraduate students (Dunning et al., 2003), students applying for graduate studies (Ryvkin et al., 2012), clerks (Edwards et al., 2003), etc.

A different interpretation of the empirical findings was offered by Krueger and Mueller (2002)[38] who attributed the pattern, instead, to the existence of "regression-to-the-mean" in combination with the "better-than-average" phenomenon (people generally overestimate their performance). Regression-to-the-mean means that it is unlikely that all students who are in the lowest (highest) performance group will once again be in the lowest (highest) performance group in subsequent testing, because some fraction of them happened to perform poorly (well) by pure "bad luck". As a result, their performance and predictions of their own performance will be inflated or deflated toward the mean.

Regression-to-the-mean arises in situations where the variables are imperfectly correlated. If it is not taken into consideration, it may be attributed incorrectly to the overall treatment effect of an intervention. Burson et al. (2006) agree with Krueger and Mueller (2002) that all students face difficulty with accurate self-assessment but they attribute the existence of the pattern to task difficulty. They manipulate the levels of difficulties of the tasks in their sessions and find that poor performers had a tendency to overestimate their performance only in tasks that they perceived as easy, whereas in difficult tasks this pattern disappeared. Moore and Cain (2008) go one step further and examine the relationship between task difficulty and the nature of the task. The authors find evidence that people are inaccurate about their performance on difficult tasks that are skill-based (as opposed to tasks whose success is based on chance). They attribute people's inflated beliefs to the fact that, on skill-based tasks, people simply have better information

[38] See also Ackerman et al. (2002), Kruger and Funder (2004), Burson et al. (2006), etc.

about their own preferences than about others' preferences which leads to regressive beliefs about others.

Ehrlinger et al. (2008) attempt to answer whether the Dunning-Kruger effect is a psychological phenomenon or a statistical artifact. In five experiments the authors ask students to predict their absolute and relative positions in various environments: with and without competition; in the lab as well as in the field; in an academic environment with undergraduate students and participants recruited at a Trap and Skeet (shooting) competition, or with and without financial incentives for accuracy of subjects' predictions. The authors use *counterfactual regression analysis* (Winship and Morgan, 1999) to discuss the presence of regression-to-the-mean. Overall, they do not find supportive evidence for the arguments of Krueger and Mueller (2002) and Burson et al. (2006). Their results instead support the original paper by Kruger and Dunning (1999), though not to the full extent.

Research presented here aims to contribute to the debate about the existence of the unskilled-and-unaware phenomenon but it differs in various domains. It contributes to the literature by bringing evidence from a randomized control trial with five repeated measures of student performance. It means that, as opposed to a statistically created counterfactual by Ehrlinger et al. (2008), part of my sample was randomly put into a control group. The performance of the control group mimics "what if" scenarios. Random allocation of students results in all groups being equally influenced by regression-to-the-mean (Barnett, van der Pols and Dobson, 2005). Repeated measures of student performance allow me to compare the behavior of students who remained in the bottom performance quartile during the entire duration of the experiment with students who

improved and as a result, departed from the original bottom performance quartile. Furthermore, it was implemented with primary and secondary school students in a developing country as opposed to the common use of undergraduate students in developed countries.

2.2.2 How to Improve Self-Assessments?

Miscalibrations in assessment of one's own performance may be costly. It has been shown that overconfidence may have economic consequences. It has played a role in stock market bubbles (Odean, 1998), entrepreneur failures (Camerer and Lovallo, 1999), suboptimal predictions in marketing management (Mahajan, 1992) and even in wars and strikes (Levy, 1983; Howard, 1983). Students are no exception (Kennedy et al., 2002; Clayson, 2005, etc.)[39]. Inflated beliefs regarding their own performance may lead students to exert suboptimal levels of effort and result in worse performance. Overconfident subjects seem to engage at a higher rate in competition (Camerer and Lovallo, 1999; Moore and Kim, 2003) and are less likely to take advice from others (Gino and Moore, 2007). Dunlosky and Rawson (2012) show that overconfidence may lead to underperformance in students. Therefore, it is important to understand who the students with the most inaccurate self-assessments are and how we can improve the level of their own assessment calibration.

[39] An exception is research conducted by Azmat et al. (2015). The authors studied the effects of rank performance feedback on university student academic performance and they found most of the students were underconfident.

If people persistently overestimate their performance, what could be done to help them improve their accuracy? Moore and Cain (2008) suggest that "myopic interpersonal comparisons" may be most effectively eliminated if the subjects receive sufficient information about the performance of others (p.207)[40]. Burson et al. (2006) also see the lack of understanding of how well their peers do in a given task as an explanation for students' inaccurate predictions. The provision of feedback helps to improve judgements and decision making (Engelmann and Strobel, 2000; Duffy and Hopkins, 2005; Krajč, 2008; Ryvkin et al., 2012; Moore and Cain, 2007; Miller and Geraci, 2011; Hacker et al., 2008)[41].

Although subjects who receive absolute and/or relative feedback about their own performance and/or performance of others typically improve their self-assessments, they remain overconfident. Camerer and Lovallo (1999) offer a possible explanation - people inflate their beliefs about their own performance because they neglect information about others and focus on their own performance only. The authors studied the connection between overconfidence and "entry into competitive games and markets". They were particularly interested in testing the hypothesis that manager overconfidence could be responsible for business failures. They specified a new competition-specific mechanism – "reference group neglect[42] [which] predicts that when agents compete based on skill, they

---

[40] It is typical that people identify themselves with groups and derive utility from the image of the group (Crocker and Luhtanen, 1990, Aberson et al., 2000; Hewstone et al., 2002).

[41] Krajc (2008) and Ryvkin et al. (2012) studied the effect of feedback on calibration among applicants for graduate studies. Students received feedback about their own performance and were monetarily incentivized for the precision of the estimates of their performance. Feedback lowered miscalibration of self-assessment. Hacker et al. (2000) provided undergraduate students with repeated feedback regarding their own performance in three exams during a semester-length course. Only well performing students calibrated their self-assessments and became more accurate. Poorly performing students did not improve. Further studies with relative feedback see Ehrlinger et al., 2008; Moore and Cain, 2008.

[42] Known in the literature also as egocentrism – e.g., Hoelzl and Rustichini (2005), Moore and Kim (2003), Windschitl et al. (2003).

will be insufficiently sensitive to the quality of competition" (p.315-316). Accordingly, people may not incorporate sufficiently the qualities of their opponents and therefore inaccurately evaluate their own performance relative to others. In existing studies, the ranking of an individual typically depends solely on his/her own performance. In particular, it does not depend on the performance of others. Would group-dependence in performance eliminate reference group neglect and help to decrease students' inaccuracy of their self-assessments? In the current study, students are evaluated in groups and the ranking is therefore based on the performance of all group members[43].

In summary, the current paper contributes to the literature by identifying whether being involved, evaluated and incentivized in a group environment with detailed feedback about one's own performance and the performance of group members can influence one's own confidence and in what ways. Moreover, it brings actual field evidence from primary and secondary schools in Southern Uganda to the debate on the unskilled-and-unaware phenomenon.

2.3 Data and the Final Sample

The final sample consists of 53 schools located in rural areas (31 primary and 22 secondary schools) out of which 19 are public, 24 are private and 10 are community

[43] Note that the design of this experiment differs from studies interested in the prevalence of inaccurate beliefs with respect to social groups (i.e., confidence in groups). Healy and Offenberg (2007) and Healy and Pate (2011) find in lab and field experiments respectively that "confidence in one's group parallels confidence in oneself" (p.24). After decomposition of the results by gender, they find significant gender differences in attitudes towards groups. Females show greater confidence in their groups, whereas males in themselves. Cacault and Grieder (2016) find inflated confidence in the relative ability of a group a subject identifies with compared to the confidence in another group. In this study, I analyze the confidence in oneself while being evaluated in groups with repeated feedback about group performance.

schools. The sample comprised 146 classes from six grades (P6 and P7 in primary schools, S1 through S4 in secondary schools) containing a total of more than 5000 students who were repeatedly tested and out of whom 3513 students were present in all five testing rounds.

The main outcome variables are student performance in Math and in English and their confidence gap. The confidence gap measures a student's miscalibration of his/her own performance[44] and is calculated as the difference between the expected number of points and the real performance. Students could achieve a maximum of 50 points in Math and 50 points in English, and they were reminded of this at every visit. The exams were constructed to be consistent with school curriculums and were conducted in addition to regular examinations. All the exam questions were taken from Primary/Secondary Leaving Examination questions published by the Ugandan National Examination Board.

To predict their performance, students were asked the question, "How many points do you think you will truly obtain from the exam you have just completed/you are about to complete?" The question was asked separately for Math and for English. Students could circle only one out of 10 options (interval 1: "0 – 5 points", interval 2: "6-10 points", …, interval 10: "46-50 points")[45]. A student is considered to be over-estimating his/her performance if his/her real score is below the minimum value of the point-interval (positive confidence gap); accurately calculating his/her performance if his/her real score is within the interval range he/she estimates (confidence gap around 0); or under-

[44] Based on the classification by Moore and Healy (2008) miscalibration of one's own performance is called overestimation.

[45] I take the mid-point of each category in order to transmit the categories into points (i.e., 1: 3 points, 2: 8 points, …, 10: 48 points). An alternative way is to use the average value within each interval. The results suggest similar findings.

estimating his/her performance if his/her real score is above the maximum value of the point interval (negative confidence gap).

$$individual \begin{cases} over-estimating & if\ score_i < minI_j \\ well-calibrating & if\ minI_j \leq score_i \leq maxI_j \\ under-estimating & if\ score_i > maxI_j \end{cases}$$

where I_j is student i's choice of interval range, j=1, 2, …, 10.

Overall, 1% of students did not answer the question regarding the estimation of their performance or they answered it incorrectly (they selected more than one option) and I excluded the observations from further analysis. I also excluded data from one secondary school that went bankrupt during the 2012 academic year and whose students were reallocated to different secondary schools.

2.4 Results: Overconfidence and the Existence of the Unskilled-and-Unaware Phenomenon

First, I relate the findings from this experiment to the existing results on overconfidence and the existence of the Dunning-Kruger effect. The evaluation of the treatment effects of different incentives on student overconfidence is discussed in section 2.5.

2.4.1 Overconfidence

Most of the students (81%) of the sample were overconfident[46], whereas approximately 11% of students estimated their performance correctly and 8% of students underestimated their performance. In the baseline survey, the average student thought he/she ranked in the 94th percentile while his/her actual overall score from combined Math and English was in the 57th percentile[47]. In other words, the average student overestimated his/her performance by 104.7%. A similar pattern can be found for each subject separately, by level of study (primary versus secondary schools) and by gender[48] (see Appendix B2.2 for the relationship between initial performance and initial confidence gap).

One possible explanation is that students might face difficulty in estimating their performance in an activity that they have not performed before. Therefore, I looked at the confidence rates among the control-group students in the second testing round. I found that these students were even more overconfident, i.e., 95.1% overestimated their score, 2.9% correctly estimated their score, and 2.1% underestimated it[49]. In other words, the average control-group student overestimated his/her overall performance by 144%.

---

[46] Only 10 students scored higher than 40 points out of 50 in both, Math and English and only one student scored more than 45 points in Math (no one in English). Therefore, there is space for overconfidence for all students. See Appendices A2.1, A2.4 and A2.5 for comparison of students by the level of their confidence.

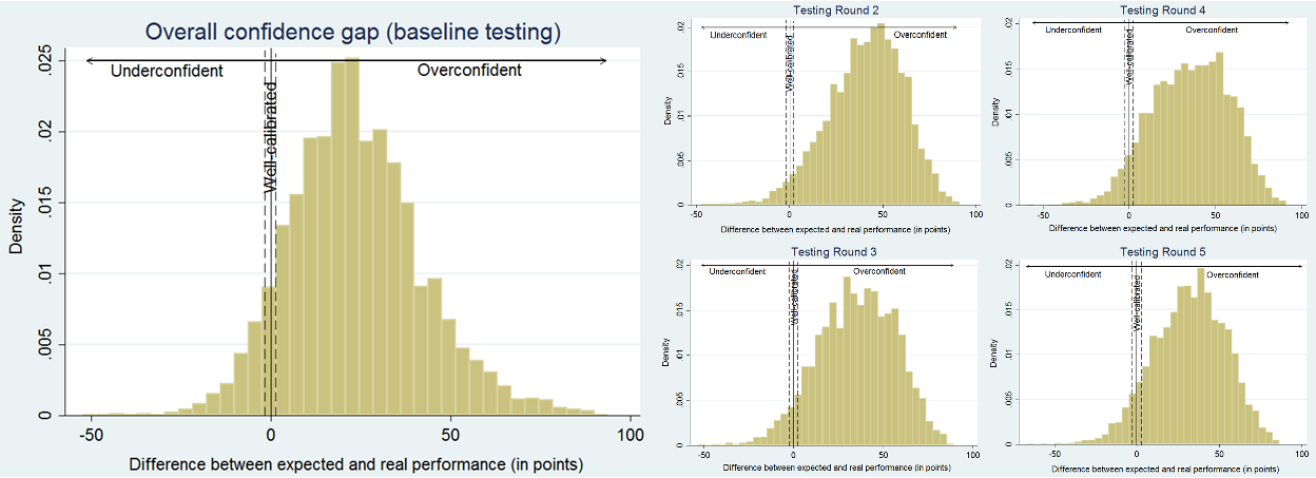
[47] Median student in terms of his/her performance expected to be ranked in 93rd percentile. For the analysis, I will follow common practice in the literature on overconfidence and provide comparisons for average students (in terms of their performance).

[48] In Math (English), the average student thought he/she scored in the 80th (the 84th percentile), while his/her actual score was in the 32nd (66th percentile). Students overestimated their performance on average by 99.1% in Math and by 57% in English. Similarly in the subsequent testing round the average student overestimated his/her performance in Math by 131% and 72% in English.

[49] A fraction of students received feedback in Round 2 before they revealed their expectations about their own performance. The results are similar for the overall sample in Round 2 as well as the restricted sample

Similarly, in the endline testing, 95.7% of control group students were overconfident and overestimated their performance by 131.5%. Thus, being exposed to a task repeatedly (i.e., repeated testing in the Math and English exam without feedback) does not seem to help students to close their confidence gaps (see Figures 2.1 and 2.2).

Figure 2.1: Distribution of the gap between expected and real performance, by testing rounds



Note: Confidence gap is the difference between expected and real performance. An underconfident student has a negative confidence gap (he/she expects a lower number of points compared to his/her real performance). A student with well-calibrated expectations estimates approximately the same score as he/she receives and his/her gap is close to zero. An overconfident student expects a higher score compared to his/her real performance (and therefore his/her confidence gap is positive).

It is not only that the fraction of overconfident control-group students increases over time but the average size of the confidence gap significantly increases by 54.2% in overall performance (79.4% in Math, and 16.4% in English, see also Figure 2.3)[50]. Being exposed to the provision of feedback decreases confidence bias, but overall students remain predominantly overconfident (see further discussion in section 2.5).

---

when only the control group (with no treatment) is taken into account. See Appendix A2.6 and A2.7 for more details.

[50] The results are based on the data of the control group students who mimic behavior in the absence of any of the treatments.

