Do incentives improve test scores? New evidence from a field experiment

Uwe Dulleck, Martin Kocher, Jayanta Sarkar[*], Dipanwita Sarkar[†]

Preliminary version (please do not cite without permission)

Abstract

Recent research shows student effort in the ‘production function’ of test scores is sub-optimal (Levitt *et al*, 2012), and that provision of incentives to enhance academic performance could be a potentially cost-effective strategy. However, not much is known about the ‘optimal’ structure of incentives, and whether incentives work under real test scenarios. In this paper, we examine the relative efficacy of a battery of individual and group-based incentives during NAPLAN tests at a high school in Queensland. The results suggest gains in performance across all incentive treatments for Year 7 students, and for some treatments for Year 9 students. Among all treatment, we find group-based incentives was most effective for Year 7 students, which suggests a strong role of ‘social incentives’. Furthermore, we find heterogeneity of gains in test scores across ability distribution – the high-ability students in Year 7 and low-ability students in Year 9 demonstrated highest gains under the group-based incentives. The insights from this field experiment are not only important from an educational policy perspective, but have wider implications for any organisation trying to increase effort-levels of its employees.

Key words: Test scores, Social incentive, Field experiment

JEL codes: I20, I21, C93

[*] Correspondence: School of Economics and Finance, QUT Business School, Queensland University of Technology, Gardens Point Campus, Brisbane, Queensland 4000. email: jayanta.sarkar@qut.edu.au; phone: +61 7 31384252; Fax: +61 7 3138 1500.

[†] Uwe Dulleck and Dipanwita Sarkar are at Queensland University of Technology, and Martin Kocher is at University of Munich.

1. Introduction

Improving academic performance of school children has always been an agenda of critical importance for researchers and policymakers throughout the world. In recent time, the ubiquity of the relative performance indicators in international assessments, such as the Program for International Student Assessment (PISA), have catalyzed the process of educational reforms in many countries to push this agenda further.[1] School systems in some countries, however, are in urgent need of reform.

Educational policies and reforms are often crafted to improve the effort and engagement levels of students to maximize educational outcomes. To many it may sound surprising that students, who are the main beneficiary of education will have to be coaxed and prodded to put in their best efforts when it is in their best interest to do so anyway. This, unfortunately, is not the case for many students. Recent research informs that student effort in the 'production function' of test scores is sub-optimal (Levitt *et al*, 2012). For example, performance of students in countries like the US and Australia have registered sharp decline in international assessments in recent years, despite a plethora of reforms such as teacher-incentives, reduced class-size, gender segregation, etc.

Recently, provision of incentives to enhance academic performance has gained popularity as a potentially cost-effective strategy. The argument is that short-term financial and non-financial incentives that is tied to a student's academic performance will elicit the desired levels of engagement and effort that correlate with high test scores. These incentives may increase performance when underperformance is due to lack of motivation, heavy discounting of the returns to schooling, or lack of accurate information on the overall benefits of schooling.[2] Most studies focus on private incentives that are tied to individual performance (Barrow et al., 2014; Bettinger, 2012; Fryer, 2011; Angrist et al., 2009; Kremer et al., 2009). These incentives can take the following forms: First, these rewards are earned whenever an individual performs better than a threshold level, regardless of the extent of improvement. This type of incentive scheme could be more appealing to the students who would otherwise perform just below the threshold. Second,

[1] For example, Germany completely revamped their traditional education system, *Gymnasium*, in response to PISA reports in 2008, United Kingdom extended the compulsory school leaving age by one year; since 2001 the United States has implemented the *No Child Left Behind* Act.

[2] On the other hand, if underperformance is due to lack of resources or due to external barriers (e.g., effective teachers, motivated students, engaged parents, or peer dynamics), then incentives may have little impact.

incentives could be tied to the ‘extent’ of improvement – that is rewards increase with level of performance above the threshold, which could induce students performing above the threshold to try harder than just satisficing. Thus, the rewards based on ‘absolute’ and ‘marginal’ performance generate incentives for individual students of all types – both low and high performers. But, these individual schemes may have limited effects because they are unable to take advantage of the peer-dynamics. Insights from the literature on peer-effects suggest that social interactions and social pressure play a significant role in altering behaviour and performance. Therefore, one pertinent question is whether we can elicit greater effort from students by making the incentives more potent by using group-based incentives. A growing strand of literature highlights the role of peer-effects – that is how social interaction or social pressure can influence behaviour and action of members in a peer group.[3] The objective of this paper is to investigate the effectiveness of group-based incentives relative to individual incentives among school students in a real test setting.

To this end, we conduct an experiment during a nation-wide mandated test in a public high-school in Australia. In a first field experiment of its kind, we use a within-subjects design to implement three treatments, two of which offer individual incentives and one offers group-based incentive. The treatments are evaluated against a non-incentivized baseline scenario to identify the relative effects of these incentives on student test scores in a nationwide academic achievement test.

We find improvements among year 7 students in all subject areas (language convention, writing, reading, numeracy). Against the baseline area of language convention, we find significant within-subject gains over the last test performance (in year 5) in reading, writing, and numeracy test performances for grade 7 students. The achievement was most pronounced in numeracy – that is, under the group-incentive scheme, and lowest in writing – under the absolute incentive scheme.

The grade 9 students responded to the incentives differently. The incentives based on marginal gain were effective in improving grades, while their response to group-based incentives were smaller in magnitude compared to the grade 7 students.

An important issue pertains to the distribution of the gains from the incentive program. More specifically, from a policy perspective it is important to know whether there is heterogeneity of the treatment effects across ability of the students. To this end, we incorporated student

[3] See Sacerdote (2001), Zimmerman (2003), Bandiera et al. (2005, 2010, 2013), Falk and Ichino (2005), Foster (2006), Lyle (2007), Kremer and Levy (2008), Carrell et al. (2009), Mas and Moretti (2009), and Carrell et al. (2011).

performances in prior diagnostic tests as indicators of ability in Reading and Numeracy conducted by the school during the year immediately prior to our experiment. After controlling for other potential influences (such as gender, forms in a grade reflecting the characteristics of the form-teacher) we find that the gains were not equal across the students of different abilities. The results of quantile regression indicate that for year 7, the highest gains were concentrated among the high-ability students across the three treatments. However, in year 9, the low-ability students demonstrated higher gains, and these gains were only achieved under the group-based incentive scheme.

The field experiment took place in a natural setting at a high school. More importantly, student performance was measured in a nationwide test that each student is required to take at various stages of his/her school years. Unlike many experimental studies, these tests were created and administered completely independently of the experiment. Hence, our results can be safely generalized in real world.

The overall superiority of group-based incentives has wider significance for policy. Social influence seems to play a strong role in eliciting greater efforts from individual group members. This is because the individual incentives that are no longer independent – they are linked. The feedback between these links generate a social multiplier effect, which is greater in magnitude than the sum of the individual effects. Group-based policies are ubiquitous in real-world – for example, team-based incentives are commonly given in firms, the military, sports, health and wellness programs, and even in academic institutions. The additional improvement in individual performance can stem from positive externalities of learning from each other, and also from fear of letting down the other group-members (Lencioni, 2002). Our results suggest that incentive structure that take advantage of the peer-effects can be quite effective in eliciting student efforts in school.

2. Experimental design

The Australian school curriculum mandates public schools to conduct the National Assessment Program – Literacy and Numeracy, popularly known as the NAPLAN tests in a view to measure progress made at the school level against national standards.[4] These tests are conducted at every

[4] The purpose of the NAPLAN tests is to assess student knowledge and skills in numeracy, reading, writing, spelling, punctuation and grammar. NAPLAN tests are complementary to the various formal and informal testing programs

public school for students at grades 3, 5, and 7. Our research team was consulted by a high-school in the state of Queensland, Australia to help them design an incentive program to improve outcomes in the NAPLAN tests conducted in 2016. In return, we used the de-identified data made available by the school for this research. The target intervention groups were students in grades 7 and 9.

