Cultural Context in Standardized Tests*

Isabella Dobrescu Richard Holden† Alberto Motta
Adrian Piccoli Philip Roberts Sarah Walker

December 12, 2021

Abstract

We report results from a field experiment on cultural context in standardized tests among 6th- and 8th-grade school students in Australia. The National Assessment Program Literacy and Numeracy (NAPLAN) is a series of basic-skills tests given to Australian students. In our experiment, 1135 students in Dubbo – a regional area in the North-Western part of the state of New South Wales – were randomly assigned to either a regular NAPLAN test or a *contextualized* test designed specifically for this experiment by the NSW Aboriginal Education Consultative Group — a not-for-profit Aboriginal organization. The contextualized test was specifically designed to mimic the regular test, but adapted to the local context of Dubbo. We evaluate effects on tests scores in numeracy for grades 6 and 8, and reading for grade 6. In numeracy, we do not find robust evidence of an impact on test scores. In reading, we find qualitatively large effects. The average treatment effect for reading is 0.27 s.d., with higher effects for Indigenous students (0.30 s.d.) than non-Indigenous students (0.24 s.d.) Together these results imply that cultural context may be important for performance on certain types of basic-skills tests.

*This study was pre-registered in the AEA RCT Registry: *AEARCTR-0007309*. We are grateful to numerous members of the NSW Department of Education for helping make this project possible. We owe special thanks to the Aboriginal Education Consultative Group for writing the contextualized tests and for their support and encouragement with this project. Carol Taylor kindly provided the psychometric test validity analysis; Fabio Martinenghi provided exceptional research assistance. All authors thank the Economics of Education Knowledge Hub @UNSWBusiness for financial support.

†Corresponding author. UNSW Business School, richard.holden@unsw.edu.au. Dobrescu: UNSW Business School, dobrescu@unsw.edu.au. Motta: UNSW Business School, motta@unsw.edu.au. Piccoli: UNSW Business School. Roberts: University of Canberra, . Walker: UNSW Business School, s.walker@unsw.edu.au.

1 Introduction

Standardized tests are controversial. They hold out the promise of admitting students to college based on academic merit rather than family background alone, and they provide a means to assess student progress throughout primary and secondary schools, and offer a basis for interventions and additional resourcing to help students that are underperforming.

But such tests have an ugly history. In the United States the Scholastic Aptitude Test (SAT) was first developed by a noted eugenicist Carl C. Brigham when he was a psychology professor at Princeton in the 1920s.

Yet perhaps the most important early champion of the use of SATs in college admissions was Harvard President James Conant, who did so for considerably more noble reasons. Harvard began using the SAT for scholarship applicants in 1934, and then mandated them for all students seeking admission in 1935. Conant wrote approvingly of standardized tests in the *Atlantic Monthly* in 1943, and ultimately oversaw the establishment of the Educational Testing Service in 1948. Conant's stated goal was to provide equal educational opportunity, regardless of privilege. Indeed, his Harvard presidency was met with much consternation by Boston's Brahmins—who saw little wrong with inherited privilege.[1]

Today there is considerable doubt that the SAT has achieved Conant's goal of making admission to college—particularly elite colleges—more accessible. It is widely acknowledged that standardized test scores are strongly correlated with socioeconomic status. And there are good reasons to believe that standardized tests typically exhibit cultural bias.

This bias can come from a number of different sources. Psychologists refer to the target object of a test as a *construct*. Examples of constructs are intelligence, anxiety and self-esteem. A question (an *item*) is said to be *biased* if it does not measure the construct alone because other factors confound it—for example, cultural affiliation (Reynolds and Suzuki, 2012). In the cross-cultural assessment literature, psychologists have identified three forms of bias (Van de Vijver and Tanzer, 2004;

[1] See here.

Triandis, 2000).

First, *construct bias* occurs when the construct of interest "is not identical across cultural groups", yet the same test is administered across groups (Van de Vijver and Tanzer, 2004). For instance, most general intelligence tests focus on skills that are highly valued in Western culture, such as logic and memory. But they ignore social aspects of intelligence, which are more heavily emphasized in non-Western cultures, thus generating a bias against members of such cultures (Van de Vijver and Tanzer, 2004). To avoid this bias, whenever the construct changes across cultures, the test should change accordingly.