Figure 2.2: Fractions of students based on their confidence level, by testing round

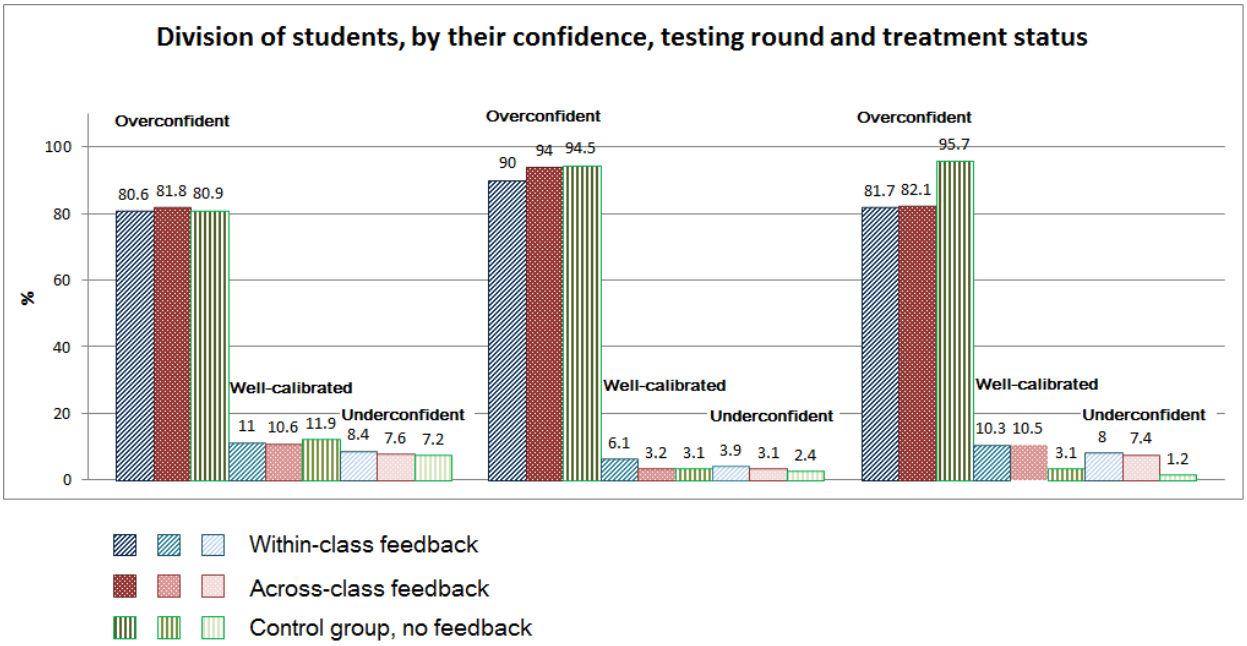
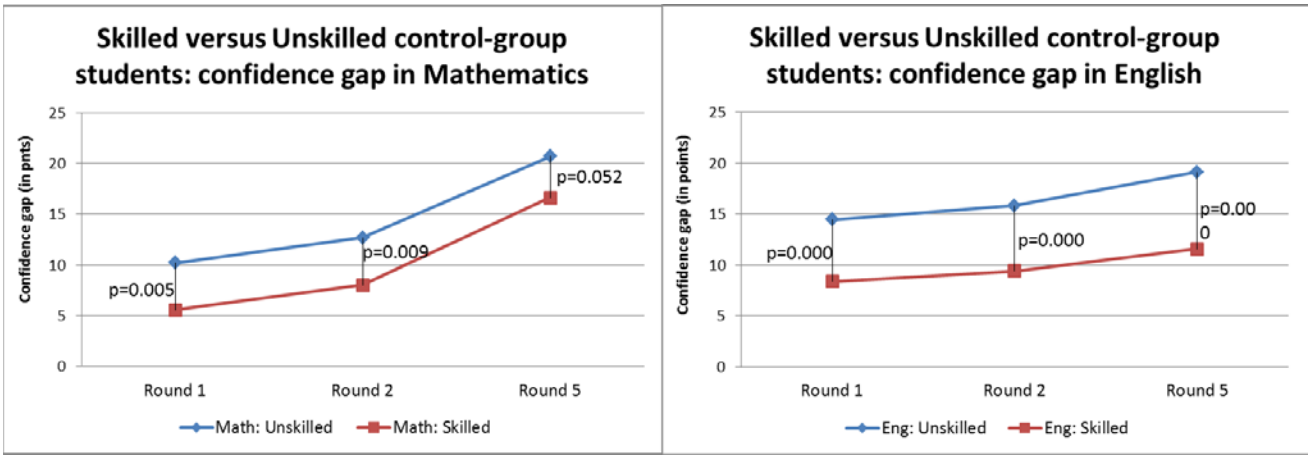


Figure 2.3: The change of the confidence gap in Math and English of bottom and top performing control-group students over time



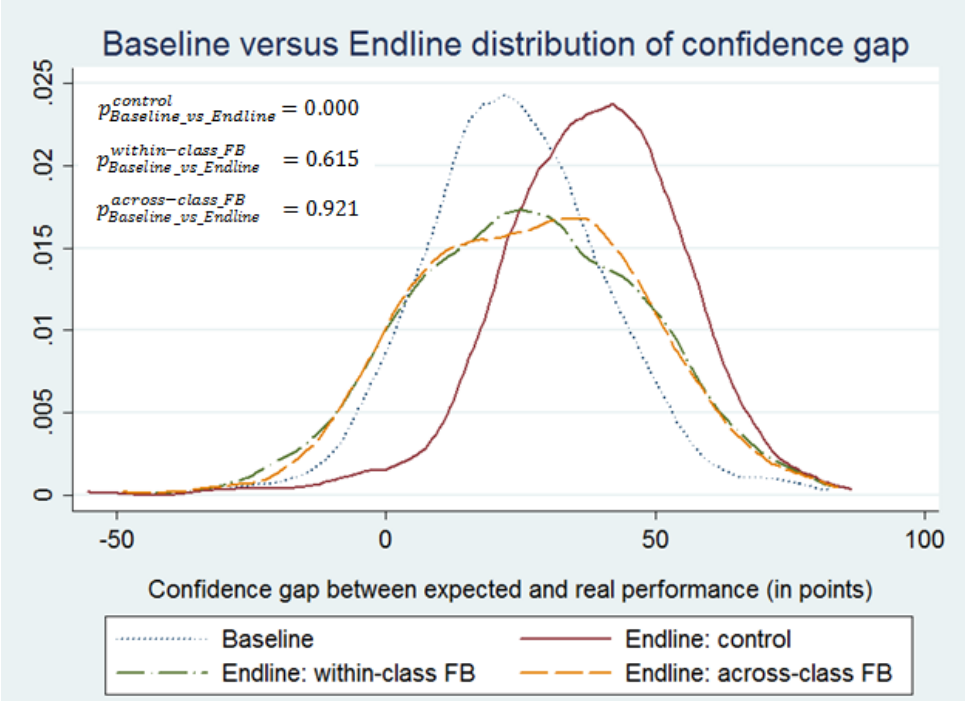
A similar pattern can be found for bottom- and top-performing students (see Figure 2.4), in which the p-values show statistical difference between the two groups of students. In the baseline survey, both skilled and unskilled students from the control group

miscalibrated their performance, although unskilled students to a significantly greater extent).

I used the Wilcoxon matched-pairs signed-ranks test (WSR) to test the significance in differences between the anticipated and actual performance of students. The null hypothesis of the test is that the observations come from the same distribution. The WSR test rejects the null hypothesis in all testing rounds (i.e., round 1 to round 5), in all cases at the 1% significance level. Students' estimates of their own scores were inaccurate (see Appendix A2.3).

Result 1: *Students of primary and secondary schools are mostly overconfident and the overconfidence persists over time, even after repeated feedback.*

Figure 2.4: Comparison of the confidence gap of control-group students in the baseline and endline testing



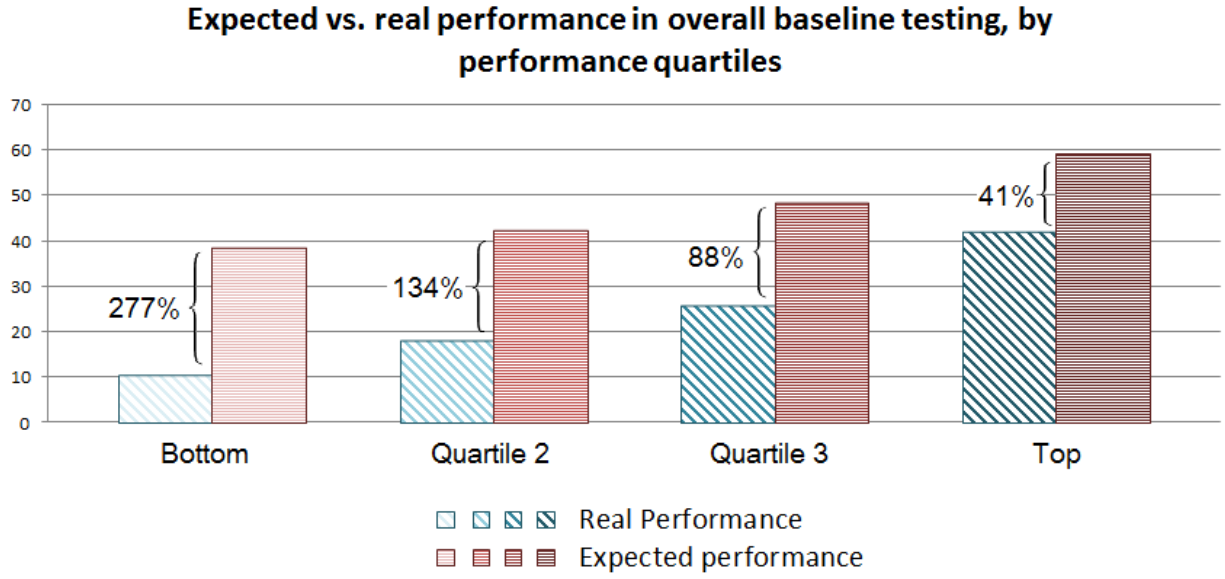
Note: FB stands for feedback.

2.4.2 The Unskilled-and-Unaware Phenomenon

2.4.2.1 Bottom versus Top Performing Students and their Expectations

Both top as well as bottom performing students overestimate their actual performance but the magnitude of miscalibration is significantly higher among bottom performers. The average gap between perceived and actual performance (in combined Math and English) is 10.9 points greater for the bottom performers compared to the top performers (see Figure 2.5). Bottom quartile students overestimated their actual performance by an extraordinary 277%, while top quartile students' overestimation was only 41%. In other words, the bottom (top) performers predicted themselves to be in the 87th (98th) percentile while they scored in the 11th (91st) percentile.

Figure 2.5: Difference in the confidence gap by performance quartiles



These results for bottom performing students are similar to the results of Kruger and Dunning (1999). In their case students scored in the 10th to 12th percentile (depending on the task) but predicted themselves to rank in the 60th percentile. The results for top performers, however, differ: skilled students in the current study do not underestimate their performance but are significantly more accurate in their predictions. Similar results can be found in Burson et al. (2006) and Ehlinger et al. (2008).

Result 2: *Bottom-performing students grossly overestimate their performance.*

Result 3: *Top-performing students do not underestimate their performance (as found by Kruger and Dunning, 1999) but overestimate significantly less compared to the unskilled ones.*

2.4.2.2 Behavioral Bias or Statistical Artefact?

The main argument of regression-to-the-mean is that a fraction of students appears in the bottom or top performance quartile by chance and their performance will regress towards the mean in subsequent testing. I take advantage of the design of this randomized experiment with repeated measures of student performance and their confidence levels to contribute to the debate about whether the Kruger-Dunning effect is a behavioral bias or a statistical artefact.

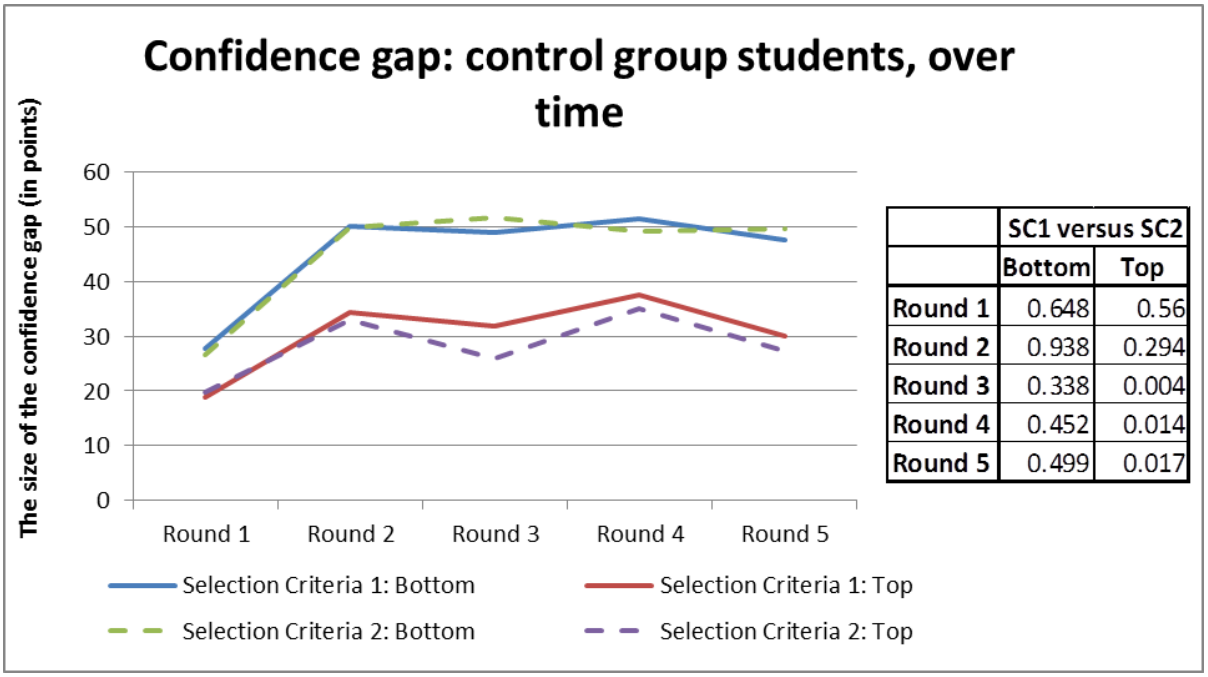
All groups (control or treatment groups) should be equally influenced by regression-to-the-mean (Barnett, van der Pols and Dobson, 2005) due to random assignment of students into groups. Repeated measures of student performance help distinguish between students who remained in the bottom performance quartile during the

entire duration of the experiment ("non-switchers") from those who departed from the original bottom performance quartile ("switchers"). Comparison of the confidence gap (and its evolution) of the switchers and non-switchers should help us understand whether there is scope for regression to the mean. As shown in Figure 2.6, I compare the confidence gap of bottom and top performing students based on two selection criteria:

Selection criterion 1 (non-switchers and switchers pooled): bottom or top performing students are those who scored in bottom or top quartile in baseline Math and English, regardless of their performance in subsequent rounds.

Selection criterion 2 (non-switchers only): bottom or top performing students are those who scored in the bottom or top quartile in every test of Math and English.

Figure 2.6: Evolution of the gap between expected and real performance, by two selection criteria of students into bottom or top performers' groups and by treatment status



	SC1 versus SC2	
	Bottom	Top
Round 1	0.648	0.56
Round 2	0.938	0.294
Round 3	0.338	0.004
Round 4	0.452	0.014
Round 5	0.499	0.017

Note: students selected into the bottom or top quartile according to criteria 1 include students who deviated from the bottom or top quartile in subsequent testing rounds; whereas these "switchers" are excluded from criteria 2. SC1/SC2 stands for Selection Criteria 1/2.

Students who stayed in the bottom (or top) performance quartile during the whole academic year scored significantly worse compared to students who ranked in the bottom-quartile in the baseline survey and then switched. In terms of the confidence gap the results are not that straightforward. The differences between control-group switchers and non-switchers in the confidence gaps are mainly insignificant in both subjects. Switchers and non-switchers react to feedback provision similarly with one exception: in Math switchers reduce their inaccurate estimates significantly more compared to non-switchers. Both groups of students seem to overestimate their performance in a similar way. The results do not seem to be driven by regression-to-the-mean (see Appendix A2.8).

Result 4: *Bottom performing students predict their performance in a similar way under the two scenarios – with or without "switchers" being included.*

2.5 Results: The Effects of Repeated Feedback Provision on Student Confidence

We have seen that students do not improve their inaccurate estimations of their own performances by being exposed to a task repeatedly. On the other hand, do they improve if they receive complex feedback about their own performance, the performance of other group members and the position of the group?

2.5.1 Average Treatment Effects of Incentives on Student Confidence

Table 2.1 provides a comparison of the size of the confidence gap in the baseline and endline testing divided by treatment and control groups. There is no statistical difference between the expected gap in the baseline testing among the two treatment groups and control group. The groups are on average the same. In the endline testing, however, students in the two treatment groups overestimated their performance by very similar smaller amounts 31.5 to 32.8% depending on the treatment compared to the control-group students (see Table 2.1 for the results in Math and English).

Table 2.1: Size of the confidence gap in baseline versus endline testing, by subject and treatment or control group

Dependent variable:	Confidence gap			
Subject:	Math		English	
Testing round:	Baseline testing	Endline testing	Baseline testing	Endline testing
Aggregated Control group	14.00	25.11	15.58	18.13
Aggregated within-class feedback group	14.49 {3.5%} [p=0.553]	17.54 {30.1%} **[p=0.000]**	16.26 {4.4 %} [p=0.306]	11.75 {35.2%} **[p=0.000]**
Aggregated across-class feedback group	14.33 {2.4%} [p=0.676]	17.19 {31.5%} **[p=0.000]**	15.89 {2.0%} [p=0.629]	11.47 {36.7%} **[p=0.000]**

Note: the table presents the average size of the gap between expected and real performance by treatment status; the percentage difference of the treatment group compared to the control group in curly brackets and the result of test of differences between treatment and control group in square brackets.

The results are in line with existing literature in which feedback led to improved self-assessment, and to improved judgements and decision making (for example, Engelmann and Strobel, 2000; Duffy and Hopkins, 2005; Ryvkin et al., 2012; Moore and

Cain, 2007; Miller and Geraci, 2011; Hacker et al., 2008). It is important to note, however, that students remain overconfident, and surprisingly the size of the gap remains high over time (between 99.6 and 102.3% of real performance among the treated and 154.3% among the control group students).

Students seem to calibrate their self-predictions more accurately in English compared to Math if they are provided with feedback. The difference can be explained by lack of metacognitive skills as proposed by Kruger and Dunning (1999). Students in the control group, whose performance is a proxy for student evolution in the absence of the treatments, stagnated in Math over the entire academic year with their absolute performance actually decreasing by 0.33%. By contrast, their absolute score in English increased by 50.25% indicating that in this subject it may be easier for students to realize their mistakes and better estimate the accuracy of their responses.