The ultimate goal of our study is to examine the efficacy of various incentive structures to elicit higher levels of perseverance and determination to do well in a real test scenario. Therefore, the experiment is designed to eliminate the role of efforts and expectations – both from students and teachers - going into the tests. Following Levitt et al. (2012), we offer students a reward for performance that is announced immediately prior to an incentivised test. This allows us to isolate the confounders due to discount rates, opportunity costs, planning failures, or human capital accumulation (e.g., studying for the test).

The timeline for the payment of incentives was constrained by the real timeline of the test and the announcement of the results in Queensland. Therefore, unlike Levitt et al. (2012) we only measure the effects of longer-term incentives that are payable only with a significant time lag of 6 weeks.

A week before the NAPLAN tests these target students were provided with the results of their previous NAPLAN test, and were told by the principal to improve on it. The school also computed their target score adjusting for the natural gain over two years – 40 points in each subject area – that can be attributed to increasing maturity, ability, and experience with NAPLAN tests. Therefore, a score of, say, 410/500 in Writing in grade 5 would be equivalent to 450/500 in Year 7; a score exceeding 450 would be considered as an improvement over previous performance. Therefore, the upper bound imposed by the maximum attainable score (here 500) makes outperformance harder for better students, and impossible for students who scored 470 or more in the same subject area in their previous test. Therefore, the empirical estimates of the effects of the incentives presented here are biased downward.

The particular NAPLAN tests took place over three days – 10 -12 May, 2016, testing students in four subject areas: Language Conventions (May 10), Writing (May 10), Reading (May 11), and Numeracy (May 12). The incentive structure comprised of the following control and treatment

that the schools regularly conduct. The results of the tests are viewed as point-in-time indicators of student progress and achievement across Australian schools. Performance in these tests enable schools to monitor academic growth over time for each student, and inform future planning and curriculum development.

interventions. Each control and treatment was administered to a single NAPLAN subject area. Students were given a printout of their last NAPLAN result, with their position as a “dot”, their actual score, along with the score-bands in which they belonged in each of the test subject areas.

Ability or quality of students were measured using performance in Reading and Numeracy tests conducted in 2015 as part of the Progressive Achievement Test (PAT) conducted by the school. We used the PAT performance for 2015 to obtain the most recent available objective measure of ability. However, of not all students who took the NAPLAN tests in 2016 took part in the PAT in 2015 (summary statistics to be provided).

2.1 Baseline

The Language Convention (LC) test on May 10 formed the control group (baseline treatment) for both grade 7 and 9 cohorts. Just before the LC test, a pre-recorded video message from the school Principal was shown to the grade 9 (grade 7) students in the examination room, urging them to improve on their last NAPLAN score in the LC test. The message was as follows:

> To grade 9 students: “*You are about to take the NAPLAN LANGUAGE CONVENTIONS test. You also took this test in Year 7, the results of which were given to you last week. Please try to improve upon your score in this area*. *All the very best!*”
>
> To grade 7 students: “*You are about to take the NAPLAN LANGUAGE CONVENTIONS test. You also took this test in Year 5, the results of which were given to you last week. Please try to improve upon your score in this area*. *All the very best!*”

These messages were replayed twice and the invigilators ensured that everybody in the examination rooms paid attention and that the messages were understood by every student.

2.2 Treatments

There were three treatments in total, each designed to incentivize student performance differently. These can be described as follows. In each treatment, the corresponding statement was repeated at least twice, or until it was clear that everybody understood the nature of the incentive.

Treatment 1 (Absolute gain): This treatment aimed to provide rewards based on absolute gain, where amount of reward is not sensitive to performance beyond a certain level. The *WRITING* assessment was chosen for this treatment. The following pre-recorded video message by the school

Principal was played in the examination rooms just before the *WRITING* test (held after LC test on May 10) for grade 9 (grade 7) students:

> To grade 9 students: "*You are about to take the NAPLAN WRITING test. You also took this test in Year 7, the results of which were given to you last week along with a target score for WRITING. If your score today is higher than your target score - that is, your WRITING score from Year 7 + 40, you will receive $20. You will be paid within a week after the NAPLAN results are announced. All the very best!*"
>
> To grade 7 students: "*You are about to take the NAPLAN WRITING test. You also took this test in Year 5, the results of which were given to you last week along with a target score for WRITING. If your score today is higher than your target score - that is, your WRITING score from Year 5 + 40, you will receive $20. You will be paid within a week after the NAPLAN results are announced. All the very best!*"

Treatment 2 (Marginal gain): This treatment aimed to measure higher marginal effort and provided equal incentives to all students irrespective of how they performed in the past test. The payment amounts are made comparable to the other treatments. The *READING* assessment was chosen for this treatment. The following pre-recorded video message by the school Principal was played in the examination rooms just before the *READING* test (held on May 11) for grade 9 (grade 7) students:

> To grade 9 students: "*You are about to take the NAPLAN READING test. You also took this test in Year 7, the results of which were given to you last week along with a target score for READING. You will receive $4 for every percentage point increase over your target score – that is, your READING score from Year 7 + 40. The maximum payment you can receive will be $20 (up to 5 percentage point increase in your score). You will be paid within a week after the NAPLAN results are announced. All the very best!*"
>
> To grade 7 students: "*You are about to take the NAPLAN READING test. You also took this test in Year 5, the results of which were given to you last week along with a target score for READING. You will receive $4 for every percentage point increase over your target score – that is, your READING score*

> *from Year 5 + 40. The maximum payment you can receive will be $20 (up to 5 percentage point increase in your score). You will be paid within a week after the NAPLAN results are announced. All the very best!*"

Treatment 3 (Group incentive): This treatment was designed to measure the effectiveness of group-based incentive and compare its effects with those of the individual incentives. The *NUMERACY* assessment was chosen for this treatment. The following pre-recorded video message by the school Principal was played in the examination rooms just before the *NUMERACY* test (held on May 12) for grade 9 (grade 7) students:

> To both grade 9 and grade 7 students: "*You are about to take the NAPLAN NUMERACY test. If your class has the highest average score in NUMERACY this year within the school, each one of you will receive $20. You will be paid within a week after the NAPLAN results are announced*."

The NAPLAN results were announced 6 weeks after the last (Numeracy) test. The students received their rewards within 7 days of the announcement of the NAPLAN results.

3. Conceptual framework

To fix ideas, we develop a simple theoretical framework to understand the relative costs and benefits associated with the various incentive schemes. We then form hypotheses about the relative attractiveness of the incentive schemes from an individual student's perspective.

Let S_i be the expected score and T_i be the adjusted target score in a NAPLAN test area for a student i, who earns a reward as long as $S_i - T_i > 0$ under incentive schemes that are ties to individual performance – that is, under Treatments 1 and 2. Let V_i be the intrinsic utility from achieving $S_i - T_i > 0$. Let $B > 0$ be the utility from the reward and $C_i > 0$ be the utility cost of effort associated with per unit of ($S_i - T_i$). The value of B = $5 is fixed in the Treatment 2 (marginal gain). The value of reward in Treatment 1 (Absolute gain) is $20, or 4$B$. Given this, the net utility from Treatment 1 is thus given by

$$U_{i,A} = V_i + 4B - C_i, \text{ given } (S_i - T_i\ ) > 0 \qquad (1)$$

A student responds to Treatment 1 if $U_{i,A} > 0$.

The net utility in Treatment 2 is given by:

$$U_{i,M} = V_i + (S_i - T_i)(B - C_i), \text{ given } (S_i - T_i\ ) > 0 \qquad (2)$$

A student responds to Treatment 2 if $U_{i,M} > 0$, a sufficient condition for which is $B - C_i > 0$.