Second, *method bias* occurs when the two (or more) sampled groups are not comparable. This might be due to either sampling issues, or to fundamental differences in the populations of interest. Economists call this issue "failure of the common support condition" (Lechner, 2008). An example of method bias is the case where the education levels of the two groups compared are radically different.

Third, *item bias* occurs whenever a given question (or item) uses concepts that are not equally familiar across the cultures of interest. It occurs when correctly answering a question *requires additional skills other than the one being measured, and these skills vary across groups* (Ackerman, 1992). It also occurs when individuals of different groups vary in what they perceive to be the most socially appropriate answer to a given question.

While there is a vast body of work that categorizes different types of bias and also seeks to measure the extent of various forms of bias, there is a paucity of causal evidence on the topic. Simply observing that there is a correlation between standardized tests and members of various social or ethnic groups does not imply that *no* test can be unbiased, and it does not pin down what type of bias is present in an existing test.

Item bias is particularly problematic for two core uses of standardized tests. As a screening device (say for college admissions) a test subject to item bias will lead to under-admission of certain cultural groups for reasons unrelated to the construct. If a test seeks to measure the ability to perform well as an undergraduate, but suffers from item bias, then it doesn't achieve its desired goal. It incorrectly rejects certain

applicants from one cultural group simply because they are from that cultural group, despite them having an equal ability to perform well as an undergraduate as members of other cultural groups who are admitted.

As a diagnostic tool for improving K-12 education, tests that are item biased are also problematic. In this case they will indicate that students from a certain cultural group are underperforming relative to their true ability. This can result in the application of the wrong educational intervention. For instance, it could lead to repetition of old material in a culturally biased way rather than presentation of new material in a culturally contextual fashion. The latter is likely to lead to better educational outcomes.

In addition to this, a number of higher-education institutions are moving to *test-blind* admissions, whereby they will not consider SAT/ACT scores even if they are submitted. The University of California system is scheduled to move to this system in 2023-24 at the latest. However, if standardized tests are eschewed in college admissions because of cultural bias then some other set of criteria will be used. It is unclear that criteria such as extra-curricular activities are less culturally biased. A much larger number of institutions have adopted *test-optional* policies under which applicants are not required to submit SAT/ACT scores with college applications. But the evidence to date suggests that such policy have little effect. For instance, (Saboe and Terrizzi, 2019) find that SAT-optional policies have no effect on racial or socioeconomic diversity, the gender ratio of institutions, or the quality of the student population, and no sustained increase in numbers of applications.

All of this suggests that perhaps the best way to remove cultural bias in college admissions is to attempt to de-bias tests like the SATs rather than abolish them.

The experiment we conduct in this paper isolates the magnitude of item bias against Australian school students from rural and remote areas, and who are indigenous.

Understanding the type and extent of cultural bias in standardized tests is important for two reasons. First it provides a picture of how students from certain cultural groups are actually performing in mastering certain subject matter. Second, it provides a step toward understanding how educational materials (such as textbooks,

handouts, and multi-media content) may also be culturally biased.

1.1 This paper

The specific goal of this paper is to provide causal evidence on the extent of cultural bias in standardized tests. To do so we conducted a field experiment among 6th- and 8th-grade school students in Australia. The National Assessment Program Literacy and Numeracy (NAPLAN) is a series of basic-skills tests given to Australian students. We partnered with the New South Wales Department of Education to conduct the experiment, in which 1135 students in Dubbo–a regional town in the North-Western part of the state of New South Wales–were randomly assigned to either a regular NAPLAN test or a *contextualized* test designed to mimic the regular test, but adapted to the local context of Dubbo. That local context includes being part of a "rural and remote" community about 250 miles (400 km) from Sydney, with a population of just over 70,000, having a median household income 79% of the statewide median, and a comparatively high Indigenous population (18.6% compared to 2.9% statewide).