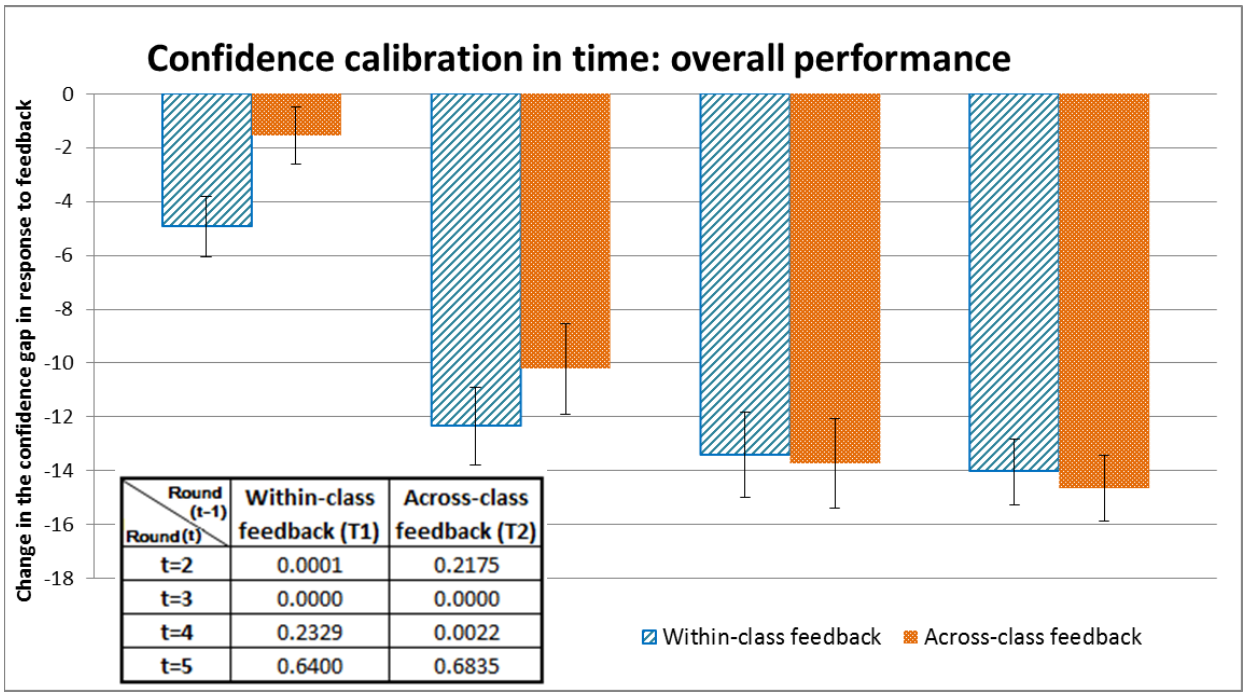
There are no significant differences in the effect sizes with respect to whether within- or across-class feedback is provided, suggesting that the type of feedback does not play a significant role. Both sets of information could be either comparably important, or students could neglect information about others and focus solely on the feedback regarding his/her own performance. The latter would be in line with "reference group neglect" by Camerer and Lovallo (1999).

Does calibration of the confidence gap evolve over time? Figure 2.7 shows the evolution of the average treatment effect of feedback provision on the confidence gap. Each bar represents by how much, on average, treated students decreased their confidence gap compared to the control group students. The table shows the results of the t-test for

whether the changes between subsequent testing rounds for each treatment group separately are significantly different from 0. In other words, it helps to understand the value added of each additional exposure to testing and feedback.

The results show that students react immediately to the provision of feedback. (Note that treatment 2 was implemented with a delay of one testing round due to logistical issues.[51]) The confidence gap for pooled performance shrank by 14.7% in response to the first provision of within-class feedback in testing round 2 and 25.8% in response to across-

Figure 2.7: Changes in the gap between expected and real performance in time: overall performance (Math and English pooled)



Round (t-1) / Round (t)	Within-class feedback (T1)	Across-class feedback (T2)
t=2	0.0001	0.2175
t=3	0.0000	0.0000
t=4	0.2329	0.0022
t=5	0.6400	0.6835

[51] Students were supposed to receive the first feedback approximately three weeks after our first visit. While we were able to evaluate all exams of the within-class feedback group, we lacked the time to evaluate more than 2900 students in the across-class feedback group. Since feedback to the across-class treatment group is based on the results of all students in that group, we failed to have the results on time.

class feedback first provided in testing round 3.[52] The treatment effects are similar in Math and English (see Appendices C2.1 and C2.3).

In order to understand the size of the treatment effect, I calculated the standardized differences in means between treatment and control group (or Cohen's d; Cohen, 1988). Using Cohen's (1988) classification of the effect sizes,[53] one can see small treatment effects in the short run, and medium-to-large effects in the long run, depending on the intervention. Cohen's d for aggregated treatment effects can be found in Table 2.2.

Table 2.2: Effect sizes (Cohen's d) of aggregated treatments, by round and by subject

Cohen's d		Aggregated within-class comparative feedback (Effect Size)	Aggregated across-class comparative feedback (Effect Size)	Aggregated financial rewards (Effect Size)	Aggregated reputational rewards (Effect Size)
Round 2	Sum	0.246	0.089	NA	NA
	Math	0.186	0.093	NA	NA
	English	0.237	0.059	NA	NA
Round 3	Sum	0.502	0.429	NA	NA
	Math	0.513	0.439	NA	NA
	English	0.416	0.345	NA	NA
Round 4	Sum	0.562	0.594	NA	NA
	Math	0.537	0.573	NA	NA
	English	0.531	0.545	NA	NA
Round 5	Sum	0.664	0.716	-0.217	-0.057
	Math	0.603	0.677	-0.168	-0.099
	English	0.619	0.627	-0.240	-0.009

Note: Cohen's d classification (Cohen, 1988): small effect (0.2), medium effect (0.5), large effect (0.8)

[52] The immediate effect of within-class feedback is significantly weaker compared to the effect of across-class feedback. The difference can be attributed either to information spread or extra exposure to tests due to the delay.

[53] Cohen's d classification (Cohen, 1988): small effect (0.2), medium effect (0.5), large effect (0.8)

Result 5: *Provision of group feedback reduces the confidence gap.*

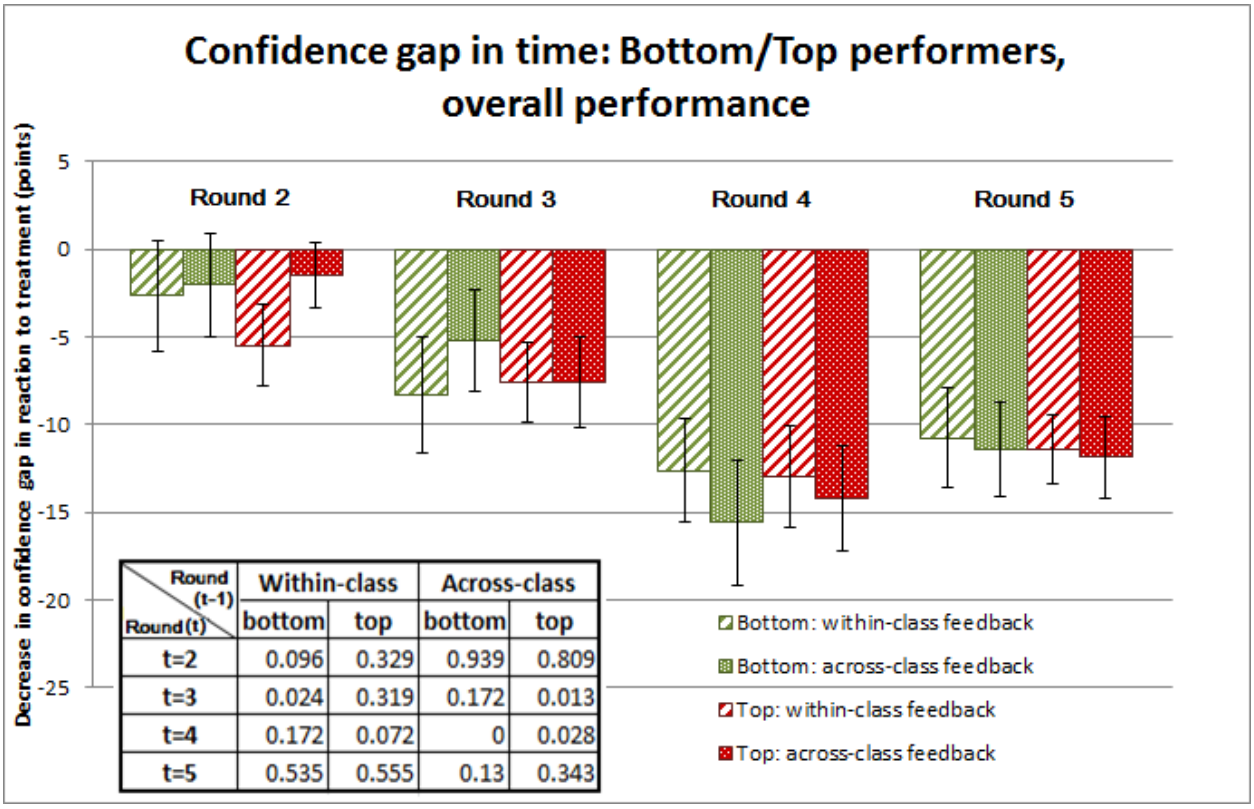
Results 6: *The accuracy of calibration significantly improves in response to the first two feedback provisions with no significant value added of further treatment.*

Result 7: *The type of feedback does not play a role in reducing miscalibration.*

2.5.2 Skilled and Unskilled Students and their Abilities to Realize their Competencies

Do skilled (or top-quartile performers) and unskilled (or bottom-quartile performers) students differ in their ability to improve the assessment of their own performance as predicted by Kruger and Dunning (1999)? Data from this experiment suggest the opposite: both skilled and unskilled students are able to significantly decrease inaccuracy in their self-assessment in a similar way. Figure 2.8 summarizes the average treatment effects in different testing rounds by students' ability (bottom or top-quartile performers). The bars represent the decrease (in points) in the average confidence gap of the treated students compared to the control group students. The small table contains p-values obtained from testing equality of the average treatment effects in two subsequent rounds for each treatment or control group separately. In the baseline testing there were no significant differences in Math between bottom and top performing students in the treatment versus control groups. In English, bottom (top) performing students miscalibrated their performance significantly more (less) compared to the control group. The baseline differences are taken into account in estimating the confidence gaps in rounds 2 through 5 (see Appendix B2.1 for randomization balance).

Figure 2.8: Evolution of the confidence gap of bottom and top performing students



Round (t-1) / Round (t)	Within-class		Across-class	
	bottom	top	bottom	top
t=2	0.096	0.329	0.939	0.809
t=3	0.024	0.319	0.172	0.013
t=4	0.172	0.072	0	0.028
t=5	0.535	0.555	0.13	0.343

Both unskilled and skilled students decrease the confidence gap similarly over time. In the endline testing bottom performers lowered the gap by 10.8 points, whereas top performers decreased it by 11.8 points (the difference is insignificant). The result goes against the predictions of Kruger and Dunning (1999) who predicted that "because of their difficulty recognizing competence in others, incompetent individuals will be unable to use information about their choices and performances of others to form more accurate impressions of their own ability" (p.1122). The type of feedback does not play a significant role.

While the top performing students react immediately to both types of treatment, bottom performers react with a one-round delay to within-class feedback provision. In all other terms the miscalibrations of bottom and top performing students are similar. Both

groups reach the maximum in terms of the level of the confidence gap calibration in testing round 4.

Result 8: *Both the unskilled and the skilled students are able to improve in terms of self-assessment if they receive feedback.*

One may argue that the improvement may be driven by the students who improved their performance and who learned about their abilities over time.[54] I repeated the exercise separately for students who improved during the academic year and those who did not and found similar patterns in both cases (for the detailed results see Appendix C2.4). Greater improvements can be observed among students who improved their performance (see Appendix A2.2).

Result 9: *Even students who do not improve their performance during an academic year reduce their inaccuracy in estimating their own performance if they receive repeated feedback.*

Overall, I find rather mixed evidence for the predictions of Kruger and Dunning (1999). While the unskilled students strongly overestimated their performance, the skilled students did not underestimate their performance. Students from both ends of the performance distribution were able to reduce the gap between the expected and the real score.

[54] Note that Kruger and Dunning (1999) did not predict such a possibility for the unskilled students. They predicted improvements in calibrations only to competent students who gain insight into their metacognitive abilities. The analysis is based on an assumption that improvements in metacognitive skills come hand in hand with the improvements in student performance. This may be a strong assumption, however, but due to the absence of different measures of metacognitive skills, this was the only way to provide more insight.

2.6 Heterogeneity

2.6.1 Is the Effect Dependent on the Task Difficulty?

Burson et al. (2006) attributed the existence of the Dunning-Kruger effect to the perceived difficulty of the task. While in easy and moderate tasks bottom performers overestimated their performance, top performers were highly inaccurate in difficult tasks. The authors found that people from both ends of the ability distribution were prone to similar degrees of miscalibration while they varied the level of the difficulty of tasks used in their sessions. In my case, I asked students during the baseline testing to evaluate how difficult they found the exam in Math and in English relative to the exams they typically experience. Since the measure of the task difficulty differs from that used in Burson et al. (2006), the effect sizes cannot be directly compared. However, it is sufficient to relate the existence of the Dunning-Kruger effect and the students' perception of the difficulty.

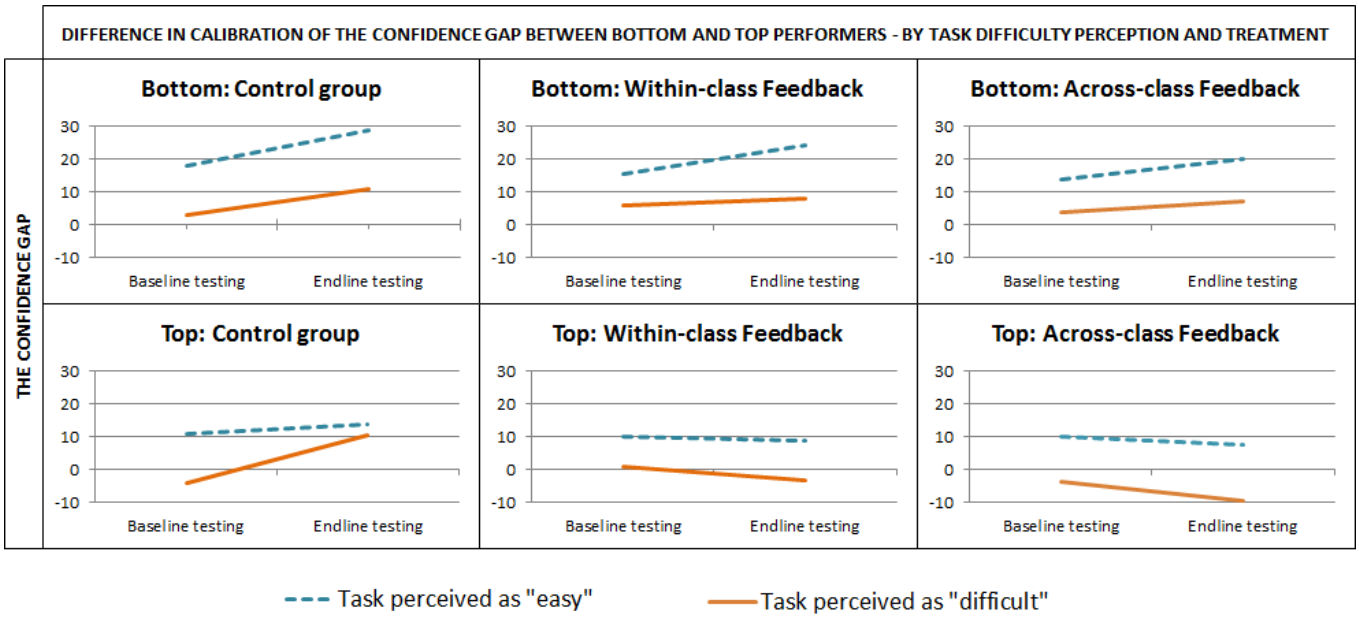
Based on the baseline testing data, the students who perceive exams as easier (relative to common tasks) come from significantly stronger backgrounds compared to the students who perceive them as more difficult.[55] As shown in Figure 2.9, on tasks perceived as easier all students overestimated their performance with the top performers being more accurate compared to the bottom performers (see also Appendix C2.8). The pattern is similar in both baseline and endline testing rounds in both Math and English. If the students perceived the task as more difficult, the unskilled students overestimated their performance and the skilled students under-estimated their performance with the

[55] In Math (English), the average score of the students who perceived the task as rather difficult was 8.945 (8.145) points, while the average score of students who perceived it as easy was 12.687 (12.704).

difference between over and under-estimation insignificant (p-value=0.178 in Math and p-value=0.175 in English).