The utility from Treatment 3 is modelled differently as it involves group-performance based incentive, unlike (1) and (2). A student's intrinsic utility from achieving the target (doing the best s(he) can) is still V_i, the same as in (1) and (2). The costs and benefits are, however, different. Let the individual utility from the group reward is B_G, the associated individual effort-cost is C_G. However, in addition to the individual utility from reward, a group member also gains utility from the event that everybody else in his/her group will get the same reward too. Let $0 < p < 1$ be student i's subjective probability of the event that everybody else in his/her group will be trying to achieve the common goal (achieving the highest average score in Numeracy test in their grade), and let $\gamma > 0$ be the utility of student i from everyone in the group earning B_G. Thus, the net utility under Treatment 3 is given by:

$$U_{i,G} = V_i + pB_G - C_{i,G} + \gamma pB_G \qquad (3)$$

A student responds to Treatment 3 if $U_{i,G} > 0$, a sufficient condition for which is $(1+\gamma)pB_G - C_{i,G} > 0$.

From (1) – (3), we are able to formalize the relative attractiveness of Treatments 1 – 3 from a student's perspective. Given that the net utility of each Treatment depends on personal cost of effort relative to the expected benefits from rewards, as well as expected scores, we first classify students in terms of the magnitudes of these values: We classify students into the following three categories:

'Low' ability: For these students expected score is low, such that $(S_i - T_i - 1) < 0$. In addition, utility cost of marginal effort is higher than the utility gain from marginal reward, so that $(1 - C_i/B) < 0$ and large. Thus, $(S_i - T_i - 1)(1 - C_i/B) > 0$, and is large;

'Medium' ability: For these students expected score is high enough, so that $(S_i - T_i - 1) > 0$. In addition, disutility of effort (C_i) is marginally less than the utility of reward (B), that is, $(1 - C_i/B)$ is positive but small, such that the magnitude of $(S_i - T_i - 1)(1 - C_i/B)$ is small.

'High' ability: For these students expected score S is high, so that $(S_i - T_i - 1) > 0$. In addition, effort cost (C_i) is very low relative to reward (B) so that $(1 - C_i/B) > 0$ is large. Thus, $(S_i - T_i - 1)(1 - C_i/B) > 0$, and its magnitude is large.

Proposition 1: *Suppose $B - C_i \geq 0$. Students of 'high' and 'low' abilities would prefer incentive based on marginal gains (Treatment 2) relative to that based on absolute gains (Treatment 1), whereas a students of 'medium' ability would prefer incentives tied to absolute gains over that tied to marginal gains.*

Proof: Treatment 2 is more attractive than Treatment 1 if $U_{i,M} > U_{i,A}$, or $(S_i – T_i)(B – C_i) > 4B – C_i$, or $(S_i – T_i – 1)(1 – C_i/B) > 3$. For 'high' and 'low' ability students, the product $(S_i – T_i – 1)(1 – C_i/B)$ is likely to be large, so that $(S_i – T_i – 1)(1 – C_i/B) > 3$ is likely to hold. These students are likely to prefer Treatment 2 to Treatment 1. For the 'medium' ability students $(S_i – T_i – 1)(1 – C_i/B)$ is small, and the inequality is unlikely to be satisfied, and for them Treatment 1 would be more attractive than Treatment 2. Obviously, when $(1 – C_i/B) = 0$, as is likely to hold for some 'low' ability students, Treatment 1 is preferred to Treatment 2.

Proposition 2: *Suppose $B – C_i \geq 0$. Given subjective probability p, the 'high' and 'low' ability students are likely to prefer Treatment 2 (marginal gains) to Treatment 3 (group-performance), unless if the preference for group earning, γ is higher than a threshold value. The 'medium' ability are likely to prefer Treatment 3 to Treatment 2.*

Proof: From (2) and (3), Treatment 3 is preferred to Treatment 2 if $U_{i,G} > U_{i,M}$, or if $(1+\gamma)pB_G – C_G > (S_i – T_i)(B – C_i)$, or if $(1+\gamma)p > (S_i – T_i)\left(\frac{B}{B_G} - \frac{C_i}{B_G}\right) + \frac{C_G}{B_G}$, or if $(1+\gamma)p > (S_i – T_i)\left(\frac{1}{4} - \frac{C_i}{4B}\right) + \frac{C_G}{4B}$, since $B = \frac{1}{4}B_G$. In our experimental setting, there is no reason to believe that the subjective cost of putting marginal effort under Treatment 3 would be different to that under Treatment 2. That is, it is likely that $C_G \cong C_i$. Given this, the condition can be rewritten as $(1+\gamma)4p > (S_i – T_i – 1)\left(1 - \frac{C_i}{B}\right) + 1$. Now, the right hand side of this inequality is large for the 'high' and 'low' ability students, and small for the 'medium' ability students. Hence, Treatment 2 is more likely to be preferred to Treatment 3 by the 'high' and 'low' ability students, unless the value of γ is large so as to satisfy the above condition. The rest of the students are likely to prefer Treatment 3 to Treatment 2.

Proposition 3: *Given the subjective probability p, a student with high utility value from team-spirit (high γ) is likely to prefer Treatment 3 (group-performance based) incentives over Treatment 1 (absolute-performance) incentives.*

Proof: From (1) and (3), Treatment 3 is preferred to Treatment 1 if $U_{i,G} > U_{i,A}$, or if $(1+\gamma)pB_G – C_G > 4B – C_i$. Since $B_G = 4B$, we have $(1+\gamma)p – 1 > (C_G – C_i)/B_G$ or, $1+\gamma > \frac{1}{p}\left(1 + \frac{C_G - C_i}{B_G}\right)$. As argued in Proposition 2, it is likely that $C_G \cong C_i$. Therefore, Treatment 3 is preferred to Treatment 1 if $(1+\gamma)p > 1$, regardless of student ability.

To summarize, we expect the effects of the incentives to differ across student ability (or academic capacity). The relative attractiveness of the Treatments depends on the magnitude of the preference parameter γ, and can be described as follows:

For the ‘high’ ability students: Treatment 2 ≻ Treatment 3 ≻ Treatment 1; if $(1+\gamma)p > 1$
Treatment 2 ≻ Treatment 1 ≻ Treatment 3, otherwise

For the ‘medium’ ability students: Treatment 3 ≻ Treatment 1 ≻ Treatment 2; if $(1+\gamma)p > 1$
Treatment 1 ≻ Treatment 3 ≻ Treatment 2, otherwise

For the ‘low’ ability students: Treatment 2 ≻ Treatment 3 ≻ Treatment 1; if $(1+\gamma)p > 1$
Treatment 2 ≻ Treatment 1 ≻ Treatment 3, otherwise

4. Results

The improvement in NAPLAN test performance from grade 5 to grade 7 for each NAPLAN subject area is presented in Figure 1 which shows the mean percentage change in score and 95% confidence intervals. Students attained significantly higher scores in their grade 7 tests in all three treatments, while their performance in the baseline treatment remained similar to grade 5 levels. Although all three types of incentives seemed to have positive effects in eliciting greater effort, the improvement was highest when group effort was rewarded, followed by reward for marginal gain, and lowest when absolute gain in scores were rewarded.