The contextualized tests were designed specifically for this experiment by the Aboriginal Education Consultative Group (AECG)—the peak body for Aboriginal education in New South Wales. They designed a set of standardized tests in reading and numeracy for students in Year 5 (primary school) and Year 7 (secondary school) that closely mimic the actual NAPLAN tests used across the state, but used items and language culturally relevant for students living in Dubbo. The tests were designed to be of equal difficulty, as we illustrate in Section 2.

We evaluate the impact of the contextualized tests on numeracy scores for students in Years 6 and 8, and reading scores for students in Year 6. In the numeracy tests, we do not find robust evidence of a treatment effect. In Year 8, treatment has no impact on numeracy scores in the pooled sampled of students, as well as separate sub-samples for Indigenous and non-Indigenous students. In Year 6, there is no evidence of an impact on Indigenous test scores and weak evidence of a negative impact on non-Indigenous scores, namely through an increase in the number of questions not attempted. In reading, however, we find qualitatively large effects for all students. The average treatment effect is 0.27 s.d., with higher effects for Indigenous students

(0.30 s.d.) than non-Indigenous students (0.24 s.d.). Back of the envelope calculations suggest that the contextualized reading test closes the rural-urban reading gap by 33 percent and the Indigenous-non-Indigenous gap by 50 percent. Together these results imply that cultural context may be important for performance on certain types of basic-skills tests.

1.2 Measuring and correcting bias

Researchers have been concerned with the existence of bias in standardized tests since their very inception (Binet and Simon, 1916). It is no accident that Alfred Binet, one of the early developers of standardized tests, called the type of intelligence targeted by these tests "uncultured intelligence"(Binet and Simon, 1916). Indeed, his goal was to measure intelligence free of cultural aspects. Today, we often call this type of intelligence "cognitive ability".

In the 1970s, these concerns sparked a fiery national debate, as American schools and universities gradually adopted standardized testing as a fundamental part of their assessment strategy. Both scholars and the public (Raspberry, 1974) were divided about whether the benefits of such tests outweighed the risk of bias against minorities (Gallagher, 2003).[2] For a short history of standardised tests in the US, see Gallagher (2003) and Grodsky et al. (2008).

Several disciplines have studied bias in standardized tests, both empirically and theoretically. Sociologists have focused on the relationship between standardized tests and socioeconomic status. Using US data, Grodsky et al. (2008) finds that standardized tests not only reflect existing racial/ethic inequalities, but reproduce them and have been shaping them since their widespread implementation. Education researchers have—among other contributions—identified what they call "sex bias" (Faggen-Steckler et al., 1974). This typically occurs when male nouns and pronouns are used in a test much more frequently than their female equivalents. Meanwhile, psychologists and anthropologists continue their efforts in understanding and attenuating cultural bias in cross-cultural studies (Broesch et al., 2020; Zeinoun et al.,

[2] Psychologists responded to this debate with the development of new frameworks aimed at avoiding such bias (Mercer, 1978).

2021)

Economists have largely taken standardized tests as given, and used them to measure the bias of other types of student evaluation (see Carlana, 2019, for an application to gender bias). That said, cultural aspects affecting student achievement have received some attention from economists. In particular, economists have been studying the relationship between cognitive ability and risk-preference (for a review, see Dohmen et al., 2018). Dohmen et al. (2018) shows how, since both of these variables are unobservable, this has proven to be difficult, and heroic assumptions are needed to claim causality. Moreover, causality could run in either direction, or both (Benjamin et al., 2013)—if cognitive ability and risk-preference reinforce each other.

On the one hand, cognitive ability could be influenced by risk-preferences and non-cognitive skills (Cunha et al., 2010). On the other hand, risk preferences could affect cognitive ability (Benjamin et al., 2013). It is from this relationship between risk and ability that recent work focuses on how culture indirectly affects student achievement via risk-preference and patience (Hanushek et al., 2020; Holmlund et al., 2021).

From a pure measurement standpoint, the most important contributions to the study of bias in standardized tests come from psychology and psychometrics. Scholars in these fields developed latent variable models first (Thissen, 2015) in the context of Classical Test Theory and, more recently, Item Response Theory.