Result 10: *There is supportive evidence for the existence of the Dunning-Kruger effect - even if the students receive repeated feedback - if the students perceive the task to be difficult. If the students perceive the task to be easy, students are predominantly overconfident.*

Figure 2.9: The differences in the calibration of the confidence gap between bottom- and top-performers and by their perception of task difficulty



Burson et al. (2006) attribute students' inaccuracy in predictions to the lack of understanding of how well their peers do in a given task. It is therefore expected that if students are informed about the performance of others, the level of inaccuracy should decrease, the unskilled should become less over-confident and the skilled more over-

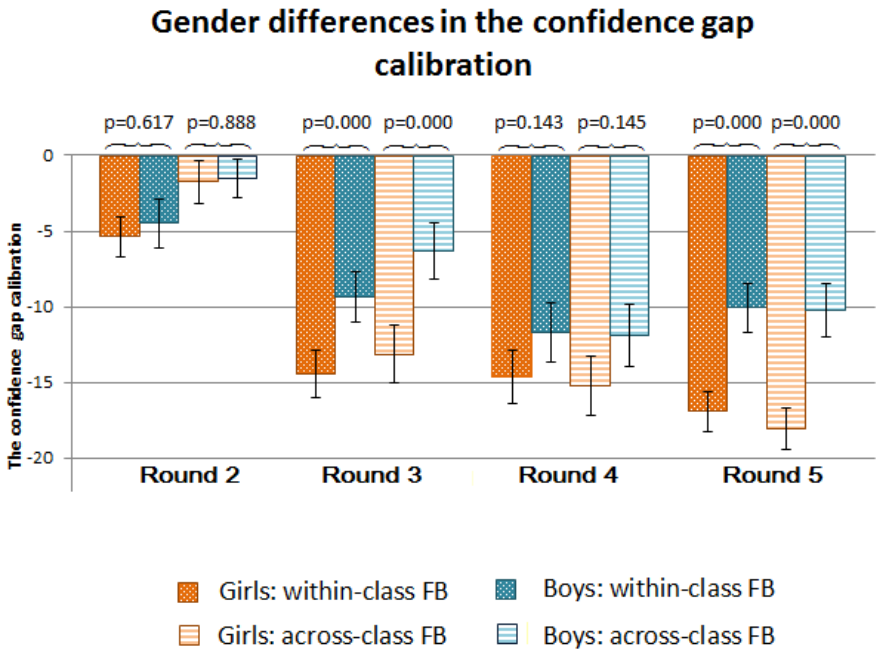
confident. Looking at the endline predictions of students in this study, I can see that the pattern persists even if students were exposed to repeated feedback. The most plausible explanation is that students only extracted information about themselves from the feedback they received and neglected information about others. Such results would be in line with "reference group neglect" introduced by Camerer and Lovallo (1999).

2.6.2 Gender Differences

Boys do not improve their performance in response to pure feedback provision while girls do (see the results in Chapter 1 of this dissertation). A possible explanation could be that boys ignore the information. In such case they would not improve accuracy in their estimations of the own performance either. In the baseline testing, the confidence gap between girls and boys is insignificant. The results show that both girls and boys use information to lower the confidence gap, for girls by 37.5%, and for boys by 24.5% in the final testing round. Boys update significantly less in all testing rounds except the second round when the gender difference is insignificant. The type of treatment does not play a significant role – neither for girls nor for boys - suggesting that they either put equal weight on within-class and across-class feedback or they extract information only about their own performance. The evolution of the confidence gap calibration in response to treatments is shown in Figure 2.10 and Appendices C2.7 and C2.10.

Result 11: *Girls decrease their inaccuracy in assessment of their performance significantly more compared to boys.*

Figure 2.10: The evolution of the confidence gap calibration – by gender and testing round



Note: p-values report the results of testing the gender differences in calibrations of confidence gap in response to different feedback provision.

2.6.3 Student Confidence in Competition for Monetary and Non-Monetary Rewards

What happens to the student confidence level if they are included in a competition for rewards[56]? Overall, one control and eight treatment groups were formed – four pure treatment groups (two feedback and two reward groups) and four groups with treatment interactions (two types of feedback combined with two types of rewards). Such design allows me to disentangle what happens to student confidence once they are included in group competition with or without perfect information about their own performance, performance of group members and that of competitors.

[56] Note that the rewards were introduced right before the last testing round; therefore, the immediate effects can be studies only for the feedback treatment groups.

The design of the experiment therefore offers a comparison of two different competitive environments. The first is based on intergroup comparison and repeated provision of feedback about group performance. It represents competition for students who are intrinsically competitive and seek comparison to others without being incentivized. The second competitive environment is based on a tournament in which students have a chance to win monetary or non-monetary rewards.

Students who receive either pure feedback or feedback together with monetary or non-monetary rewards overestimated their performance significantly less, i.e., by only 22 to 32%. On the contrary, students who received no feedback but participated in a competition increased the confidence gap by 6-8% more. The result is driven purely by an increase in girls' overconfidence (11%), whereas the boys' confidence gap remains unchanged. In other words, in the scenario with no feedback, boys are similarly inaccurate about their expectations of their performance, whether they are included in a tournament for rewards or not, but girls' expectations increase significantly with the introduction of rewards. Girls' self-assessment inaccuracy increases while their performance remains the same whereas boys' expectations remain similarly inaccurate (compared to the control group boys) and their performance significantly improves. Girls therefore may remain behind boys in terms of improvements in tournaments for rewards due to inflated expectations about their performance. The current experiment, however, does not offer further details and it would be interesting to see possible explanations in future studies. Information provision is the driving force of students' perception calibration (for details see Appendix C2.2).

***Result 12**: The inclusion of students in a group competition for rewards (without any feedback regarding their performance) may increase their inaccuracy of assessment of one's own performance, at least in the case of females.*

2.6.4 Differences by Age

One may argue that the reaction to feedback may differ by students' age, either because of more years of education, improvement in the topic or simply because they become more "mature." I interacted the age of students with treatment dummies. The results suggest that age does not play a significant role in calibration improvements. The reason could be that unlike developed countries in which the age range within a class is around 1 year, the average age range in Ugandan primary and secondary schools is 7 years (i.e., students' age in the 6th grade of primary school range from the official age of 12/13 up to 19).

Students are very often absent, with the absence rate varying between 19.2 and 29.3% during our visits, and they are often asked to repeat the classes. Only 51.1% of students indicated that they had not repeated any class before the time they were interviewed. The results presented in Appendix C2.6 compare the calibration by self-assessment of students at the official age in their respective classes (i.e., 12-13 years old in the 6th grade in primary school) to students who are older than the official age. The results suggest that students at the official age calibrate their expectations regarding their own-performance similarly compared to the students who are older than the official age. While the reasons some students delayed their school attendance do not seem to matter in calibrations of their self-assessments, the level of study (primary or secondary school) does

seem to play a significant role in the magnitude of calibration improvement (the pattern is similar). Students in secondary schools decrease the perception gap to a greater degree compared to primary school students. The difference can be attributed to higher selection criteria in secondary schools. The results are presented in Appendix C2.5.

2.7 Summary

People are often overconfident about their outcomes. The participants in this experiment - primary and secondary school students in Southern Uganda – are not an exception. Overconfidence may be costly and therefore researchers have searched for ways to improve people's self-assessment of their own performance. Feedback typically helps to increase the accuracy of the assessment and lower the gap between perceived and real performance. People, however, remain overconfident.

This paper contributes to the current literature in the following ways. First, the results shed light on what happens to students' overconfidence if students are evaluated in groups, and they repeatedly receive feedback about their own performance and the performance of their groups. The type of feedback does not play a significant role in students' accuracy of calibration of their self-assessment. A possible explanation is that students neglect information about others (as suggested by Camerer and Lovallo, 1999) even in the scenarios where they are evaluated in groups and their group score therefore depends directly on the score of other group members. The provision of feedback helps students to lower inflated self-assessments of their own performance; however, the overconfidence persists.

Students seem to significantly improve their calibrations in response to the first two feedback provisions, followed by insignificant improvements in the confidence gap with any extra treatment. Girls decrease their inaccuracy in self-assessment significantly more compared to boys. In a competitive environment, when students were randomly selected to compete for monetary or non-monetary rewards but received no feedback previously about their performance, girls became significantly more overconfident, whereas boys were not affected.

Furthermore the results of this experiment bring evidence from primary and secondary schools in Uganda to the debate regarding the existence of the unskilled-and-unaware phenomenon (also known as the Dunning-Kruger effect first documented by Kruger and Dunning, 1999). Unskilled students grossly overestimate their performance and skilled students underestimate their performance if they perceive the task to be easier compared to the tasks used at schools. If the students of this experiment perceived the task to be more difficult, both bottom and top performers were overconfident. If they perceived the task to be easier or of comparable difficulty, students were predominantly overconfident (the unskilled significantly more compared to the skilled students). The results are in line with Burson et al. (2006) who first documented the conditionality of the existence of the Dunning-Kruger effect on task difficulty. The results also show that the unskilled students improve in their accuracy even if they do not improve their performance. Such a result is against the prediction of Kruger and Dunning (1999) who expected no ability of the unskilled students to improve in response to feedback provision. I do not find any support to suggest that the results would be driven by the regression-to-the-mean.

Appendix 2

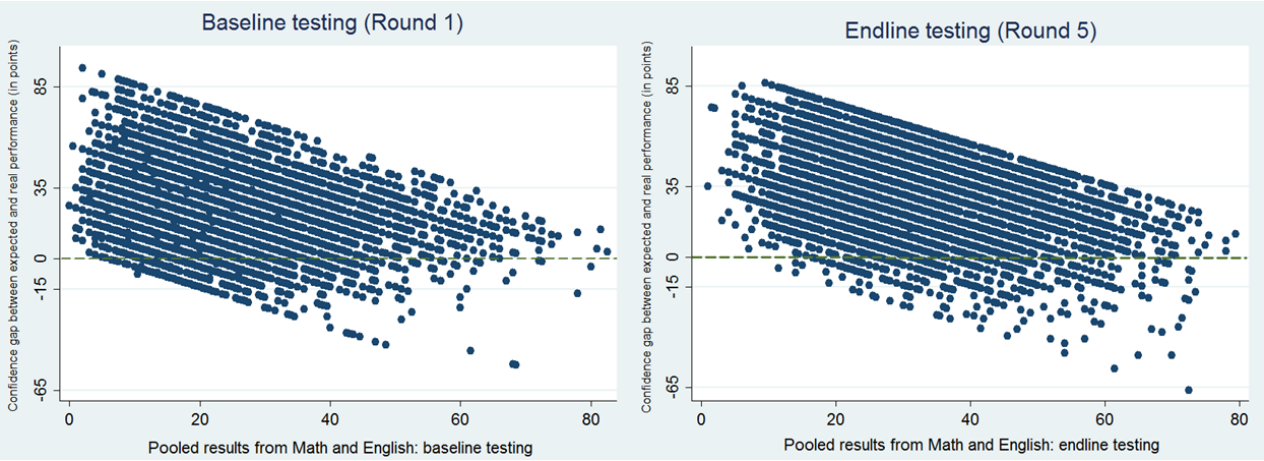
Appendix A2: Description of data

Appendix A2.1: Summary statistics by the level of miscalibration of own performance, baseline testing

	Overestimated self-assessment	Calibrated Self-assessment	Underestimated Self-assessment
Overall AM (AE), in %	81.0 / 85.3	11.2 / 8.9	7.8 / 5.8
Girls AM (AE), in %	81.1 / 85.6	11.4 / 8.8	7.5 / 5.6
Boys AM (AE), in %	80.9 / 85.0	10.9 / 8.9	8.2 / 6.1
Math: performance (points)	11.56	12.45	15.65
English: performance (points)	11.39	12.82	18.46
Math: estimation gap (in points)	14.89	0.21	-6.49
English: estimation gap (in points)	15.14	0.48	-6.56
Absenteeism (in %)	28.7	32.7	28.2
Stress (in %)	7.0	7.2	7.2
Subjective happiness (level)	10.7	11.6	11.3
Effort (in Math, level)	3.7	3.4	3.2
Effort (in English, level)	3.6	3.3	3.3

Note: first three rows describe the percentage represented by over/well/underestimated self-assessment (for the overall sample, girls or boys only); "estimation gap" stands for the difference between anticipated number of points and the corresponding real score, "absenteeism" accounts for percentage of students absent during the last visit given their presence in the baseline testing, "stress" level is measured by the Stressed Perceived Scale (a 0-4 Likert scale), "subjective happiness" by the Subjective Happiness Scale (a 7-point Likert scale); "effort" is a subjective measure of the level of effort subject exerted in Math or English (a 5-point Likert scale).

Appendix A2.2: Scattegram – Confidence gap and real performance, baseline and endline testing

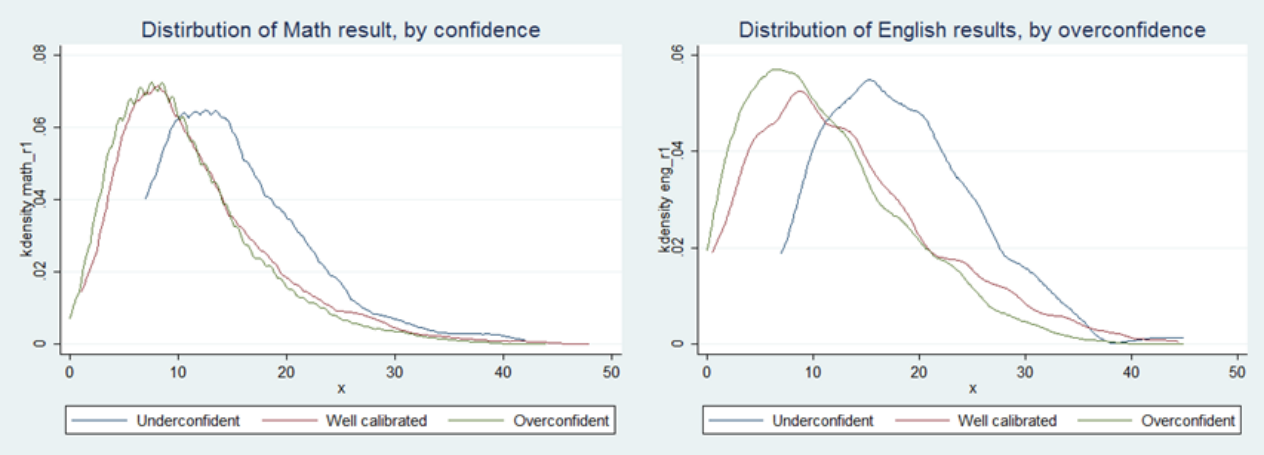


Appendix A2.3: Wilcoxon matched-pairs signed-ranks test of equality of matched pairs of observations, by round and by subject

Difference between expected and real performance		**p-value**
Round 1	Math	0.0000
	English	0.0000
Round 2	Math	0.0000
	English	0.0000
Round 3	Math	0.0000
	English	0.0000
Round 4	Math	0.0000
	English	0.0000
Round 5	Math	0.0000
	English	0.0000

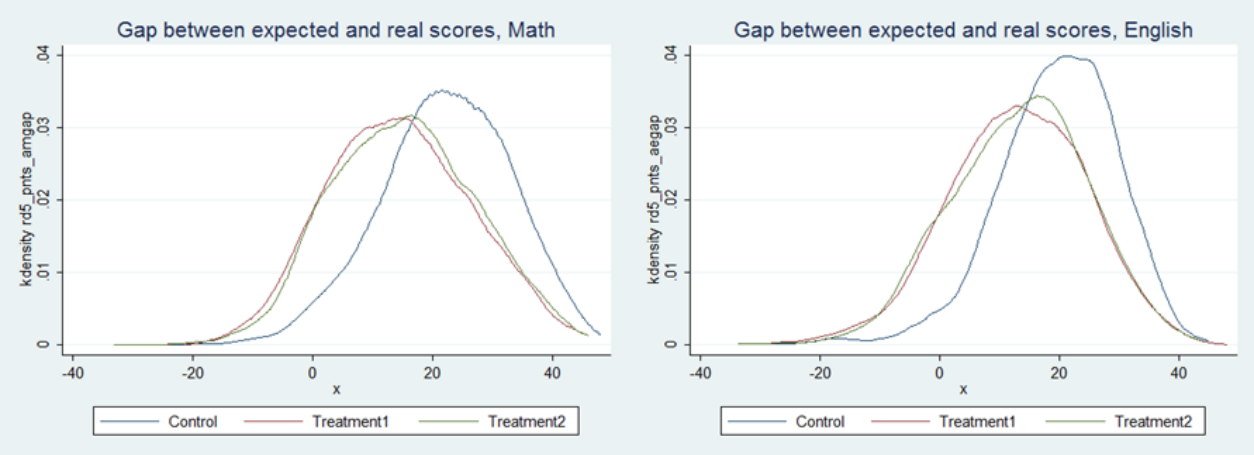
Note: the null hypothesis is that the observations come from the same distribution.