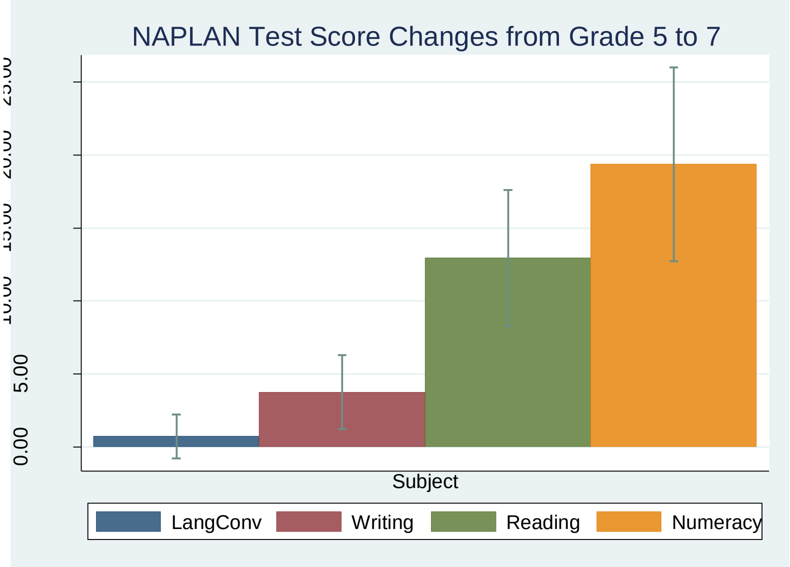


Figure 1: Average improvements from grade 5 to grade 7 in NAPLAN tests

The performance for grade 9 students are perceptively different, as shown in Fig. 2. While we find no improvement in their performance in the language convention (baseline), and writing, students improved their test scores significantly in reading and numeracy. It appears that marginal reward incentives had the largest impact compared to group incentives for the higher grade students.

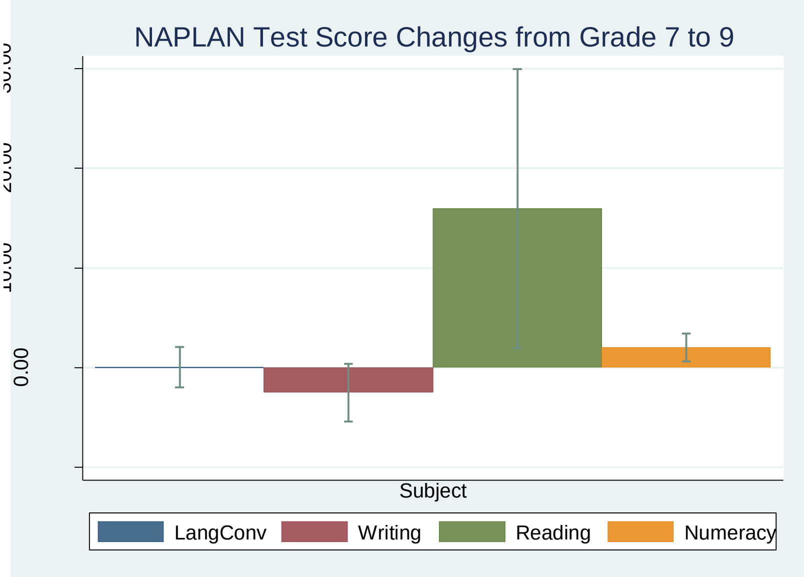
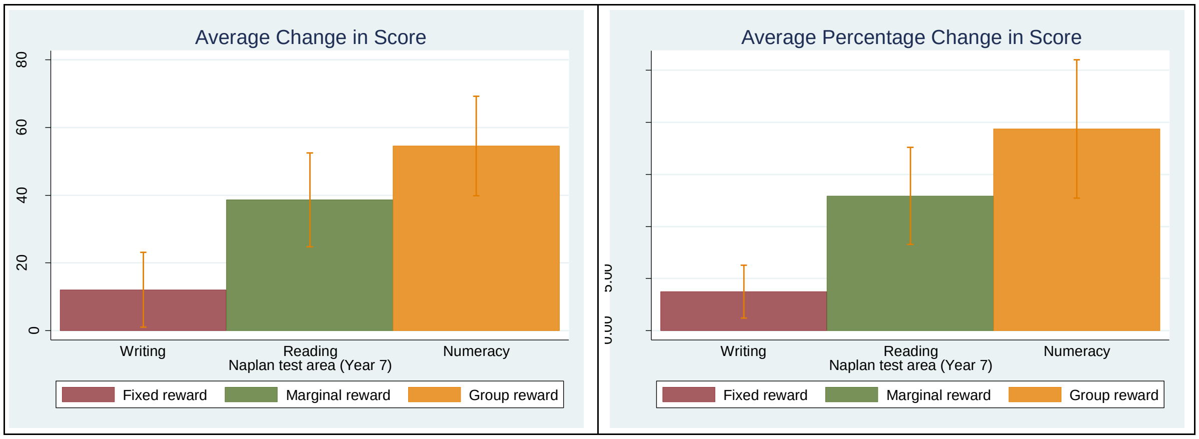


Figure 2: Average improvements from grade 7 to grade 9 in NAPLAN tests

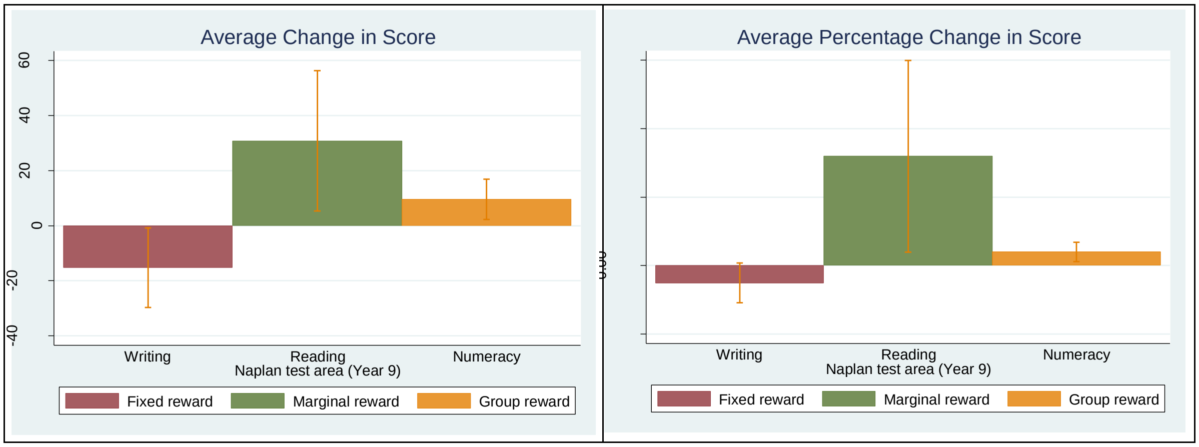
In order to evaluate the effect of each treatment relative to the baseline, we next look at the within-subject difference in test score changes from grade 5 to 7, and from grade 7 to 9, in Figure 3 and 4, respectively. Grade 7 students seemed to have responded strongly to all types of incentives relative to the 'verbal encouragement' in the baseline. All changes are statistically significant, but the magnitude of the power of incentives varied a great deal. As shown in Fig. 3, the fixed incentive in the WRITING test was the least effective, while the group-based incentives provided in NUMERACY test was most powerful, both in terms of absolute and percentage 'gain'. The importance of group-based incentives are reinforced by the grade 9 comparisons in Figure 4, but the impact of rewarding marginal gain in scores is seven times higher.

Given the positive impact of incentives in improving performance in most test areas for students of both grades, it is worthwhile to investigate who responds more to the incentives – did the incentives induce greater effort from some students more than others? We compare the kernel density plots of percentage change in NAPLAN scores for grade 7 and grade 9 student in Figure 5 and 6, respectively. Compared to the baseline, the distribution of score changes is positively skewed for the three treatment when considering grade 7 students in Figure 5. This implies gains in test scores are likely to be higher for better-performing students. In contrast, the distributions for all three treatments look very similar to that of the baseline when considering grade 9 students in Figure 6, except for the change in Reading where marginal improvement in score was incentivized. However, it seems that a few outliers are driving this difference for the outcome in Reading.