In Item Response Theory, latent variable models are single- (Rasch, 1960) and multi-parameter logistic models. They are used to measure an unobservable feature of the examinee—a *latent construct*—analyzing the outcomes of standardized tests and test questions. They model the probability of answering a question correctly as a function of the latent construct (such as cognitive ability), and of the properties of the item (for good introductions, see Mellenbergh, 1989; Thissen, 2015).

These models can be interpreted causally (see Stenner et al., 2013; Rabbitt, 2018), but only under strong functional-form assumptions. Indeed, without an external (*exogenous*) source of variation, credible evidence of cultural bias cannot be produced. Randomized control trials (RCTs) are the best tool to address this issue since they

guarantee an exogenous source of variation by randomly splitting a sample of individuals into a "treatment" group and a "control" group.

The remainder of the paper proceeds as follows. Section 2 outlines our experimental design, including the context in which our intervention took place. Section 3 describes our data and empirical approach. Section 4, which is the heart of the paper, present the results of our experiment and explores possible mechanisms driving these results. Section 5 contains some brief concluding remarks.

2 Experimental Design

2.1 Context and Experimental Site

2.1.1 Education in Australia

The provision of public education in Australia is the responsibility of state and territory governments and their respective departments of education. Public schools, which provide free education from Kindergarten to Year 12, enroll approximately two-thirds of Australian students, with the remainder attending non-government schools.[3]

While state and territory governments are responsible for the regulation and provision of primary and secondary education within their jurisdictions, all schools in Australia are expected to follow a national curriculum set by the Australian Curriculum, Assessment and Reporting Authority (ACARA), which mandates specific achievement standards for students, regardless of background or location within Australia. To evaluate the achievement of these standards, ACARA administers an annual standardized test, the NAPLAN, to all students in Years 3, 5, 7, and 9.[4] NAPLAN tests, administered each May in all schools across the country, are intended to provide a snapshot of students' current abilities in reading, writing, language

[3]Non-government schools include independent schools connected to religious institutions, as well as schools driven by a pedagogical philosophy or specific needs. These schools receive funding from both state/territory governments and the federal government. All schools are regulated by their state/territory curriculum and assessment authorities.

[4]NAPLAN stands for National Assessment Program Literacy and Numeracy.

(spelling, grammar and punctuation), and numeracy.

ACARA sets national minimum standards for the NAPLAN, which represent the expected learning outcomes for students in each year level and subject. While the rationale for the national curriculum is premised on "improving the quality, equity, and transparency of the Australian education system", there are large disparities in the proportion of students from different backgrounds who meet the minimum standards in NAPLAN. For example, in the 2019 outcomes for Year 5 Numeracy, 85.9 percent of Indigenous students in the state of New South Wales were at or above the minimum standard versus 96.5 percent of non-Indigenous students.[5] The disparity is similar between students in remote areas and major cities, with 88.1 percent of remote students meeting the minimum standard versus 96.5 in major cities. In Reading, the Year 5 disparities were comparable, with 84.6 percent of Indigenous versus 95.9 percent of non-Indigenous students, and 85.7 percent of remote versus 95.9 percent of metro students at or above the minimum standard.

Australian policy makers have recently started to consider potential interventions to address these disparities. In New South Wales, the Department of Education, in cooperation with the Aboriginal Education Consultative Group (AECG), has focused specifically on improving the cultural relevance of curriculum for students of different backgrounds. Of particular concern is whether the design and delivery of assessments that are specific to the cultural context of students in remote and regional areas can improve their educational outcomes.

2.1.2 Dubbo

We conduct a randomized evaluation of a culturally contextualized test in Dubbo, New South Wales, a regional community approximately 400 kilometers north west of metropolitan Sydney. According to 2016 census figures, Dubbo is home to nearly 70,000 people, 10,500 of whom identify as Aboriginal and Torres Strait Islander background (Indigenous hereafter). Median weekly household income is $1,176 AUD (roughly $865 USD), with 2.5 people per household on average and a median age of 40. These figures are lower for Indigenous residents, with a median weekly household

[5]Indigenous students are those who identify as Aboriginal or Torres Strait Islander background.

income of $1,085 AUD, 3.1 people per household, and a median age of 22.[6] On average, 42 percent of non-Indigenous adults, and only 27 percent of Indigenous adults, in Dubbo have completed high school.[7]