Appendix A2.4: Comparison of baseline distribution of Math and English score, by confidence level



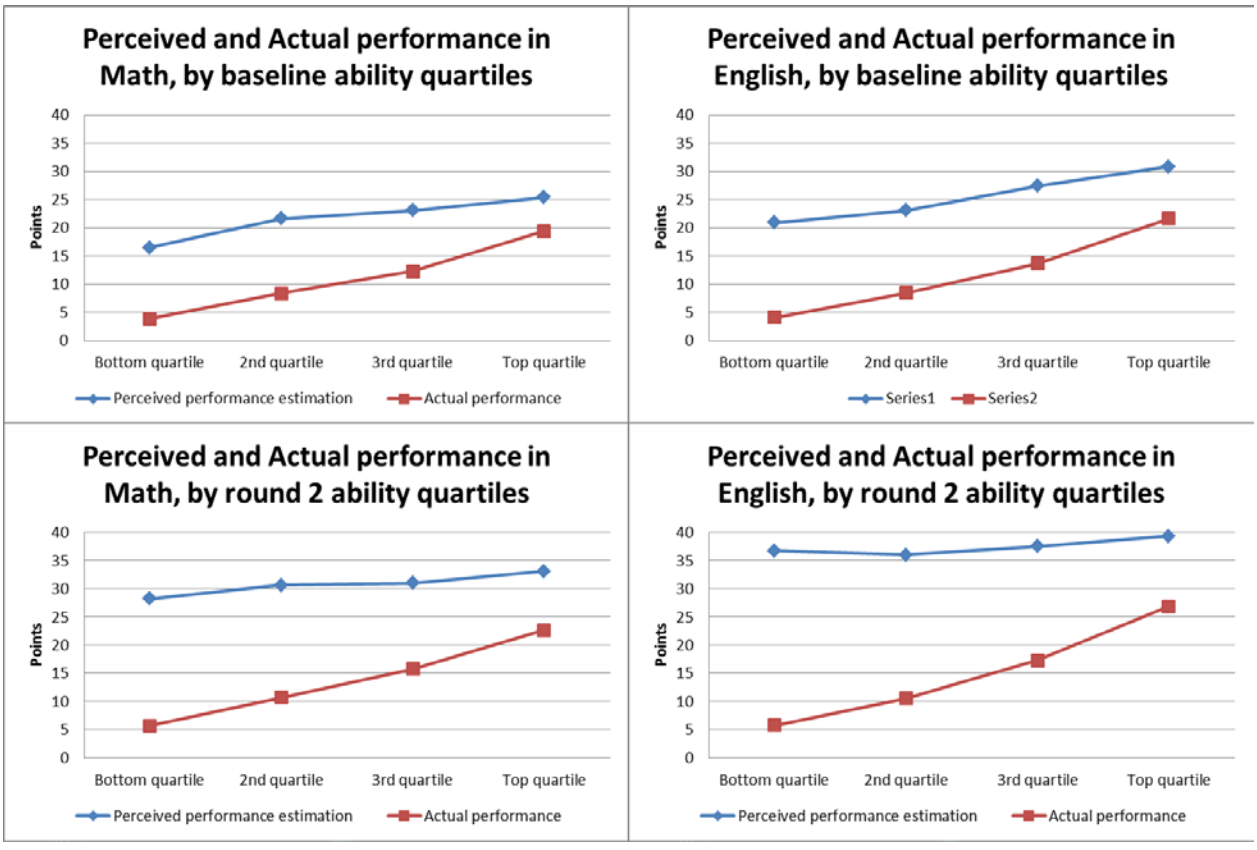
Note: A student is considered to be underconfident if his/her real performance was higher than the upper limit of the interval he/she predicted his/her real score would belong to, well-calibrated if the real score was within the predicted interval, and overconfident if the real score was below the lower limit of the interval he/she predicted for his/her real score. In both Math and English underconfident students performed better than overconfident students and students with well-calibrated expectations.

Appendix A2.5: Comparison of the distribution of the confidence gap in the final testing round, by subject and treatment status

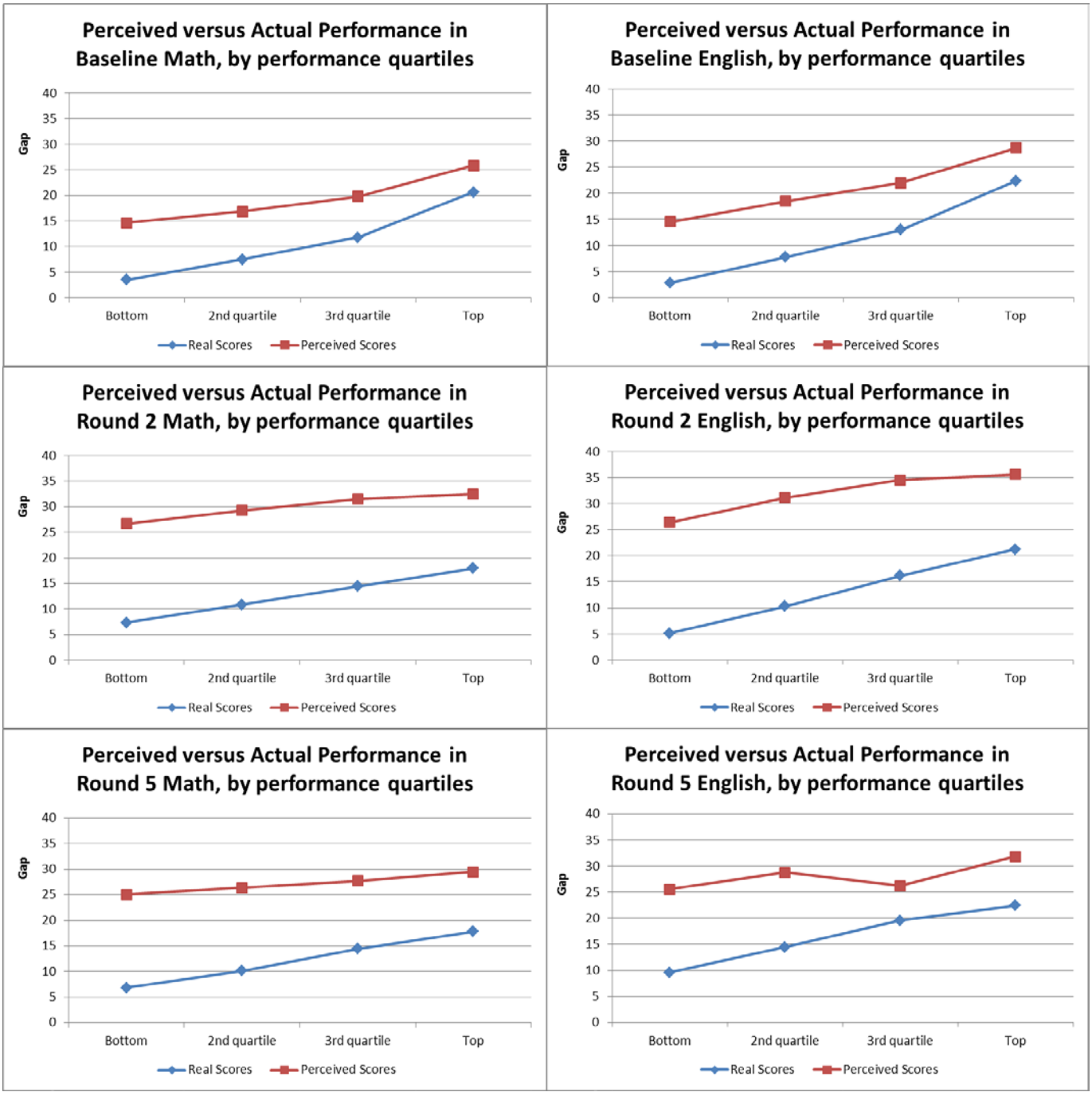


Note: T1 stands for within-class feedback, T2 for across-class feedback and Control for control group without any feedback. A student is considered to be underconfident if his/her real performance was higher than the upper limit of the interval he/she predicted his/her real score would belong to, well-calibrated if the real score was within the predicted interval, and overconfident if the real score was below the lower limit of the interval he/she predicted for his/her real score. The confidence gap of the control group students is significantly higher compared to the treated students.

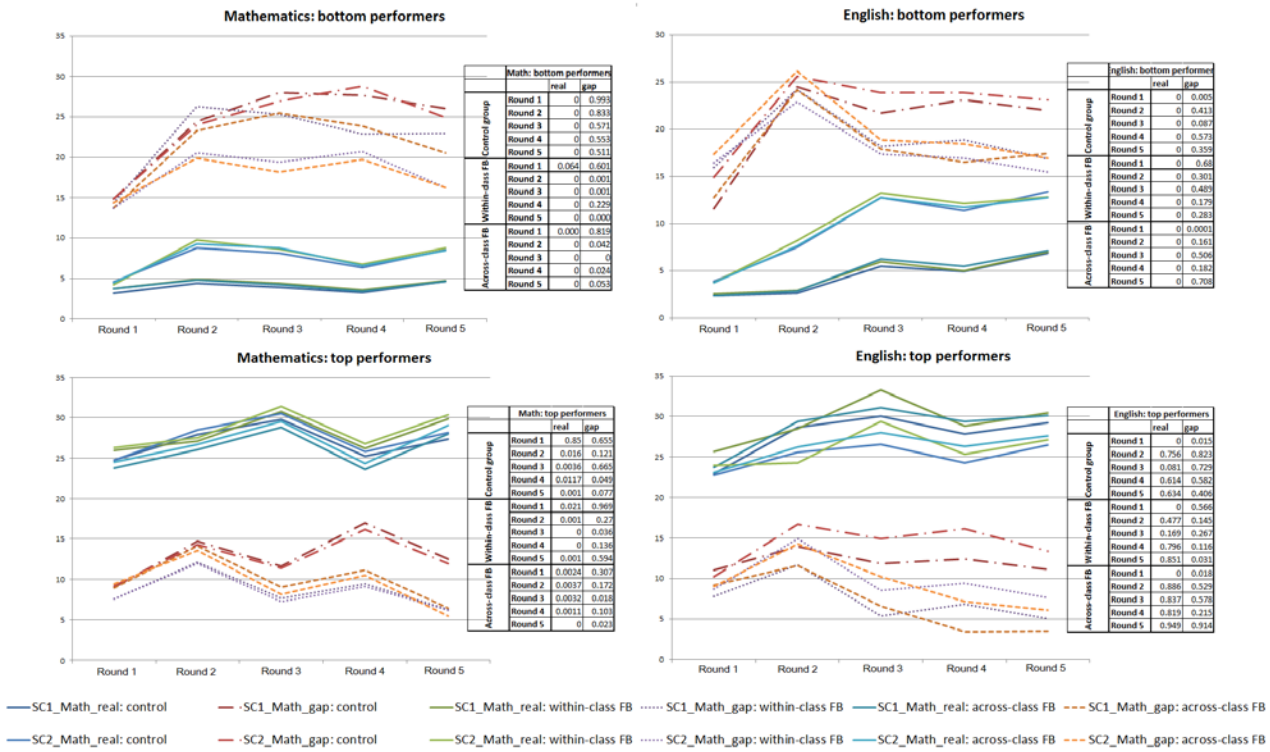
Appendix A2.6: Perceived versus actual performance – division into bottom/top quartile based on the performance in Round 2



Appendix A2.7: Evolution of perceived versus actual performance in time, by subject and testing round



Appendix A2.8: The evolution of the performance and the confidence gap of bottom and top quartile students who stayed in or switched from the initial performance quartile – by treatment status



Note: SC1 stands for scenario 1 in which only non-switchers are taken into account (i.e., bottom (top) performers who scored in the bottom/top quartile in all testing rounds). SC2 stands for scenario 2 in which only switchers are taken into account (students who performed in the bottom (top) quartile in the baseline survey but performed in a better (worse) quartile in subsequent testing(s)). Students in both scenarios expect similar scores.

Appendix B2: Randomization balance and baseline testing

Appendix B2.1: Comparison of mean characteristics of students in treatment and control groups, randomization into feedback treatment and no-feedback control groups

Dependent variable:	Confidence gap (baseline testing)			
Subject:	Math		English	
Sample:	Bottom performers	Top performers	Bottom performers	Top performers
EXAM SCORE				
Within-class feedback (T1)	3.773	20.238	4.512	20.959
Across-class feedback (T2)	3.668	19.967	4.488	20.488
Control (C)	3.240	20.059	4.447	19.422
Joint p-value	[p=0.269]	[p=0.906]	[p=0.967]	**[p=0.060]**
PERCEIVED SCORE				
Within-class feedback (T1)	15.046	29.004	29.264	28.448
Across-class feedback (T2)	11.518	29.861	25.831	29.679
Control (C)	13.841	28.549	23.607	30.068
Joint p-value	[p=0.261]	[p=0.902]	**[p=0.000]**	[p=0.553]

Note: Mean comparisons. P-values from testing joint differences between treated and control groups are in square brackets. Bottom-/Top-performers are students whose performances fall in the bottom/top performance quartile in every testing round.

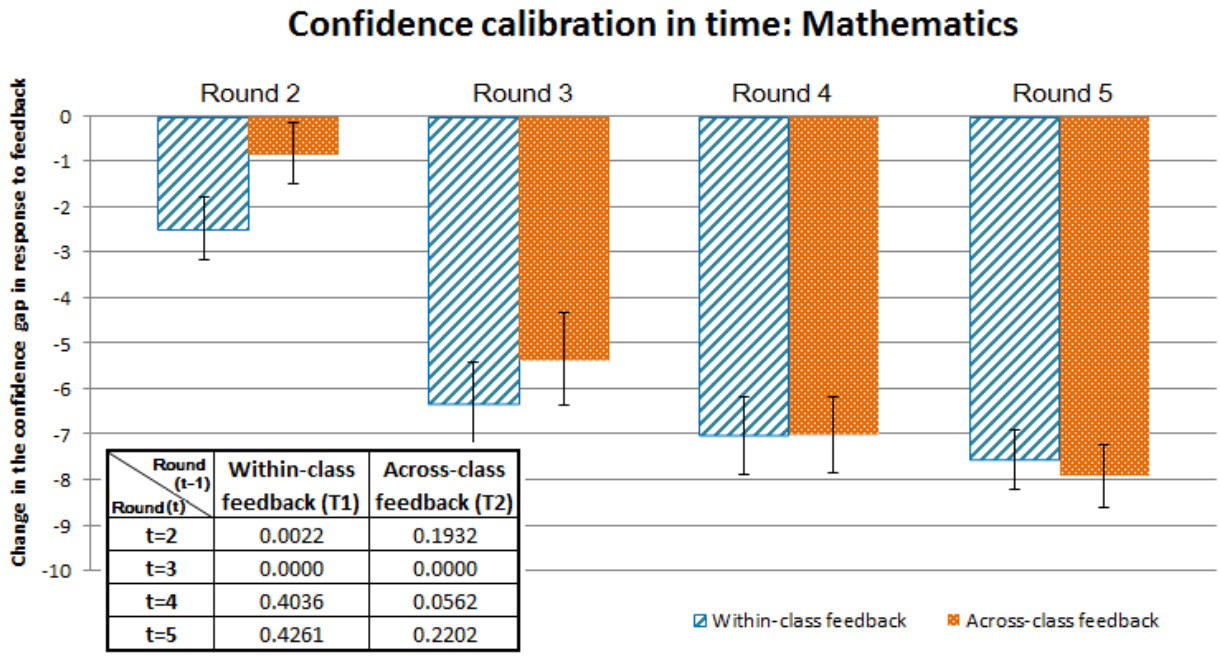
Appendix B2.2: Relation between initial performance and initial confidence gap

Dependent variable:	Initial confidence gap			
Specification:	(1)	(2)	(3)	(4)
Initial overall performance	**-0.322*** **	**-0.323*** **	**-0.320*** **	**-0.319*** **
(Math and English pooled)	**(0.035)**	**(0.035)**	**(0.038)**	**(0.038)**
Within-class feedback		0.734		1.342
		(1.492)		(1.466)
Across-class feedback		0.517		0.801
		(1.379)		(1.292)
Controlled for stratas	No	No	Yes	Yes
Number of observations	3450	3450	3450	3450

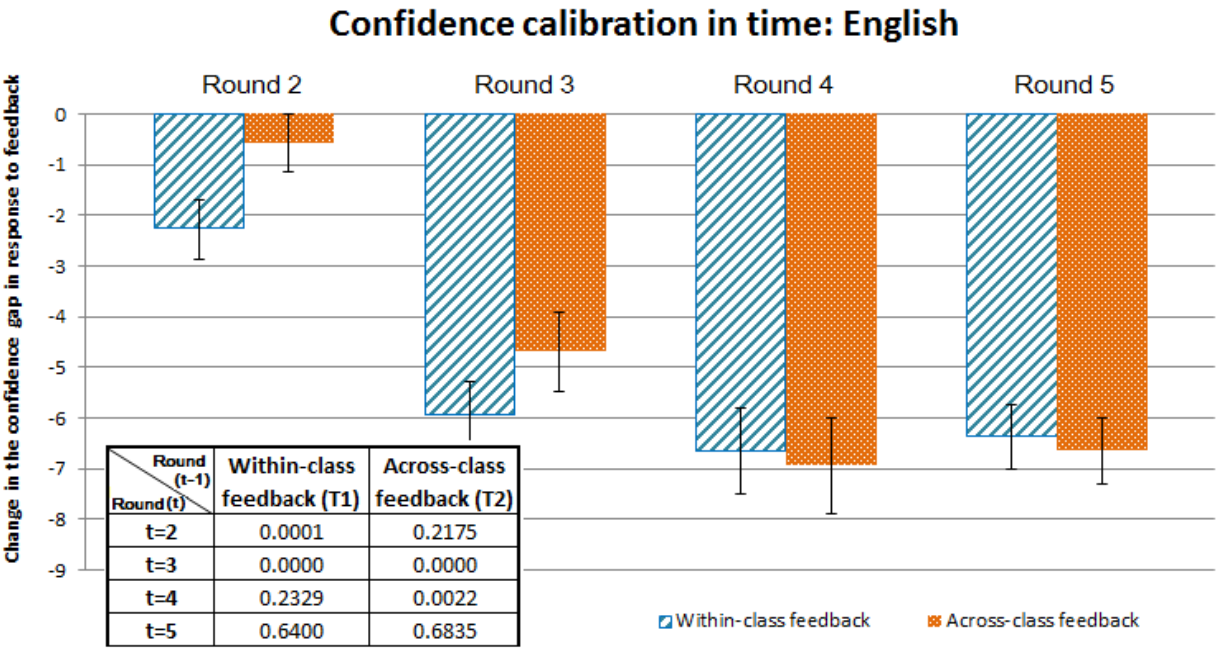
Note: OLS. Robust standard errors adjusted for clustering at class level are presented in brackets. Controlled for stratum fixed effects (area, level and school performance in national examinations). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2: Average Treatment Effects

Appendix C2.1: The evolution of the confidence gap in time: by subject and by type of treatment



Round (t-1) / Round (t)	Within-class feedback (T1)	Across-class feedback (T2)
t=2	0.0022	0.1932
t=3	0.0000	0.0000
t=4	0.4036	0.0562
t=5	0.4261	0.2202



Round (t-1) / Round (t)	Within-class feedback (T1)	Across-class feedback (T2)
t=2	0.0001	0.2175
t=3	0.0000	0.0000
t=4	0.2329	0.0022
t=5	0.6400	0.6835

Note: OLS. Error bars show robust standard errors adjusted for clustering at class level. Controlled for stratum fixed effects (area, level and school performance in national examinations). The table shows test results from testing significance in gaps in two subsequent rounds (i.e., it tests how significant the improvements in calibrations are in the two following rounds).