(a) (b)

Figure 3: Change in test scores for each treatment relative to the baseline (Language Conventions) among grade 7 students



(a) (b)

Figure 4: Change in test scores for each treatment relative to the baseline (Language Conventions) among grade 9 students

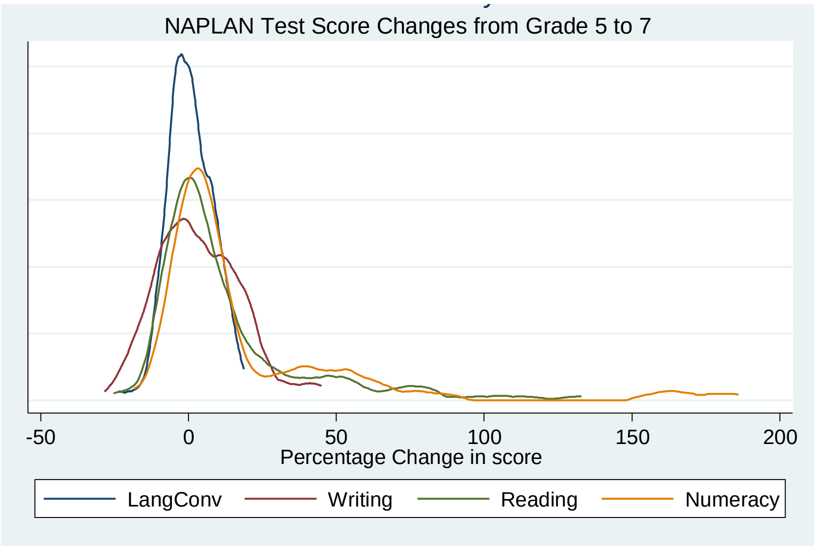


Figure 5. Kernel density plots for Grade 7 students

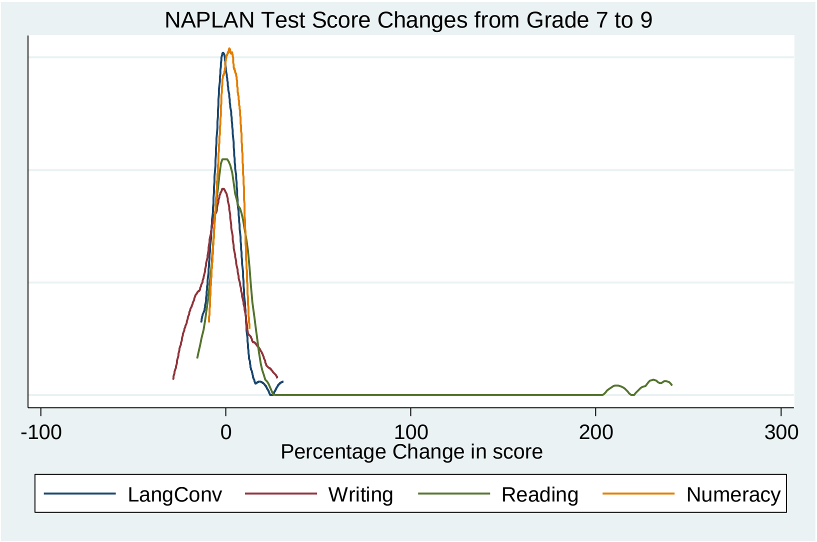


Figure 6. Kernel density plots for Grade 9 students

Alternatively, we also use previous year's Progressive Assessment Test (PAT) scores in two subject areas most relevant to the NAPLAN test areas - Mathematics and Reading, as indicators of 'ability'. In Figure 7a and 8a in the Appendix, we assign students to the top, middle and bottom thirds of the ability distributions to examine the extent to which incentives may affect students differently.[5]

[5] These tests are routinely conducted by schools in accordance with the ACER (Australian Council for Education Research), an independent, not-for-profit research organisation.

3.1 OLS, random effect and fixed effect estimates

Given that we have a within-subjects design, results from regressions that control for random effects and fixed effects are reported along with the ordinary least squares estimates, in Table 1 for grade 7 and Table 2 for grade 9 students. We control for student gender, section they are enrolled in, and PAT-reading and math scores. Results in Table 1 shows that students in grade 7 responded strongly to both incentives for marginal gains and group effort, by improving their test scores by 10 to 14 percentage points in OLS regressions and 8 to 11 percentage points when controlling for random effects. Allowing for individual fixed effects further reduces the effect sizes to 5 and 11 percentage points for Reading and Numeracy, respectively. The impact is always higher for Numeracy where group effort is rewarded.

Table 1. OLS, Random effects, Fixed effects Regressions for Percentage Change in NAPLAN Score from Grade 5 to Grade 7

	OLS	OLS	OLS	OLS	RE	FE
Male	1.940	2.506	3.955	3.723	4.075	
	(2.687)	(2.779)	(2.615)	(2.645)	(2.952)	
Absolute gain reward	1.916	1.073	1.377	1.329	-0.980	-4.201*
(Writing)	(1.503)	(1.622)	(1.604)	(1.633)	(1.833)	(2.164)
Marginal gain reward	11.391***	11.149***	9.805***	10.317***	8.048***	4.746***
(Reading)	(2.193)	(2.221)	(2.152)	(2.187)	(1.896)	(1.731)
Group reward	17.265***	16.902***	14.529***	14.190***	11.548***	10.506***
(Numeracy)	(3.103)	(3.336)	(3.051)	(3.093)	(2.596)	(2.162)
Section						
ENG071B	15.545***	12.067**	10.316*	9.089	10.066	
	(4.461)	(5.548)	(5.349)	(5.734)	(6.173)	
ENG071C	9.925***	5.828	3.778	2.683	2.642	
	(2.553)	(3.903)	(3.182)	(3.848)	(4.199)	
ENG071D	10.513***	6.139	6.348	5.577	6.672	
	(3.853)	(4.125)	(4.944)	(4.663)	(5.024)	
ENG071E	5.216**	0.965	1.133	-0.106	-0.253	
	(2.117)	(3.348)	(3.252)	(3.790)	(4.093)	
ENG071F	22.396***	17.016*	19.036**	16.908*	17.715*	
	(7.814)	(9.370)	(8.849)	(9.567)	(10.661)	
PAT-Reading Score		-0.474		-0.192	-0.257	
		(0.300)		(0.304)	(0.362)	
PAT-Math Score			-0.387	-0.340	-0.315	
			(0.248)	(0.255)	(0.274)	
Constant	-8.321***	4.405	1.281	5.304	8.142	6.758***
	(1.905)	(7.807)	(6.165)	(8.571)	(9.455)	(0.741)
Observations	471	423	420	410	410	471
R-squared	0.143	0.162	0.151	0.154	0.152	0.068

Notes: Robust standard errors in parenthesis. OLS regressions clustered on student id. */**/*** denote significance at 0.1/0.05/0.01 levels. Baseline subject area (Language Conventions) is the reference category. F-stat reported for OLS and FE models, while Chi-sq reported for RE model.

These findings are somewhat changed when looking at the role of incentives for grade 9 students. While incentives for rewarding marginal gains in test scores still elicits significantly higher effort, group incentives work to a much smaller extent for the higher grade students. Specifically, the gain in Reading score is two (RE) to three (FE) times higher than those experienced by grade 7 students. In contrast, the gain is Numeracy is only significant when controlling for individual fixed effects, such that the magnitude of gain is merely a third of that attained by grade 7 students.