Across New South Wales, there are 110 school networks, each led by a Director, with approximately 20 schools per network.[8] The schools in our intervention come from Macquarie Network and Mudgee Network, two adjacent school networks in the greater Dubbo area. There are 20 schools in the Macquarie network, consisting of 10 primary, 4 secondary, 1 central school (K-12), 1 distance education school (K-12), 3 schools for Specific Purposes, and 1 environmental education center, and 19 schools in the Mudgee Network, including 9 primary, 4 secondary, 5 central schools (K-12), and 1 environmental education center. In both of these networks, primary schools range in student size from 700 in larger schools to 20 in small rural schools. Secondary schools range from 2,000 students (across three campuses in the more densely populated town of Dubbo) to 300 students in smaller, more rural high schools. Central schools, which enroll students from Kindergarten through Year 12, range in size from 300 to 150 students.

The share of Indigenous students varies by school network and level. In the Macquarie network, 42 percent of primary and 38 percent of secondary students come from an Indigenous background. In the Mudgee network, 27 percent of primary and 24 percent of secondary students are from an Indigenous background. The slightly lower share of Indigenous secondary students in both networks reflects a more general disparity in secondary school retention rates between Indigenous and non-Indigenous students across Australia.

[6]Figures obtained at the Statistical Area (SA) 3 level of the 2016 Census for the Dubbo SA3. The SA3 is a statistical unit of analysis used by the Australian Bureau of Statistics, which is designed to capture regional data. Most SA3s have between 30,00 and 130,000 people (Australian Bureau of Statistics: https://abs.gov.au).

[7]For comparison, the median weekly household income of residents in Greater Sydney is $1,750 AUD, with 2.8 people per household, a median age of 36, and 62 percent high school completion rate (Sydney Greater Capital City Statistical Area, Australian Bureau of Statistics: https://abs.gov.au).

[8]School networks are akin to school districts in the US context, noting that the administration of school networks is centralized at the state level.

2.2 Intervention and sampling design

In 2020, the AECG designed a set of standardized tests in reading and numeracy for students in Year 5 (primary school) and Year 7 (secondary school) that used items and language determined to be culturally relevant for students living in regional and remote parts of Australia, and in particular, those from Indigenous backgrounds. The content of the tests draws directly from the 2019 NAPLAN exam, but with questions that incorporate the local context of regional Dubbo.

Our intervention implements these tests in Year 5 reading and numeracy, and Year 7 numeracy. The Year 5 numeracy test contains 39 questions, while the reading test contains five reading passages and 33 questions. The Year 7 numeracy test contains 38 questions. An independent psychometric validity check revealed that the main test constructs - i.e., content, gender bias, and reading load - were effectively identical between the two (NAPLAN and contextualized) test versions for both numeracy and reading.[9] Figure 1 provides an example of a Year 5 Numeracy question, with the original NAPLAN question on the left and the culturally contextualized question on the right. Similarly, Appendix Figures 1 and 2 show examples of Year 5 reading materials.

We implemented a randomized evaluation of the culturally contextualized tests in Term 1 of the 2021 academic year across 12 primary and 8 secondary schools in two adjacent school networks in the greater Dubbo area.[10] Because NAPLAN tests were not administered in 2020, due to disruption caused by the COVID-19 pandemic, we conducted the evaluation among Year 6 and Year 8 students in 2021 who ordinarily would have taken the NAPLAN in Years 5 and 7 the previous year, but did not.[11]

[9]The one exception was related to one of the six reading passages in the initial version of Year 6 reading test that was made more difficult in the contextualized test, yielding it an overall readability age of 13-14 compared to the NAPLAN version that had a readability age of 11-12. This test section was adjusted in the final reading test version (that was deployed to Year 6 students) as per assessor suggestions. The full psychometric report is available here.

[10]The academic calendar in New South Wales consists of four terms with two-week breaks between terms and a five-week break at the end of the year. In 2021, Term 1 ran from January 27 to April 1. Term 2 ran from April 19 to June 25; Term 3 from July 12 to September 12; and Term 4 from October 5 to December 17.