Appendix C2.2: OLS estimates of the effects of different motivation schemes on the student confidence gap – short-term versus long-term effects

Dependent variable:	Confidence Gap					
Time perspective:	Short-term effects		Long-term effects		Long-term effects	
Subject:	Math	English	Math	English	Math	English
PANEL A: AGGREGATED TREATMENT EFFECTS						
Within-class feedback (T1)	**-2.475*** **	**-2.278*** **	**-7.573*** **	**-6.385*** **	**-7.685*** **	**-6.413*** **
	(0.682)	**(0.593)**	**(0.659)**	**(0.635)**	**(0.639)**	**(0.621)**
Across-class feedback (T2)	-0.827	-0.577	**-7.927*** **	**-6.658*** **	**-7.875*** **	**-6.625*** **
	(0.674)	(0.572)	**(0.676)**	**(0.647)**	**(0.663)**	**(0.634)**
Financial Rewards (Finrew)					1.361	1.336
					(0.864)	(0.913)
Reputational Rewards (Reprew)					1.223	-0.148
					(0.835)	(0.815)
Baseline Score Anticipation	**0.319*** **	**0.381*** **	**0.244*** **	**0.264*** **	**0.243*** **	**0.264*** **
	(0.023)	**(0.023)**	**(0.022)**	**(0.018)**	**(0.018)**	**(0.021)**
Controlled for stratas	Yes	Yes	Yes	Yes	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of observations	3417	3376	3417	3376	3417	3376
PANEL B: INTERACTION OF ALL TREATMENTS						
PURE TREATMENTS						
Within-class social comparison					**-7.261*** **	**-5.748*** **
(T1_solo)					**(0.938)**	**(0.936)**
Across-class social comparison					**-7.694*** **	**-6.310*** **
(T2_solo)					**(0.935)**	**(0.945)**
Financial Rewards					1.262	1.580
(Fin_solo)					(1.149)	(1.024)
Reputational Rewards					**2.052***	0.889
(Rep_solo)					**(1.077)**	(1.133)
TREATMENT INTERACTIONS						
Within-class feedback with					**-6.274*** **	**-5.336*** **
financial reward (T1_fin)					**(1.425)**	**(1.483)**
Across-class feedback with					**-5.879*** **	**-4.031*** **
financial reward (T2_fin)					**(1.001)**	**(1.027)**
Within-class feedback with					**-6.536*** **	**-6.277*** **
reputational reward (T1_rep)					**(1.106)**	**(1.083)**
Across-class feedback with					**-6.697*** **	**-7.000*** **
reputational reward (T2_rep)					**(1.429)**	**(1.235)**
Baseline Score Anticipation					**0.244*** **	**0.266*** **
					(0.022)	**(0.021)**
Controlled for stratas					Yes	Yes
Prob > F					0.0000	0.0000
Number of observations					3417	3376

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, school performance at national examination and grade level). * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.3: Average treatment effects of different incentive schemes on the confidence gap, by bottom versus top performers and by subject

Dependent variable:	Confidence gap					
Subject:	Math			English		
Sample:	Bottom performers	Top performers	Difference (Bottom versus Top)	Bottom performers	Top performers	Difference (Bottom versus Top)
ROUND 1 (BASELINE)						
Within-class feedback	2.771	-1.506	**4.277****	**3.309****	**-3.679*****	**6.989*****
	(1.754)	(1.328)	**(1.995)**	**(1.366)**	**(1.230)**	**(1.706)**
			[p=0.034]			**[p=0.000]**
Across-class feedback	1.946	-1.443	**3.389****	0.032	**-3.101*****	**3.133****
	(1.647)	(1.156)	**(1.704)**	(0.841)	**(1.137)**	**(1.302)**
			[p=0.049]			**[p=0.017]**
Differences (Bottom/Top:	0.825	-0.063		**3.277****	-0.578	
Within versus Across	(2.295)	(1.526)		**(1.382)**	(1.510)	
FB)	[p=0.720]	[p=0.967]		**[p=0.019]**	[p=0.702]	
ROUND 2						
Within-class feedback	**3.835****	**-8.024*****	**11.859*****	**-2.624***	**-8.748*****	**6.124*****
	(1.758)	**(1.222)**	**(2.002)**	**(1.544)**	**(1.016)**	**(1.700)**
			[p=0.000]			**[p=0.000]**
Across-class feedback	0.143	**-6.434*****	**6.577*****	-1.942	**-8.608*****	**6.666*****
	(1.948)	**(1.055)**	**(2.223)**	(1.296)	**(0.971)**	**(1.462)**
			[p=0.004]			**[p=0.000]**
Differences (Bottom/Top:	3.692	-1.589		-0.682	-0.139	
Within versus Across	(2.443)	(1.324)		(1.992)	(1.210)	
FB)	[p=0.133]	[p=0.232]		[p=0.733]	[p=0.908]	
ROUND 5 (ENDLINE)						
Within-class feedback	-1.676	**-13.277*****	**11.600*****	**-4.413*****	**-12.989*****	**8.577*****
	(1.729)	**(1.246)**	**(2.033)**	**(1.171)**	**(1.064)**	**(1.463)**
			[p=0.000]			**[p=0.000]**
Across-class feedback	**-4.690*****	**-14.133*****	**9.445*****	**-2.953*****	**-14.437*****	**11.485*****
	(1.411)	**(1.456)**	**(1.962)**	**(0.986)**	**(1.011)**	**(1.299)**
			[p=0.000]			**(p=0.000]**
Differences (Bottom/Top:	3.012	0.856		-1.460	1.448	
Within versus Across	(2.134)	(1.812)		(1.372)	(1.332)	
FB)	[p=0.160]	[p=0.637]		[p=0.289]	[p=0.279]	

Note: OLS. Bottom-/Top-performers are students whose performances fall in the bottom/top performance quartile in all five testing rounds. The table provides comparison in Round 1 (baseline testing), Round 2 (short-term effects), and Round 5 (endline testing, long-term effects). Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Columns 4 and 7 describe difference between bottom versus top performers and test results for the difference in square brackets. Differences within bottom/top performers by the type of feedback are presented in the last row in each section. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.4: Comparison of the calibration patterns of the students who did or did not improve their performance over time

Dependent variable:	Confidence gap					
Subject:	Math			English		
Sample:	Bottom performers	Top performers	Difference (Bottom versus Top)	Bottom performers	Top performers	Difference (Bottom versus Top)
PANEL A: ONLY STUDENTS WHO DID IMPROVE THEIR PERFORMANCE						
Within-class feedback	-5.716	**-4.892****	0.824	-8.848	**-4.396*****	4.452
	(4.161)	**(2.129)**	(4.658)	(6.015)	**(1.494)**	(6.085)
			[p=0.860]			[p=0.469]
Across-class feedback	**-8.355****	**-5.223****	3.132	-6.027	**-10.101*****	-4.075
	(3.925)	**(2.347)**	(4.349)	(6.499)	**(2.924)**	(7.168)
			[p=0.474]			[p=0.573]
Differences	2.639	0.331		-2.822	**5.705***	
(Bottom/Top: Within	(2.767)	(2.215)		(3.639)	**(2.954)**	
versus Across FB)	[p=0.343]	[p=0.882]		[p=0.443]	**[p=0.061]**	
PANEL B: ONLY STUDENTS WHO DID NOT IMPROVE THEIR PERFORMANCE						
Within-class feedback	-1.973	**-4.691*****	-2.718	**-6.469*****	**-5.335*****	1.134
	(2.987)	**(1.541)**	(3.308)	**(1.505)**	**(1.152)**	(1.889)
			[p=0.413]			[p=0.549]
Across-class feedback	**-4.486***	**-6.391*****	-1.905	**-4.588*****	**-6.322*****	-1.734
	(2.597)	**(2.018)**	(3.261)	**(1.562)**	**(1.270)**	(1.935)
			[p=0.561]			[p=0.372]
Differences	2.513	1.699		-1.881	0.987	
(Bottom/Top: Within	(2.709)	(1.867)		(1.333)	(1.165)	
versus Across FB)	[p=0.356]	[p=0.365]		[p=0.161]	[p=0.399]	

Note: Note: OLS. Bottom-/Top-performers are students whose performances fall in the bottom/top performance quartile in all five testing rounds. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Column 4 and 7 describe difference between bottom versus top performers and test results for the difference in square brackets. Differences within bottom/top performers by the type of feedback are presented in the last row in each section. Panel A and B differ in selection of students into the bottom/top quartile. While bottom/top performing students in the Panel A improved their performance between baseline and endline testing, students in Panel B did not improve their performance. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.5: OLS estimates of the effects of different motivation schemes on student confidence gap – by level of study

Dependent variable:	Confidence gap					
Subject:	Math			English		
Study level:	Primary	Secondary	difference	Primary	Secondary	difference
PURE TREATMENTS						
Within-class comparison	**-6.619*** **	**-8.184*** **	[p=0.428]	**-5.114*** **	**-6.459*** **	[p=0.521]
(T1_solo)	**(0.030)**	**(1.144)**		**(1.527]**	**(1.437)**	
Across-class comparison	**-7.525*** **	**-7.239*** **	[p=0.872]	**-5.370*** **	**-7.022*** **	[p=0.392]
(T2_solo)	**(1.245)**	**(1.259)**		**(1.087)**	**(1.594)**	
Financial Rewards	1.740	-2.871	[p=0.117]	**1.873***	-1.745	[p=0.138]
(Fin_solo)	(1.634)	(2.444)		**(1.001)**	(2.222)	
Reputational Rewards	**3.256** **	-2.739	[p=0.029]	**2.188***	**-5.139** **	[p=0.009]
(Rep_solo)	**(1.421)**	(2.359)		**(1.312)**	**(2.455)**	
TREATMENT INTERACTIONS						
Within-class feedback with	**-5.541*** **	**-9.772*** **	[p=0.217]	**-4.658** **	**-8.325*** **	[p=0.205]
financial rewards (T1_fin)	**(1.945)**	**(2.819)**		**(1.927)**	**(2.158)**	
Across-class feedback with	**-4.718*** **	**-9.909*** **	[p=0.042]	-2.115	**-9.054*** **	[p=0.015]
financial rewards (T2_fin)	**(1.382)**	**(2.146)**		(1.928)	**(2.105)**	
Within-class feedback with	**-5.459*** **	**-9.618*** **	[p=0.064]	**-4.615*** **	**-10.540*** **	[p=0.027]
reputational rewards (T1_rep)	**(1.409)**	**(1.747)**		**(1.639)**	**(2.069)**	
Across-class feedback with	**-5.624*** **	**-8.758*** **	[p=0.229]	**-6.205*** **	**-9.119*** **	[p=0.190]
reputational rewards (T2_rep)	**(2.040)**	**(1.615)**		**(1.678)**	**(1.463)**	
Baseline perception	**0.205*** **	**0.283*** **		**0.272*** **	**0.248*** **	
	(0.030)	**(0.031)**		**(0.028)**	**(0.032)**	
Baseline Score	**-0.307*** **	**-0.457*** **		**-0.190*** **	**-0.467*** **	
	(0.069)	**(0.047)**		**(0.065)**	**(0.054)**	
Controlled for stratas	Yes	Yes		Yes	Yes	
Prob > F	0.0000	0.0000		0.0000	0.0000	
Number of observations	3417	3417		3376	3376	

Note: OLS. Primary stands for primary school (level P6 and P7) and secondary for secondary schools (levels S1 up to S4). Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Column 4 and 7 describe difference between bottom versus top performers and test results for the difference in square brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.6: OLS estimates of the effects of different motivation schemes on student confidence gap – students at official age versus older

Dependent variable:	Confidence gap					
Subject:	Math			English		
School Age:	Official	Unofficial	difference	Official	Unofficial	difference
PURE TREATMENTS						
Within-class comparison	**-7.628*** **	**-6.745*** **	[p=0.460]	**-5.153*** **	**-6.379*** **	[p=0.349]
(T1_solo)	**(1.040)**	**(1.144)**		**(0.981)**	**(1.244)**	
Across-class comparison	**-8.693*** **	**-6.408*** **	[p=0.052]	**-6.376*** **	**-6.078*** **	[p=0.817]
(T2_solo)	**(1.062)**	**(1.112)**		**(1.062)**	**(1.179)**	
Financial Rewards	0.625	1.930	[p=0.350]	2.234	0.585	[p=0.431]
(Fin_solo)	(1.404)	(1.314)		(1.474)	(1.392)	
Reputational Rewards	1.547	**2.809***	[p=0.459]	-0.064	2.052	[p=0.218]
(Rep_solo)	(1.171)	**(1.607)**		(1.181)	(1.598)	
TREATMENT INTERACTIONS						
Within-class feedback with	**-6.686*** **	**-5.818*** **	[p=0.609]	**-6.017*** **	**-4.527*** **	[p=0.463]
financial rewards (T1_fin)	**(1.393)**	**(1.855)**		**(1.900)**	**(1.632)**	
Across-class feedback with	**-7.539*** **	**-3.733*** **	[p=0.016]	**-4.502*** **	**-3.799** **	[p=0.738]
financial rewards (T2_fin)	**(1.259)**	**(1.269)**		**(1.318)**	**(1.610)**	
Within-class feedback with	**-7.059*** **	**-6.083*** **	[p=0.464]	**-6.839*** **	**-5.649*** **	[p=0.476]
reputational rewards (T1_rep)	**(1.256)**	**(1.331)**		**(1.247)**	**(1.460)**	
Across-class feedback with	**-7.752*** **	**-5.212** **	[p=0.142]	**-7.399*** **	**-6.715*** **	[p=0.673]
reputational rewards (T2_rep)	**(1.368)**	**(2.136)**		**(1.299)**	**(1.613)**	
Baseline perception	**0.237*** **	**0.257*** **		**0.261*** **	**0.269*** **	
	(0.024)	**(0.032)**		**(0.027)**	**(0.029)**	
Baseline Score	**-0.421*** **	**-0.333*** **		**-0.359*** **	**-0.426*** **	
	(0.045)	**(0.061)**		**(0.040)**	**(0.052)**	
Controlled for stratas	Yes	Yes		Yes	Yes	
Prob > F	0.0000	0.0000		0.0000	0.0000	
Number of observations	3417	3417		3376	3376	

Note: OLS. Official stands for official age of student in his/her class (i.e., all students who started primary at the age of 6 and did not repeat any class). Unofficial stands unofficial age and contain students older than official age. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Column 4 and 7 describe difference between bottom versus top performers and test results for the difference in square brackets.

* significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.7: OLS estimates of the effects of different motivation schemes on student confidence gap – by gender

Dependent variable:	Confidence Gap					
Subject:	Math			English		
Gender:	Girls	Boys	Difference (boys vs girls)	Girls	Boys	Difference (boys vs girls)
NON-INTERACTED TREATMENT EFFECTS						
Within-class comparison (T1_solo)	**-8.788*** (1.163)**	**-5.177*** (1.149)**	[p=0.008]	**-6.733*** (1.162)**	**-5.353*** (1.133)**	[p=0.341]
Across-class comparison (T2_solo)	**-9.825*** (1.014)**	**-5.119*** (1.209)**	[p=0.000]	**-7.476*** (1.083)**	**-5.128*** (1.241)**	[p=0.084]
Financial Rewards (Fin_solo)	**2.905*** (1.073)**	-0.992 (1.667)	[p=0.010]	**2.463*** (1.301)**	-0.863 (1.646)	[p=0.129]
Reputational Rewards (Rep_solo)	**2.911** (1.178)**	1.033 (1.791)	[p=0.353]	1.611 (1.142)	0.859 (1.842)	[p=0.695]
TREATMENT INTERACTIONS						
Within-class feedback with financial rewards (T1_fin)	**-7.387*** (1.420)**	**-4.727*** (1.731)**	[p=0.024]	**-5.909*** (1.354)**	**-4.194** (1.765)**	[p=0.264]
Across-class feedback with financial rewards (T2_fin)	**-6.969*** (1.168)**	**-3.893*** (2.065)**	[p=0.210]	**-5.482*** (1.154)**	**-4.053** (2.012)**	[p=0.549]
Within-class feedback with reputational rewards (T1_rep)	**-6.877*** (1.359)**	**-6.133*** (1.522)**	[p=0.681]	**-7.108*** (1.201)**	**-6.342*** (1.536)**	[p=0.669]
Across-class feedback with reputational rewards (T2_rep)	**-6.279*** (1.663)**	**-7.158*** (1.616)**	[p=0.581]	**-7.047*** (1.423)**	**-7.297*** (1.734)**	[p=0.899]
Baseline perception	**0.260*** (0.027)**	**0.229*** (0.031)**		**0.293*** (0.027)**	**0.256*** (0.030)**	
Baseline Score	**-0.386*** (0.051)**	**-0.395*** (0.050)**		**-0.348*** (0.039)**	**-0.316*** (0.044)**	
Controlled for stratas	Yes	Yes		Yes	Yes	
Prob > F	0.0000	0.0000		0.0000	0.0000	
Number of observations	1970	1441		1938	1432	

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Column 4 and 7 describe difference between bottom versus top performers and test results for the difference in square brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.8: Level differences in average perception gap, by student performance, treatment status and their perception of the task difficulty with respect to the control group

Dependent variable:	Confidence Gap					
Subject:	Math			English		
Difficulty perception:	Perceived as difficult	Perceived as easy	Difficult vs Easy	Perceived as difficult	Perceived as easy	Difficult vs Easy
PANEL A: BASELINE TESTING (ROUND 1)						
Bottom: Within-class feedback group (T1_bottom)	5.982	15.508	**[p=0.016]**	5.922	16.472	**[p=0.087]**
Top: Within-class feedback group (T1_top)	0.836	9.950	[p=0.183]	-8.874	8.848	**[p=0.042]**
Bottom: Across-class feedback group (T2_bottom)	3.616	13.625	**[p=0.079]**	4.978	13.769	[p=0.143]
Top: Across-class feedback group (T2_top)	-3.787	10.086	**[p=0.051]**	-0.082	9.936	**[p=0.022]**
Bottom: Control group	2.751	17.914	**[p=0.006]**	2.795	11.951	[p=0.155]
Top: Control group	-4.123	10.988	**[p=0.009]**	-1.696	11.874	[p=0.571]
Joint test (bottom-performers)	[p=0.589]	[p=0.275]		[p=0.447]	**[p=0.010]**	
Joint test (top-performers)	[p=0.515]	[p=0.636]		[p=0.539]	**[p=0.034]**	
Joint test (bottom vs. top)	[p=0.178]	**[p=0.022]**		[p=0.175]	**[p=0.052]**	
PANEL B: ENDLINE TESTING (ROUND 5)						
Bottom: Within-class feedback group (T1_bottom)	7.799	24.309	**[p=0.009]**	13.469	7.081	[p=0.466]
Top: Within-class feedback group (T1_top)	-3.143	8.875	[p=0.158]	-5.399	2.066	[p=0.335]
Bottom: Across-class feedback group (T2_bottom)	7.299	20.162	[p=0.109]	11.886	9.851	[p=0.802]
Top: Across-class feedback group (T2_top)	-9.551	7.493	**[p=0.067]**	-7.184	0.658	[p=0.157]
Bottom: Control group	10.983	28.725	[p=0.673]	15.802	14.393	[p=0.155]
Top: Control group	10.509	13.907	**[p=0.020]**	-0.571	7.321	[p=0.867]
Joint test (bottom vs. control)	**[p=0.006]**	[p=0.499]		[p=0.451]	**[p=0.000]**	
Joint test (top vs. control)	**[p=0.000]**	**[p=0.000]**		[p=0.549]	**[p=0.000]**	
Joint test (bottom vs. top)	**[p=0.066]**	**[p=0.000]**		**[p=0.005]**	**[p=0.000]**	

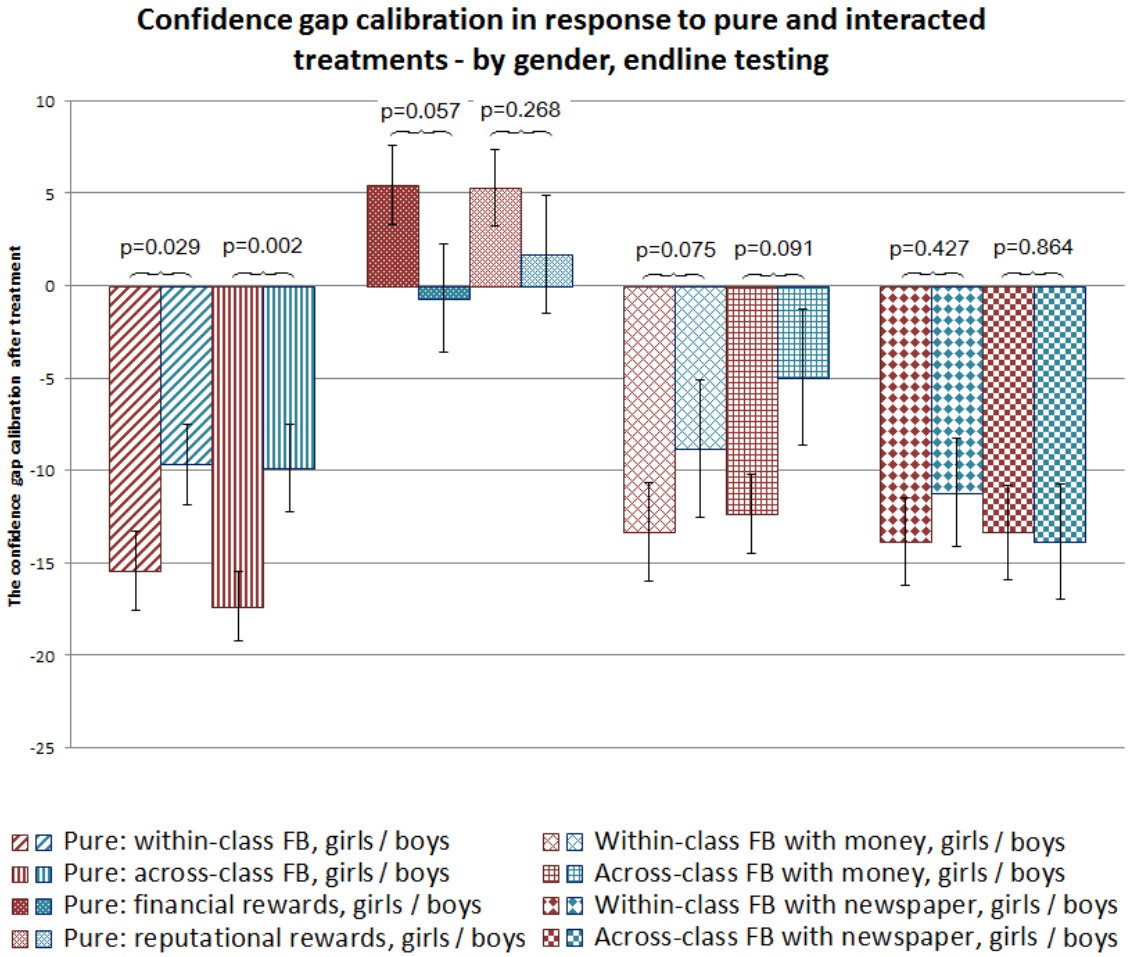
Note: Mean comparisons. Controlled for stratum fixed effects (area, level and school performance in national examinations). Column 4 and 7 show p-values from testing the differences by subject perception. Joint tests in bottom rows in each panel show p-values from testing differences between bottom and/or top and/or control group students.

Appendix C2.9: The average treatment effects – heterogeneity of the results

Dependent variable:	Confidence gap			
Specification:	(1)	(2)	(3)	(4)
Within-class feedback	**-14.05*** **	**-11.36*** **	**-14.62*** **	**-13.86*** **
	(1.215)	**(1.529)**	**(1.466)**	**(2.029)**
Across-class feedback	**-14.65*** **	**-11.68*** **	**-16.32*** **	**-15.06*** **
	(1.218)	**(1.929)**	**(1.579)**	**(1.624)**
Within-class feedback x improvement		**-0.642*** **		
		(0.084)		
Across-class feedback x improvement		**-0.711*** **		
		(0.098)		
Within-class feedback x official age			-0.476	
			(1.161)	
Across-class feedback x official age			**-2.475** **	
			(1.176)	
Within-class feedback x level of study				-2.09
				(2.252)
Across-class feedback x level of study				**-3.52* **
				(1.987)
Improvement		**-0.984*** **		
		(0.158)		
Age			**-3.951* **	
			(2.202)	
Level of study				-3.01
				(3.359)
Stratas	Yes	Yes	Yes	Yes
Initial value	Yes	No	Yes	Yes
Prob > F	0.0000	0.0000	0.0000	0.0000
Number of observations	3319	3319	3254	3319

Note: OLS. Robust standard errors adjusted for clustering at class level are in parentheses. Controlled for stratum fixed effects (area, level and school performance in national examinations). Improvement equals 1 if students improved between final and baseline testing. * significant at 10%; ** significant at 5%; *** significant at 1%

Appendix C2.10: The differences in the calibration of the confidence gap between girls and boys in response to different treatments, endline testing



Bibliography 1

Andrabi, T., Das J. and Ijaz-Khwaja, A. (2014): Report Cards: The impact of providing school and child test-scores on educational markets (October 29, 2014), HKS Working Paper No. RWP14-052.

Angrist, J., Bettinger, E., and Kremer, M. (2006): Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia, The American Economic Review, Vol. 96(3), 847-862.

Angrist, J., and Lavy, V. (2009): The effects of high stakes high school achievement awards: Evidence from a randomized trial, The American Economic Review, Vol. 99(4), 1384-1414.

Apesteguia, J., Azmat, G., and Iriberri, N. (2012): The impact of gender composition on team performance and decision-making: evidence from the field, Management Science, Vol. 58(1) January 2012, 78–93.

Arnold, H. J. (1976): Effects of performance feedback and extrinsic reward upon high intrinsic motivation, Organizational Behavior and Human Performance, Vol. 17(2), 275-288.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014): Awards unbundled: Evidence from a natural field experiment, Journal of Economic Behavior and Organization, Vol. 100, 44-63.

Auriol, E., and Renault, R. (2008): Status and incentives, The RAND Journal of Economics, Vol. 39(1), 305-326.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2015): What you know can't hurt you (for long): A field experiment on relative performance feedback, Queen Mary, University of London, Working Paper.

Azmat, G. and Iriberri, N. (2010): The importance of relative performance feedback information: Evidence from a natural experiment using high school students, Journal of Public Economics, Vol. 94(7), 435-452.

Azmat, G. and Iriberri, N. (2016): The provision of relative performance feedback: an analysis of performance and satisfaction, Journal of Economics and Management Strategy, Vol. 25(1), 77-110.

Bandiera, O., Barankay, I., and Rasul, I. (2010): Social incentives in the workplace, The Review of Economic Studies, Vol. 77(2), 417-458.

Bandiera, O., Barankay, I. and Rasul, I. (2011): Field experiments with firms, Journal of Economic Perspectives, Vol. 25(3), 63-82.

Bandiera, O., Larcinese, V., and Rasul, I. (2015): Blissful ignorance? A natural experiment on the effect of feedback on students' performance, Labour Economics, Vol. 34, 13-25.

Barankay, I. (2011): Rankings and social tournaments: Evidence from a crowd-sourcing experiment, Wharton School of Business, University of Pennsylvania Working Paper.

Benabou, R., and Tirole, J. (2003): Intrinsic and extrinsic motivation, The Review of Economic Studies, Vol. 70(3), 489-520.

Benjamini, Y., and Y. Hochberg (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society, Series B (Methodological) Vol. 57, 289–300.

Benjamini, Y., and D. Yekutieli (2001): The control of the false discovery rate in multiple testing under dependency, Annals of Statistics, Vol. 29, 1165–1188.

Besley, T., and Ghatak, M. (2008): Status incentives, The American Economic Review, Vol. 98(2), 206-211.

Bettinger, E. P. (2012): Paying to learn: The effect of financial incentives on elementary school test scores. Review of Economics and Statistics, Vol. 94(3), 686-698.

Bigoni, M., Fort, M., Nardotto, M., and Reggiani, T. (2011): Teams or tournaments? A field experiment on cooperation and competition among university students, IZA Discussion Paper Series No. 5844.

Blanes i Vidal, J., and Nossol, M. (2011): Tournaments without prizes: Evidence from personnel records, Management Science, Vol. 57(10), 1721-1736.

Blimpo, M. P. (2014): Team incentives for education in developing countries: A randomized field experiment in Benin, American Economic Journal: Applied Economics, Vol. 6(4), 90-109.

Buunk, B. P., Gibbons, F. X., and Reis-Bergan, M. (1997): Social comparison in health and illness: A historical overview. In Buunk, B.P., and Gibbons, F.X. (Eds): Health, coping and well-being: Perspectives from social comparison theory, Psychology Press, 1-23.

Buunk, B. P., and Gibbons, F. X. (2000): Toward an enlightenment in social comparison theory: Moving beyond classic and Renaissance approaches. In Suls, J., and Wheeler, L. (Eds): Handbook of Social Comparison: Theory and Research, The Nederlands: Kluwer Academic, 487-499.

Burgers, C., Eden, A., van Engelenburg, M. D., and Buningh, S. (2015): How feedback boosts motivation and play in a brain-training game, Computers in Human Behavior, Vol. 48, 94-103.

Charness, G., Masclet, D., and Villeval, M. C. (2010): Competitive preferences and status as an incentive: Experimental evidence, Groupe d'Analyse et de Théorie Économique Working Paper No. 1016.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983): A global measure of perceived stress, Journal of Health and Social Behavior, Vol. 24(4), 385-396.

Croson, R., and Gneezy, U. (2009): Gender differences in preferences, Journal of Economic Literature, Vol. 47(2), 448-474.

Deci, E. L. (1971): Effects of externally mediated rewards on intrinsic motivation, Journal of Personality and Social Psychology, Vol. 18(1), 105-115.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999): A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation, Psychological Bulletin, Vol. 125(6), 627-668.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A.P. and van der Zee, Y. (2008): Social comparison in the classroom: a review, Review of Educational Research, Vol. 78(4), 828-879.

Dolan, P., Metcalfe, R., and Powdthavee, N. (2008): Electing happiness: Does happiness affect voting and do elections affect happiness, Discussion Papers in Economics No. 2008/30.

Duffy, J., and Kornienko, T. (2010): Does competition affect giving?, Journal of Economic Behavior and Organization, Vol. 74(1), 82-103.

Dunn, O. J. (1961): Multiple comparisons among means, Journal of the American Statistical Association, Vol. 56(293), 52–64.

Dynarski, S. (2008): Building the stock of college-educated labor, Journal of Human Resources, Vol. 43(3), 576-610.

Eisenkopf, G. (2011): Paying for better test scores, Education Economics, Vol. 19(4), 329-339.

Ellingsen, T., and Johannesson, M. (2007): Paying respect, Journal of Economic Perspectives, Vol. 21(4), 135-150.

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009): Feedback and incentives, Experimental evidence, Labour Economics, Vol. 16(6), 679-688.

Falk, A. and Ichino, A. (2006): Clean Evidence on Peer Pressure, Journal of Labor Economics, Vol. 24(1), 39-57.

Festinger, L. (1954): A theory of social comparison processes, Human Relations, Vol. 7(2), 117-140.

Fordyce, M. W. (1988): A review of research on the happiness measures: A sixty second index of happiness and mental health, Social Indicators Research, Vol. 20(4), 355-381.

Fryer Jr, R. G. (2010): Financial incentives and student achievement: Evidence from randomized trials, National Bureau of Economic Research, No. 15898.