Table 2. OLS, Random effects, Fixed effects Regressions for Percentage Change in NAPLAN Score from Grade 7 to Grade 9

	OLS	OLS	OLS	OLS	RE	FE
Male	-0.388	-3.661	-3.549	-3.706	-3.706	
	(3.785)	(4.983)	(5.535)	(5.780)	(5.780)	
Absolute gain reward	-2.964*	-5.281***	-3.010	-4.149*	-4.149*	-4.507
(Writing)	(1.761)	(1.924)	(2.201)	(2.324)	(2.324)	(3.041)
Marginal gain reward	15.523**	15.325*	17.070*	17.537*	17.537*	13.481**
(Reading)	(6.985)	(8.150)	(9.444)	(9.921)	(9.921)	(6.033)
Group reward	2.543*	1.648	0.598	0.563	0.563	2.954*
(Numeracy)	(1.430)	(1.848)	(2.211)	(2.272)	(2.272)	(1.769)
Section						
MAT091D	-3.761	-7.342	1.375	-0.837	-0.837	
	(4.038)	(8.394)	(8.386)	(13.708)	(13.708)	
MAT091B	5.675	0.995	7.741	6.251	6.251	
	(6.287)	(7.925)	(9.469)	(12.639)	(12.639)	
MAT091C	2.354	2.912	6.219	5.009	5.009	
	(6.215)	(12.059)	(13.949)	(17.903)	(17.903)	
MAT091E	-4.647	-5.187	-3.710	-5.013	-5.013	
	(4.121)	(6.225)	(7.359)	(10.852)	(10.852)	
Pat Reading Score		-0.281		-0.179	-0.179	
		(0.280)		(0.305)	(0.305)	
Pat Math Score			-0.105	-0.078	-0.078	
			(0.434)	(0.684)	(0.684)	
Constant	0.027	8.624	3.362	7.406	7.406	0.995
	(4.024)	(10.196)	(13.081)	(22.609)	(22.609)	(1.477)
Observations	245	177	155	143	143	247
R-squared	0.075	0.084	0.084	0.090	0.090	0.057

Notes: Robust standard errors in parenthesis. OLS regressions clustered on student id. */**/*** denote significance at 0.1/0.05/0.01 levels. Baseline subject area (Language Conventions) is the reference category. F-stat reported for OLS and FE models, while Chi-sq reported for RE model.

3.2 Quantile regression estimates

We now turn to an evaluation of the treatment effects across the distribution of test score changes. The quantile regression estimates at the 10, 25, 33, 66, 75, and 90th percentiles are reported for grade 7 students in Table3 and grade 9 students in Table 4. Standard errors are computed using 200 bootstrap replications. As indicated from the kernel density plot in Figure 5, the distribution for gains in Writing has a larger spread implying that grade 7 students at the lower and upper tails are likely to perform differently than the baseline group. We find this is evident in Table 3 which reports the significantly negative (positive) quantile regression estimates for the 10th (90th) percentile for Writing. In contrast, students at the top of the distribution (75th and 90th percentiles) gain significantly in Reading, while the gains in Numeracy are evidently positive for all students irrespective of their position in the distribution. However, the gains are progressively larger for students ranked higher in the distribution.

The results for grade 9 students are similar but only evident for students at the bottom tail for the distribution. After controlling for their PAT-reading and math scores, we find that the gains in Numeracy accrue only for the 10th percentile, while the decline in Writing test scores occurs only for the 25th percentile.

Table 3. Quantile Regression: Percentage Change in NAPLAN Score from Grade 5 to Grade 7

	10th quantile	25th quantile	33rd quantile	66th quantile	75th quantile	90th quantile	10th quantile	25th quantile	33rd quantile	66th quantile	75th quantile	90th quantile
Male	-1.904	-0.391	0.635	1.923	2.114	-1.397	-1.299	0.186	0.915	2.334	2.135	-0.506
	(1.183)	(1.121)	(1.021)	(1.625)	(2.156)	(3.425)	(1.294)	(1.308)	(1.331)	(1.895)	(2.529)	(3.573)
Absolute gain reward (Writing)	-5.122**	-1.143	-0.927	5.612***	6.651***	9.593***	-5.150***	-2.120	-1.432	4.161	5.094*	7.255**
	(2.004)	(1.810)	(1.441)	(2.138)	(1.948)	(3.240)	(1.881)	(2.071)	(2.010)	(2.806)	(2.676)	(3.090)
Marginal gain reward (Reading)	2.039	1.899	1.748	5.865**	10.235***	33.397***	2.039	1.532	0.995	4.591	10.377***	21.358***
	(1.493)	(1.229)	(1.352)	(2.500)	(3.720)	(8.376)	(1.675)	(1.595)	(1.788)	(3.105)	(3.608)	(6.135)
Group reward (Numeracy)	4.755***	5.013***	5.202***	7.690***	12.136*	40.032***	4.554***	4.283***	4.546***	4.933***	6.152	18.476***
	(1.585)	(1.390)	(1.332)	(2.482)	(6.755)	(9.858)	(1.740)	(1.472)	(1.541)	(1.840)	(5.122)	(6.997)
Section												
ENG071B	2.985**	2.638	2.623	9.450***	12.721	28.302**	1.323	-1.329	1.097	7.418	11.764	30.489***
	(1.464)	(1.927)	(1.770)	(3.241)	(7.839)	(12.588)	(2.261)	(2.912)	(2.696)	(4.667)	(9.688)	(9.704)
ENG071C	2.142	2.078	3.312*	9.432***	12.892***	15.114**	0.288	0.097	2.599	4.332	6.667	6.204
	(2.267)	(1.935)	(1.917)	(3.336)	(4.219)	(5.986)	(2.532)	(2.577)	(2.330)	(4.036)	(5.622)	(7.264)
ENG071D	1.965	3.478**	2.435*	6.934**	11.347***	13.736*	-0.229	1.918	2.742	5.356	7.838*	12.361
	(2.061)	(1.623)	(1.261)	(2.718)	(3.143)	(8.254)	(3.430)	(2.941)	(2.444)	(3.515)	(4.522)	(11.964)
ENG071E	1.581	0.516	0.061	8.028***	10.054***	6.160*	0.714	-0.492	-0.240	3.896	4.549	0.508
	(1.820)	(1.410)	(1.641)	(2.408)	(2.614)	(3.618)	(2.319)	(2.027)	(2.084)	(3.036)	(4.632)	(6.129)
ENG071F	2.945	4.838*	4.435	12.948*	18.663	33.398	1.134	3.266	4.409	9.054	16.282	29.562
	(2.791)	(2.724)	(2.758)	(6.853)	(11.370)	(47.358)	(4.058)	(4.327)	(4.219)	(6.823)	(13.235)	(53.526)
PAT-Reading Score							-0.032	-0.017	0.022	-0.055	0.011	0.127
							(0.150)	(0.147)	(0.141)	(0.261)	(0.293)	(0.389)
PAT-Math Score							-0.044	-0.128	-0.100	-0.314	-0.415	-0.711**
							(0.116)	(0.121)	(0.118)	(0.213)	(0.310)	(0.360)
Constant	-8.927***	-6.000***	-4.531***	-3.648**	-3.387*	1.538	-6.289	-2.033	-2.425	6.493	7.166	16.753
	(1.394)	(1.285)	(1.174)	(1.588)	(2.032)	(2.880)	(4.457)	(4.750)	(4.593)	(6.517)	(7.867)	(10.544)
Observations			471						410			
R-squared	0.067	0.030	0.032	0.057	0.074	0.205	0.066	0.033	0.035	0.071	0.088	0.213

Notes: Bootstrapped standard errors in parenthesis. */**/*** denote significance at 0.1/0.05/0.01 levels. Baseline subject area (Language Conventions) is the reference category.