[11]NAPLAN tests are administered in May of each year, which, in 2020, coincided with a brief disruption in face-to-face learning in New South Wales as a result of the COVID-19 pandemic. In 2020, disruption to instruction was minimal in Australia given the extremely low prevalence of cases

In the treatment group, students received the culturally contextualized tests designed by the AECG. In the control group, students received tests in the same style as the NAPLAN. Treatment was randomly assigned at the student level from current enrollment lists, stratified by school and year level. We further stratified across two sub-samples of cultural background: (i) Indigenous, and (ii) non-Indigenous. In each sub-sample of each year in each school, half of the students were assigned to the treatment group and the other half were assigned to the control group. Table 1 illustrates the number of students randomly selected in each cell across years, subjects, and cultural background.

Tests were administered in Week 9 of Term 1 (the week of March 29, 2021).[12] Three weeks prior to the tests (i.e., on March 8, 2021), parents were notified of the study and provided consent for their child to participate in an opt-out design, consistent with standard NAPLAN procedures. No students opted-out, although on the day of the test an additional 7 students participated in the tests.[13]

In addition, some students did not attempt the tests on the day they were administered (non-compliers). Appendix Table 1 shows the breakdown of non-compliers as follows: 42 Indigenous and 41 non-Indigenous students in Year 6 reading, 40 Indigenous and 36 non-Indigenous students in Year 6 numeracy, and 93 Indigenous and 106 non-Indigenous students in Year 8 numeracy. Appendix Tables 2 and 3 show that this sample of non-compliers is balanced across most pre-intervention covariates for all years, subjects, and cultural backgrounds. Furthermore, we show in the following section that the final sample of students who did attempt the tests is balanced across all pre-intervention covariates (Tables 4 and 5).

Tests were administered by the teacher in each classroom, and for each test, stu-

throughout the country. In our sample, schools were restricted to home learning from March 24 to May 25 out of a total of 40 weeks of instruction in 2020.

[12]Schools administered the tests at their discretion during this week. For Year 6, the numeracy test was held on one day and the reading test was held on another, consistent with the style of the NAPLAN. Since treatment was randomized within school, our results should not be impacted by idiosyncrasies in the particular day of the week or order in which the tests were administered. We further control for school fixed effects in our main specifications to address unobserved heterogeneity at the school level.

[13]Of these 7 students, 4 non-Indigenous students were in Year 6 – 2 in the treatment group and 2 in control, and 3 students were in Year 8 – 1 Indigenous in the treatment, and 1 non-Indigenous and 1 Indigenous in control.

dents had 40 minutes to attempt all questions, consistent with the style in which the NAPLAN test is administered. At the end of the tests, students were asked to rate their perceived effort, as well as the relevance of the questions to their local experience, by drawing a line on a "speedometer" that went from 0 to 100 in increments of 10. Students were asked the following five subjective questions about their exam experience, which we explore as potential mechanisms:

1. *Actual effort*: Now think of the test you have just completed. How much effort did you put in?
2. *Potential effort*: How much effort would you have put in if the test counted towards your end of year school marks?
3. *Recognition*: Could you recognize people and places from your community in the examples used in the test questions?
4. *Relevance*: Were the examples used in the question relevant to you?
5. *Appreciation*: Would you like future tests and assignments to use local examples that were relevant to you?

Upon completion, the tests were retained by the New South Wales Department of Education. Department of Education staff marked the tests and calculated the number of correct answers, the number of incorrect answers, the number of questions that were not attempted, and the responses to the subjective questions at the end of the tests, for each student. An anonymized database was then shared with the research team for data analysis.

3 Data and Empirical Analysis

3.1 Descriptive statistics and balance checks

Table 2 presents the summary statistics for Year 6 students, separated by subject. On average, these students are 11.5 years old, evenly split between male and female. Almost 40 percent of students are from an Indigenous background, while over 95

percent speak English as their primary language at home. Parental education levels are relatively low; 16 percent of students have at least one parent with a Bachelor degree or higher, while for 25 percent of students, the highest level of education their parents have completed is secondary school or lower. Most students come from a lower socio-economic background, just slightly below the second quartile.