Hannan, R. L., Krishnan, R., and Newman, A. H. (2008): The effects of disseminating relative performance feedback in tournament and individual performance compensation plans, The Accounting Review, Vol. 83(4), 893-913.

Hanushek, E. A., and Woessmann, L. (2007): The role of education quality for economic growth, World Bank Policy Research Working Paper No. 4122

Hanushek, E.A. (2005): Why quality matters in education, Finance & Development, Vol. 42(2).

Hastings, J. S., Neilson, C. A., and Zimmerman, S. D. (2012): The effect of school choice on intrinsic motivation and academic outcomes, National Bureau of Economic Research No. 18324.

Hattie, J., and Timperley, H. (2007): The power of feedback, Review of Educational Research, Vol. 77(1), 81-112.

Helliwell, J. F., and Wang, S. (2012): The state of world happiness, World Happiness Report, 10-57.

Hirano, K., Imbens, G. and Ridder, G. (2003): Efficient estimation of average treatment effects using the estimated propensity score, Econometrica, Vol. 71(4), 1161–1189.

Hochberg, Y. (1988): A sharper Bonferroni procedure for multiple tests of significance, Biometrika, Vol. 75, 800–802.

Hoff, K. and Pandey, P. (2006): Discrimination, social identity, and durable inequalities, American Economic Review, Vol. 96(2), 206-211.

Holland, B. S., and Copenhaver. M.D. (1987): An improved sequentially rejective Bonferroni test procedure, Biometrics, Vol. 43, 417–423.

Holm, S. (1979): A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics, Vol. 6, 65–70.

Hoxby, C. (2000): Peer effects in the classroom: Learning from gender and race variation, National Bureau of Economic Research No. 7867.

Imbens, G.W. (2004): Nonparametric estimation of average treatment effects under exogeneity: A review, The Review of Economics and Statistics: Vol. 86(1), 4-29.

Jalava, N., Joensen, J.S. and Pellas, E. (2015): Grades and rank: Impacts of non-financial incentiveson test performance, Journal of Economic Behavior and Organization, Vol. 115, 161–196.

Juster, R. P., McEwen, B. S., and Lupien, S. J. (2010): Allostatic load biomarkers of chronic stress and impact on health and cognition, Neuroscience and Biobehavioral Reviews, Vol. 35(1), 2-16.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007): Experimental analysis of neighborhood effects, Econometrica, Vol. 75(1), 83-119.

Kosfeld, M. and Neckermann, S. (2011): Getting more work for nothing? Symbolic awards and worker performance, American Economic Journal: Microeconomics, Vol. 3(3), 86-99.

Kremer, M., Miguel, E. and Thornton, R. (2002): Incentives to Learn, National Bureau of Economic Research No. 10971.

Krueger, A. B. (1999): Experimental estimates of education production functions, The Quarterly Journal of Economics, Vol. 114(2), 497-532.

Kuhnen, C. M., and Tymula, A. (2012): Feedback, self-esteem, and performance in organizations, Management Science, Vol. 58(1), 94-113.

Kwak, D. (2010): Inverse probability weighted estimation for the effect of kindergarten enrollment age and peer quality on student academic achievement for grades K-12, Michigan State University, Working Paper.

LaLonde, R. J. (1995): The promise of public sector-sponsored training programs, Journal of Economic Perspectives, Vol. 9(2), 149-168.

Lavy, V. (2009): Performance pay and teachers' effort, productivity and grading ethics, American Economic Review, Vol. 99(5), 1979-2011.

Lazear, E.P. (2000): Performance pay and productivity, American Economic Review, Vol. 90(5), 1346-1361.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2012): The behavioralist goes to school: Leveraging behavioral economics to improve educational performance, National Bureau of Economic Research No. 18165.

Locke, E. A., and Latham, G. P. (1990): A theory of goal setting and task performance, Englewood Cliffs, NJ: Prentice-Hall, Inc., 413.

Lupien, S. J., McEwen, B. S., Gunnar, M. R., and Heim, C. (2009): Effects of stress throughout the lifespan on the brain, behavior and cognition, Nature Reviews Neuroscience, Vol. 10(6), 434-445.

Lyubomirsky, S., and Lepper, H. (1999): A measure of subjective happiness: Preliminary reliability and construct validation, Social Indicators Research, Vol. 46, 137-155.

McEwen, B. S. (2008): Central effects of stress hormones in health and disease: Understanding the protective and damaging effects of stress and stress mediators, European Journal of Pharmacology, Vol. 583(2), 174-185.

MacKerron, G. (2012): Happiness economics from 35 000 feet, Journal of Economic Surveys, Vol. 26(4), 705-735.

Markham, S. E., Scott, K., and McKee, G. A. I. L. (2002): Recognizing good attendance: a longitudinal, quasi-experimental field study. Personnel Psychology, Vol. 55(3), 639-660.

Mas, A. and Moretti, E. (2009): Peers at Work, American Economic Review, Vol. 99(1), 112-145.

Mettee, D. R., and Smith, G. (1977): Social comparison and interpersonal attraction: The case for dissimilarity. In J. M. Suls and R. L. Miller (Eds.), Social Comparison Processes: Theoretical and Empirical Perspectives, 69-101.

Moldovanu, B., Sela, A., and Shi, X. (2007): Contests for status, Journal of Political Economy, Vol. 115(2), 338-363.

Muzatti, B., and Agnoli, F. (2007): Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children, Developmental Psychology, Vol 43(3), 747-759.

Newson, R.B. (2010): Frequentist q-value for multiple-test procedures, Stata Journal, Vol. 10(4), 568-584.

Raudenbush, S. W., et al. (2011): Optimal Design Software for Multi-level and Longitudinal Research, Version 3.01.

Ray, D. (2002): Aspirations, poverty and economic change. In Banerjee, A.V. (2006): Understanding Poverty, 409-421.

Reardon, S. F., Cheadle, J. E., and Robinson, J. P. (2009): The effect of Catholic schooling on math and reading development in kindergarten through fifth grade, Journal of Research on Educational Effectiveness, Vol. 2(1), 45-87.

Ryan, R. M., and Deci, E. L. (2000): Intrinsic and extrinsic motivations: Classic definitions and new directions, Contemporary Educational Psychology, Vol. 25(1), 54-67.

Sacerdote, B. (2011): Chapter 4 - Peer effects in education: How might they work, how big are they and how much do we know thus far?, Handbook of the Economics of Education, Vol. 3, 249-277.

Schneiderman, N., Ironson, G., and Siegel, S. D. (2005): Stress and health: psychological, behavioral, and biological determinants, Annual Review of Clinical Psychology, Vol. 1, 607-628.

Šidák, Z. (1967): Rectangular confidence regions for the means of multivariate normal distributions, Journal of the American Statistical Association, Vol. 62, 626–633.

Simes, R. J. (1986): An improved Bonferroni procedure for multiple tests of significance, Biometrika, Vol. 73, 751–754.

Slavin, R. (1984). Meta-analysis in education: How has it been used?, Educational Researcher, Vol. 13(8), 6-15.

Suls, J., and Wheeler, L. (2000): A selective history of classic and neo-social comparison theory, Handbook of social comparison, The Springer Series in Social Clinic Psychology, 3-19.

Tran, A., and Zeckhauser, R. (2012): Rank as an inherent incentive: Evidence from a field experiment, Journal of Public Economics, Vol. 96(9), 645-650.

Van Dijk, F., Sonnemans, J., and Van Winden, F. (2001): Incentive systems in a real effort experiment, European Economic Review, Vol. 45(2), 187-214.

Veenhoven, R. (1988): The utility of happiness, Social Indicators Research, Vol. 20(4), 333-354.

Weiss, Y., and Fershtman, C. (1998): Social status and economic performance: A survey, European Economic Review, Vol. 42(3), 801-820.

Wolf, T. M. (1994): Stress, coping and health: enhancing well-being during medical school, Medical Education, Vol. 28(1), 8-17.

Wooldridge, J. (2007): Inverse Probability Weighted M-Estimation for General Missing Data Problems, Journal of Econometrics, Vol. 141(2), 1281-1301.

Bibliography 2

Aberson, C. L., Healy, M., and Romero, V. (2000): Ingroup bias and self-esteem: A meta-analysis, Personality and Social Psychology Review, Vol. 4(2), 157-173.

Ackerman, P. L., Beier, M. E., and Bowen, K. R. (2002): What we really know about our abilities and our knowledge, Personality and Individual Differences, Vol. 33(4), 587-605.

Akerlof, G. A., and Kranton, R. E. (2000): Economics and identity, Quarterly Journal of Economics, Vol. 115(3), 715-753.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2015): What you know can't hurt you (for long): A field experiment on relative performance feedback in higher education, Queen Mary, University of London, Working Paper.

Barber, B. M., and Odean, T. (2001): Boys will be boys: Gender, overconfidence, and common stock investment, Quarterly Journal of Economics, Vol. 116(1), 261-292.

Barnett, A. G., van der Pols, J. C., and Dobson, A. J. (2005): Regression to the mean: what it is and how to deal with it, International Journal of Epidemiology, Vol. 34(1), 215-220.

Benabou, R., and Tirole, J. (2002): Self-confidence and personal motivation, The Quarterly Journal of Economics, Vol. 117(3), 871-961.

Benoît, J. P., and Dubra, J. (2011): Apparent overconfidence, Econometrica, Vol. 79(5), 1591-1625.

Burson, K. A., Larrick, R. P., and Klayman, J. (2006): Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons, Journal of Personality and Social Psychology, Vol. 90(1), 60-77.

Cacault, M. P., and Grieder, M. (2016): How group identification distorts beliefs, SSRN Working Paper.

Camerer, C., and Lovallo, D. (1999): Overconfidence and excess entry: An experimental approach, The American Economic Review, Vol. 89(1), 306-318.

Charness, G., Kuhn, P., and Villeval, M. C. (2010): Competition and the Ratchet effect, National Bureau of Economic Research No. 16325.

Clayson, D. E. (2005): Performance overconfidence: metacognitive effects or misplaced student expectations?. Journal of Marketing Education, Vol. 27(2), 122-129.

Cohen, J. (1988): Statistical Power Analysis for the Behavioral Sciences, Vol. 2, Lawrence Earlbaum Associates Hillside, NJ.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983): A global measure of perceived stress, Journal of health and social behavior, Vol. 24(4), 385-396.

Crocker, J., and Luhtanen, R. (1990): Collective self-esteem and ingroup bias, Journal of Personality and Social Psychology, Vol. 58(1), 60-67.

Cross, K. P. (1977): Not can, but will college teaching be improved?, New Directions for Higher Education, Vol. 1997(17), 1-15.

Duffy, J., and Hopkins, E. (2005): Learning, information, and sorting in market entry games: theory and evidence, Games and Economic Behavior, Vol. 51(1), 31-62.

Dunlosky, J., and Rawson, K. A. (2012): Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention, Learning and Instruction, Vol. 22(4), 271-280.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003): Why people fail to recognize their own incompetence, Current Directions in Psychological Science, Vol. 12(3), 83-87.

Edwards, R. K., Kellner, K. R., Sistrom, C. L., and Magyari, E. J. (2003): Medical student self-assessment of performance on an obstetrics and gynecology clerkship, American Journal of Obstetrics and Gynecology, Vol. 188(4), 1078-1082.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., and Kruger, J. (2008): Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent, Organizational Behavior and Human Decision Processes, Vol. 105(1), 98-121.

Engelmann, D., and Strobel, M. (2000): The false consensus effect disappears if representative information and monetary incentives are given, Experimental Economics, Vol. 3(3), 241-260.

Gino, F., and Moore, D. A. (2007): Effects of task difficulty on use of advice, Journal of Behavioral Decision Making, Vol. 20(1), 21-35.

Gneezy, U., Niederle, M., and Rustichini, A. (2003): Performance in competitive environments: Gender differences, Quarterly Journal of Economics-Cambridge Massachusetts, Vol. 118(3), 1049-1074.

Gneezy, U., and Rustichini, A. (2004): Gender and competition at a young age, The American Economic Review, Vol. 94(2), 377-381.

Hacker, D. J., Bol, L., and Bahbahani, K. (2008): Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style, Metacognition and Learning, Vol. 3(2), 101-121.

Hacker, D. J., Bol, L., Horgan, D. D., and Rakow, E. A. (2000): Test prediction and performance in a classroom context, Journal of Educational Psychology, Vol. 92(1), 160.

Haun, D. E., Zeringue, A., Leach, A., and Foley, A. (2000): Assessing the competence of specimen-processing personnel, Laboratory Medicine, Vol. 31(11), 633-637.

Healy, A., and Offenberg, J. G. (2007): Overconfidence, social groups, and gender: Evidence from the lab and field, SSRN Working Paper.

Healy, A., and Pate, J. (2011): Can teams help to close the gender competition gap?, The Economic Journal, Vol. 121(555), 1192-1204.

Hewstone, M., Rubin, M., and Willis, H. (2002): Intergroup bias, Annual Review of Psychology, Vol. 53(1), 575-604.

Hodges, B., Regehr, G., and Martin, D. (2001): Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it, Academic Medicine, Vol. 76(10), S87-S89.

Howard, M. E. (1983): The causes of war and other essays, Cambridge, MA: Harvard University Press.

Kennedy, E. J., Lawton, L., and Plumlee, E. L. (2002): Blissful ignorance: The problem of unrecognized incompetence and academic performance, Journal of Marketing Education, Vol. 24(3), 243-252.

Köszegi, B. (2006): Ego utility, overconfidence, and task choice, Journal of the European Economic Association, Vol. 4(4), 673-707.

Krajč, M. (2008): Are the unskilled really that unaware? Understanding seemingly biased self-assessments, CERGE-EI Working Paper Series No. 373.

Krajč, M., and Ortmann, A. (2008): Are the unskilled really that unaware?, An alternative explanation, Journal of Economic Psychology, Vol. 29(5), 724-738.

Krueger, J., and Mueller, R. A. (2002): Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predicts errors in estimates of own performance, Journal of Personality and Social Psychology, Vol. 82(2), 180-188.

Kruger, J., and Dunning, D. (1999): Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments, Journal of Personality and Social Psychology, Vol. 77(6), 1121-1134.

Kruger, J., and Dunning, D. (2002): Unskilled and unaware - but why? A reply to Krueger and Mueller (2002), Journal of Personality and Social Psychology, Vol. 82(2), 189-192.

Kuhn, P., and Villeval, M. C. (2015): Are Women More Attracted to Co-operation Than Men?, The Economic Journal, Vol. 125(582), 115-140.

Larrick, R. P., Burson, K. A., and Soll, J. B. (2007): Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not), Organizational Behavior and Human Decision Processes, Vol. 102(1), 76-94.

Levy, J. S. (1983): Misperception and the causes of war: Theoretical linkages and analytical problems, World Politics, Vol. 36(01), 76-99.

Lipko, A.R., Dunlosky, J. and Merriman, W.E. (2009): Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions, Journal of Experimental Child Psychology, Vol. 103(2), 152-166.

Lyubomirsky, S., and Lepper, H. S. (1999): A measure of subjective happiness: Preliminary reliability and construct validation, Social Indicators Research, Vol. 46(2), 137-155.

Mahajan, J. (1992): The overconfidence effect in marketing management predictions, Journal of Marketing Research, Vol. 29(3), 329-342.

Marottoli, R. A., Richardson, E. D., Stowe, M. H., Miller, E. G., Brass, L. M., Cooney, L. M., and Tinetti, M. E. (1998): Development of a test battery to identify older drivers at risk for self-reported adverse driving events, Journal of the American Geriatrics Society, Vol. 46(5), 562-568.

Miller, T. M., and Geraci, L. (2011): Unskilled but aware: reinterpreting overconfidence in low-performing students, Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol. 37(2), 502-506.

Moore, D. A., and Cain, D. M. (2007): Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition, Organizational Behavior and Human Decision Processes, Vol. 103(2), 197-213.

Moore, D. A., and Healy, P. J. (2008): The trouble with overconfidence, Psychological Review, Vol. 115(2), 502-517.

Moore, D. A., and Kim, T. G. (2003): Myopic social prediction and the solo comparison effect, Journal of Personality and Social Psychology, Vol. 85(6), 1121-1135.

Niederle, M., and Vesterlund, L. (2007): Do women shy away from competition? Do men compete too much?, The Quarterly Journal of Economics, Vol. 122(3), 1067-1101.

Rullière, J. L., Pinto, L. S., and Vialle, I. (2011): Self-confidence and teamwork: An experimental test, Gate Working Paper No. 1126.

Ryvkin, D., Krajč, M. and Ortmann, A., (2012): Are the unskilled doomed to remain unaware?, Journal of Economic Psychology, Vol. 33(5), 1012-1031.

Santos-Pinto, L., and Sobel, J. (2005): A model of positive self-image in subjective assessments, The American Economic Review, Vol. 95(5), 1386-1402.

Odean, T. (1998): Volume, volatility, price, and profit when all traders are above average, Journal of Finance, Vol. 53(6), 1887-1934.

Winship, C., and Morgan, S. L. (1999): The estimation of causal effects from observational data, Annual Review of Sociology, Vol. 25, 659-706.

Zenger, T. R. (1992): Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover, Administrative Science Quarterly, Vol. 37(2), 198-219.