Table 4. Quantile Regression: Percentage Change in NAPLAN Score from Grade 7 to Grade 9

	10th quantile	25th quantile	33rd quantile	66th quantile	75th quantile	90th quantile	10th quantile	25th quantile	33rd quantile	66th quantile	75th quantile	90th quantile
Male	0.637	-0.751	-1.092	-0.230	0.492	0.352	2.209	-1.859	-2.169	0.506	1.601	2.125
	(1.741)	(1.570)	(1.284)	(1.348)	(1.426)	(2.371)	(2.875)	(2.471)	(2.187)	(2.094)	(2.826)	(4.718)
Absolute gain reward (Writing)	-10.065***	-7.085***	-4.286*	-1.302	0.349	5.045	-6.130	-6.165*	-4.903	-2.149	-3.058	-0.398
	(3.084)	(2.702)	(2.339)	(1.724)	(2.714)	(4.035)	(3.926)	(3.243)	(3.092)	(2.392)	(2.812)	(4.971)
Marginal gain reward (Reading)	1.275	2.158	2.452	3.199	5.527**	7.835	1.943	1.313	2.565	4.202	4.941	5.740
	(2.789)	(1.662)	(1.581)	(2.269)	(2.295)	(63.095)	(3.607)	(2.323)	(2.047)	(3.016)	(15.140)	(86.971)
Group reward (Numeracy)	4.363*	1.896	2.823*	3.447**	3.453**	0.543	8.445**	3.212	3.968	2.184	2.401	-1.552
	(2.406)	(1.566)	(1.604)	(1.338)	(1.453)	(2.748)	(3.973)	(2.658)	(2.397)	(2.742)	(2.943)	(3.722)
Section												
MAT091D	0.630	2.209	1.243	-1.159	-2.187	-5.269**	6.459	8.561*	5.124	-0.876	-3.012	-0.318
	(2.887)	(2.537)	(2.449)	(2.182)	(2.022)	(2.496)	(6.653)	(4.938)	(4.220)	(4.227)	(5.069)	(10.368)
MAT091B	2.094	3.870	4.065*	3.692*	1.858	1.625	2.652	7.275*	6.098*	2.600	0.742	20.118
	(2.631)	(2.551)	(2.172)	(2.102)	(2.527)	(22.026)	(5.300)	(3.820)	(3.211)	(3.989)	(5.334)	(19.319)
MAT091C	0.705	2.180	2.691	4.425**	2.902	2.928	-4.923	3.022	0.101	0.722	0.013	2.645
	(2.798)	(2.539)	(2.470)	(2.140)	(2.202)	(3.809)	(7.721)	(5.310)	(4.844)	(4.876)	(6.096)	(47.172)
MAT091E	4.248*	2.043	1.421	-0.549	-1.286	-2.822	1.649	4.209	3.157	-2.152	-2.769	-0.323
	(2.343)	(2.031)	(1.807)	(2.211)	(2.300)	(3.027)	(4.368)	(3.460)	(3.062)	(3.887)	(4.190)	(8.386)
PAT-Reading Score							-0.374	-0.130	-0.314	-0.105	-0.196	-0.001
							(0.372)	(0.286)	(0.226)	(0.243)	(0.293)	(0.423)
PAT-Math Score							0.174	0.361	0.350	-0.042	0.056	0.097
							(0.278)	(0.269)	(0.233)	(0.221)	(0.233)	(0.526)
Constant	-11.120***	-5.431***	-4.417***	0.727	2.562	8.694**	-8.718	-10.935	-5.904	4.468	6.214	4.594
	(2.484)	(1.715)	(1.488)	(1.718)	(2.043)	(3.572)	(10.347)	(8.341)	(7.052)	(8.098)	(9.754)	(19.027)
Observations			245							143		
R-squared	0.158	0.058	0.047	0.044	0.036	0.041	0.131	0.079	0.065	0.055	0.046	0.036

Notes: Bootstrapped standard errors in parenthesis. */**/*** denote significance at 0.1/0.05/0.01 levels. Baseline subject area (Language Conventions) is the reference category.

4. Concluding remarks

The study explores the effects of various incentive structures in increasing students' motivation to do better in real tests compared to their past performance in similar tests. We have examined the effects of incentives based on absolute gain, marginal gain, and performance as a group. Each type of incentives were then compared against the performance in a baseline test with no monetary or non-monetary incentives. The incentive programs were rolled out for grade 7 and grade 9 students at a public high school in Queensland during their NAPLAN tests.

We find improvements among year 7 students in all subject areas (language convention, writing, reading, numeracy). Against the baseline area of language convention, we find significant gains over the last NAPLAN performance (in year 5) in reading, writing, and numeracy test performances for grade 7 students. Student performance in topping their own previous (adjusted) NAPLAN score was most pronounced in Numeracy test, where group-based incentives were implemented. The gain in student performance was substantial in Reading tests where incentives were given based on marginal improvements. Achievement was lowest in Writing test, under incentives based on absolute gain. Gains in performance were not uniform - the highest gains were concentrated among the high-ability students in each NAPLAN test area.

The grade 9 students responded to the incentives differently. The incentives based on marginal gain were effective in improving grades, while their response to group-based incentives were smaller in magnitude compared to the grade 7 students. Moreover, unlike their grade 7 peers, the low-ability students demonstrated higher gains under group-incentive.

Overall, the results provide new insights into design of incentive structure aimed at eliciting greater student efforts in school. We show that incentive structures that reward team-based performance would likely to be most effective in real test situations when performance matters most. In our experiment, the effectiveness of group-based incentives is likely due to individual aversion to let other group-members down.

Previous research emphasised the rewards to be provided immediately after the performance, because myopic students may undervalue the incentives provided over a longer horizon. This experiment, however, showed that power of incentives is substantial even when they materialize after a month. From the insights gained from hyperbolic discounting and present-bias, it seems reasonable to argue that immediately available incentives would possibly work better.

There is concern in the education literature that these types of incentives will crowd out intrinsic motivation. However, extrinsic rewards can be a powerful tool to promote intrinsic motivation and habit formation (Lepper et al 1973, Cameron et al. 2005, Bettinger 2010). If extrinsic rewards increase students' returns to achievement in tests, then properly structured extrinsic rewards could potentially build (rather than crowd out) intrinsic motivation to plan and prepare better for tests. For example, students may not have the knowledge of the steps necessary to improve their achievement on a test. However, they may be able to effectively respond to incentives on intervening tasks such as learning the daily lessons, completing homework, or focusing on a practice test.

The study has a few limitations. First, the study has been conducted in a low-SES school, and similar incentives may be less powerful elsewhere. Second, we are unable to say anything about whether or not the incentives affected later test performance. Neither are we able to investigate the presence of any spillover effects on the later cohorts. These caveats limit the longer-term policy implications of the study. However, the clean design of our experiment allowed clear identification of how students respond to different monetary and non-monetary incentives. The findings from this study inform us about the superiority of group-based incentives as effective instruments for policy, which are much less costly than individual-based incentives provided by some schools. This insight can be used to design optimal cost-effective short-term incentive schemes to elicit higher effort from students in competitive assessments.

References

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos (2009). "Incentives and Services for College Achievement: Evidence from a Randomized Trial." American Economic Journal: Applied Economics, 1, 136–163.

Azmat, Ghazala, Nagore Iriberri (2010) "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students", Journal of Public Economics, 94(7-8): 435-452.

Bandiera, Oriana, Iwan Barankay and Imran Rasul (2005). "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, 120, 917–962.

Bandiera, Oriana, Iwan Barankay and Imran Rasul (2010). "Social Incentives in the Workplace." *Review of Economics Studies*, 77, 417–458.

Bandiera, Oriana, Iwan Barankay and Imran Rasul (2013). "Team Incentives: Evidence from a Firm Level Experiment." *Journal of European Economic Association*, 11, 1079–1114.