In the Year 6 numeracy test, out of the 39 questions in total, students answered 18.2 correctly, 17.7 incorrectly, and did not attempt to answer 3.2 questions. In the Year 6 reading exam, students answered 16.7 questions correctly, 13.6 incorrectly, and did not attempt to answer 1.7 questions out of the 33 in total. Self-reported levels of real and potential effort were high across both tests, with students reporting over 80 out of 100 for real effort and nearly 90 out of 100 for potential effort if the test counted toward their end of year grades. For both tests, students reported around 60 out of 100 for recognition of local context and relevance, and over 70 out of 100 for appreciation (i.e., they would like future tests and assignments to use local examples that are relevant to them).

Table 3 presents the summary statistics for Year 8 students. On average, these students are 13.5 years old and slightly male-skewed (55 percent). Indigenous students represent 38 percent of the sample, while 95 percent of students speak English as their primary language at home. Again, parental education levels are relatively low; 18 percent of students have a parent with a Bachelor degree or higher, while for 23 percent students, the highest level of education their parents have completed is secondary school or lower. These students also come from a socio-economic background that is just below the second quartile of the national distribution.

Out of the 38 questions in total for Year 8 numeracy, students answered 17.1 correctly, 17.9 incorrectly, and did not attempt to answer 2.0 questions. Again, self-reported levels of real and potential effort were relatively high, with students reporting over 75 out of 100 for real effort and 88 out of 100 for potential effort if the test counted toward their end of year grades. Levels of recognition and appreciation were slightly lower than in Year 6, with the average student reporting around 56 for recognition of local context, 51 for relevance, and 66 out of 100 for appreciation.

In Tables 4 and 5 we conduct balance tests across years, subjects, and cultural

backgrounds for each of the pre-intervention covariates reported in the summary statistics to verify that the treatment assignment is random. The exercise shows that treatment and control is balanced on all dimensions in each of the samples. These tables also show the disparities between Indigenous and non-Indigenous students. In Year 6, for example, 40 percent of Indigenous parents have completed secondary school or lower as the highest level of educational attainment, compared to 15 percent of non-Indigenous parents. Further, the socio-economic status of Indigenous students falls between the first and second quartile compared to non-Indigenous students who sit squarely in the second quartile.

3.2 Estimation strategy

We estimate the average treatment effect of the culturally contextualized tests by regressing test outcomes (Y_{is}), for student i in school s, on assignment to the treatment group (T_i). In all specifications, we include school fixed effects (μ_s) to control for school-level unobservables that may be correlated with test scores, and cluster standard errors (ε_{is}) at the school level. Given the small number of schools (12 primary and 8 secondary), we implement a Wild cluster bootstrap and report p-values in the main tables.

While pre-intervention covariates are balanced across treatment and control, we include student-level controls (X_i) in some specifications, including age, Indigenous background, parents' education level, socio-economic status (SES), whether the student comes from an English-speaking background, and previous NAPLAN scores, to verify coefficient stability. Socio-economic status (SES) and English-speaking background are missing for some students (19 students in Year 6 reading, 16 students in Year 6 numeracy, and 16 students in Year 8 numeracy), while previous NAPLAN scores are missing for roughly 13% of the sample (64 students in Year 6 reading, 66 students in Year 6 numeracy, and 96 students in Year 8 numeracy). There are various reasons why some students in our sample are missing prior NAPLAN information, including those who were absent on the day of the 2018 exam, those who were exempt or withdrawn from the 2018 exam, or those who could not be linked to the administrative data provided to the researchers by the NSW Department of Educa-

tion. In Tables 4 and 5, we show that the share of students for whom 2018 NAPLAN information is available is balanced across treatment and control, and among these students, prior NAPLAN scores are balanced across treatment and control, as well. Nonetheless, we run two specifications with controls: one excluding SES, English, and previous NAPLAN scores to maintain a sample that is comparable to the estimation without controls, and one including SES, English, and previous NAPLAN, which drops all observations for which this information is missing.

Our estimating equation is the following:

$$Y_{is} = \alpha + \beta T_i + \gamma X_i + \mu_s + \varepsilon_{is} \tag{1}$$

The estimated coefficient for β represents the average treatment effect (ATE) of the culturally contextualized tests. We estimate the ATE for three outcomes, all standardized to mean zero, standard deviation one: (i) number of correct answers, (ii) number of incorrect answers, and (iii) number of questions not attempted.