Barrow, Lisa, Lashawn Richburg-Hayes, Cecilia Elena Rouse, and Thomas Brock (2014). "Paying for Performance: The Educational Impacts of a Community College Scholarship Program for Low-Income Adults." Journal of Labor Economics, 32, 563–599.

Bettinger, Eric (2012). "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." Review of Economics and Statistics, 94, 686–698.

Bettinger, Eric (2010) "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores", NBER Working Paper 16333.

Cameron, Judy, W. David Pierce, K.M. Banko, and A. Gear. (2005). "Achievement- Based Rewards and Intrinsic Motivation: A Test of Cognitive Mediators." Journal of Educational Psychology, 97(4): 641-655.

Carrell, Scott E., Richard L. Fullerton, and James E. West (2009). "Does Your Cohort Matter? Measuring Peer Effects in College Achievement." *Journal of Labor Economics*, 27, 439–464.

Carrell, Scott E., Mark Hoekstra, and James E. West (2011). "Is Poor Fitness Contagious? Evidence From Randomly Assigned Friends." *Journal of Public Economics*, 95, 657–663.

Falk, Armin and Andrea Ichino (2006). "Clean Evidence on Peer Effects." *Journal of Labor Economics*, 24, 39–58.

Foster, Gigi (2006). "It's Not Your Peers, and It's Not Your Friends: Some Progress Toward Understanding the Educational Peer Effect Mechanism." *Journal of Public Economics*, 90, 1455–1475.

Fryer, Roland G. (2011) “Financial Incentives and Student Achievement: Evidence from Randomized Trials”, The Quarterly Journal of Eonomics, 126(4): 1755-1798.

Kremer, Michael, and Dan Levy (2008). “Peer Effects and Alcohol Use among College Students.” *Journal of Economic Perspectives*, 22, 189–206.

Kremer, Michael, Edward Miguel, and Rebecca Thornton (2009). “Incentives to Learn.” Review of Economics and Statistics, 91, 437–456.

Lepper, Mark, David Greene, and Richard E. Nisbett. (1973). “Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the `Overjusti_cation' Hypothesis.” Journal of Personality and Social Psychology, 28(1): 129-137.

Levitt Stephen, John List, Susanne Neckerman, Sally Sadoff (2012). “The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance”, NBER WP 18165 (http://www.nber.org/papers/w18165).

Lyle, David S. (2007). “Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point.” *Review of Economics and Statistics*, 89, 289–299.

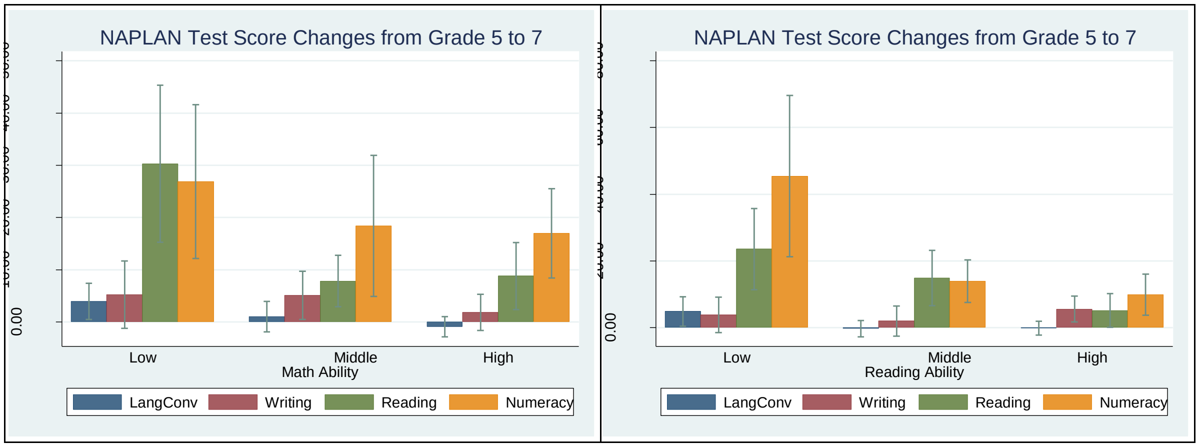
Mas, Alexandre and Enrico Moretti (2009). “Peers at Work.” *American Economic Review*, 99, 112–145.

Sacerdote, Bruce (2001). “Peer Effects with Random Assignment: Results for Dartmouth Roommates.” *Quarterly Journal of Economics*, 116, 681–704.

Zimmerman, David J. (2003). “Peer Effects in Academic Outcomes: Evidence from a Natural Experiment.” *Review of Economics and Statistics*, 85, 9–23.

<u>Appendix</u>

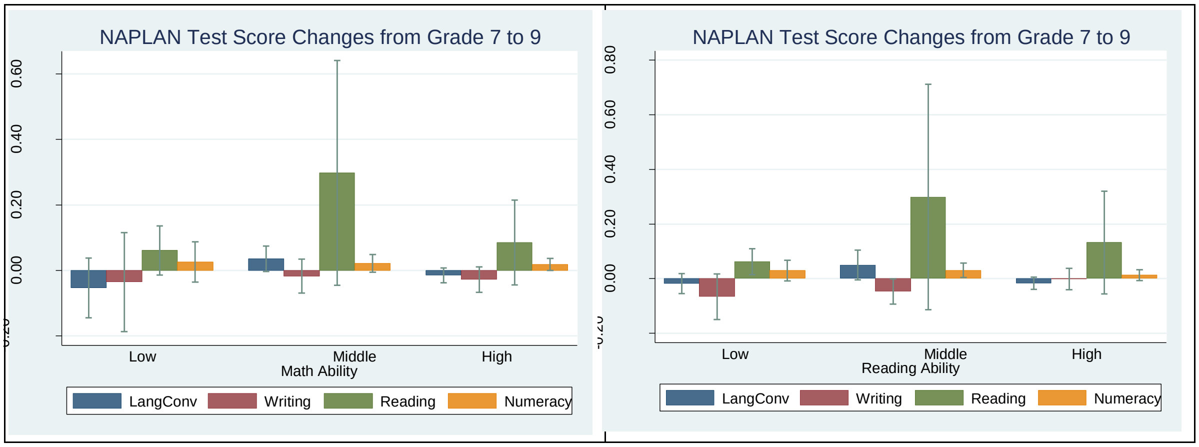
As seen in Figure 1a, overall we find that improvements for grade 7 students are found to be highest for low ability students. Importantly, these improvements were specifically in the areas of Reading and Numeracy made by students who were weakest in reading and mathematics to begin with. The incentives for group performance boosted performance by more than 4 times for the bottom third compared to the top third of the distribution of PAT-reading scores. Likewise, the incentives for marginal rewards improved performance by almost 4 times for the lowest PAT-reading ability students and almost 5 times for the lowest PAT-math ability students. Interestingly, even the performance in the baseline subject improved significantly for the lowest ability students in both the PAT-reading and math ability distributions. Therefore, incentives are found to have a stronger positive impact for low ability students, with marginal reward and group-based rewards resulting in largest gains.



(a) (b)

Figure 1a: (a) Improvement in test scores among grade 7 students in four NAPLAN test areas by PAT-math ability; (b) Improvements of test scores among grade 7 students in four NAPLAN test areas by PAT-reading ability

The same, however, cannot be said about the grade 9 students. Figure 2a shows that although the gains were not statistically significant irrespective of the ability of students. Two exceptions were the significant gains in Reading scores among low Reading-ability students and Numeracy scores among high Math-ability students.



(a) (b)

Figure 2a: (a) Improvement in test scores among grade 9 students in four NAPLAN test areas by PAT-math ability; (b) Improvements of test scores among grade 9 students in four NAPLAN test areas by PAT-reading ability