For each outcome, we estimate equation (1) separately for Year 6 reading, Year 6 numeracy, and Year 8 reading. Within each year-subject, we estimate the ATE for the pooled sample of Indigenous and non-Indigenous students, and separately for each cultural background sub-sample. The results are presented in the following section.

4 Results

4.1 Year 6 student performance

4.1.1 Reading

The first panel of Table 6 presents the results for Year 6 reading. We begin by estimating the effect on the pooled sample of Indigenous and non-Indigenous students. Columns (1)-(3) show a large increase in the number of correct answers, which remains stable as we add controls. The coefficients imply that the ATE of the culturally contextualized reading test is a 0.27 to 0.30 standard deviation increase in the num-

ber of correct answers, which is statistically significant at the 1 percent level in all specifications. In parallel, we find a 0.26 to 0.23 standard deviation reduction in the number of incorrect answers in columns (4)-(6), while in columns (7)-(9), there is no evidence that the contextualized test impacts the number of questions not attempted by students.

Next, we estimate the effect for Indigenous students, only. In columns (1)-(3), the ATE is a 0.3 standard deviation increase in the number of correct answers, which is statistically significant at the 5 percent level and robust to the inclusion of controls. The magnitude of this effect is slightly higher in the specification with full controls (0.38 standard deviations), which we interpret with caution due to the missing observations. In parallel, we find a 0.29 standard deviation reduction in the number of incorrect answers and no evidence that the contextualized test affects the number of questions not attempted.

Finally, we estimate the ATE for non-Indigenous students and again find a large and positive impact of the contextualized test on reading test scores. The coefficients are stable across specifications and imply a 0.24 standard deviation impact on the number of correct answers, significant at the 5 percent level. Similarly, the coefficients in columns (4)-(6) imply a 0.24 to 0.20 standard deviation reduction in the number of incorrect answers, with no impact on the number of questions not attempted in columns (7)-(9). Notably, when we compare the coefficients for Indigenous versus non-Indigenous, we cannot reject the null hypothesis that they are equal, suggesting that the ATE is statistically equivalent across groups.

These findings are qualitatively meaningful. First consider the pooled sample of Indigenous and non-Indigenous students, which we leverage to calculate lower bounds on the extent to which the contextualized reading tests help to close the rural-urban learning gap in New South Wales. If fact, the magnitude of the ATE suggests that implementing the contextualized reading tests would increase Dubbo reading scores by over 12 points on the NAPLAN scale, which corresponds to a 33 percent reduction in the rural-urban Year 5 reading gap in New South Wales.[14]

[14]To see this note that in the control group of the pooled sample, the mean number of correct answers is 16. The standard deviation is 5.1. Since the ATE for full sample is 0.27 sd, this would increase the number of correct questions to 17.4 and hence raises the NAPLAN score by 12 points.

We conduct a similar exercise for Indigenous students, which suggests even larger qualitative effects. In the Indigenous sample, the mean number of correct answers is 14.2, which sits in "Band 4" of the NAPLAN assessment scale and just meets the 2019 minimum standard for Year 5 reading.[15] The ATE suggests that implementing the contextualized test would increase reading scores for Indigenous students by nearly 25 points on the NAPLAN scale, pushing these students from "Band 4" to "Band 5", which exceeds the minimum standard for Year 5 reading in NAPLAN. Within sample, the ATE also suggests that the contextualized test closes the Indigenous to non-Indigenous reading gap by 50 percent.[16]

Finally, for non-Indigenous students in the control group, the mean number of correct answers is 17.2, which is above the minimum standard in Year 5 reading and corresponds to "Band 5" on the NAPLAN assessment scale. The ATE for this group suggests that the contextualized test increases reading scores by approximately 12 points on the NAPLAN scale, which is just shy of "Band 6".

4.1.2 Numeracy

The second panel of Table 6 presents the results for Year 6 numeracy. First we estimate the ATE for the pooled sample of Indigenous and non-Indigenous students. In columns (1)-(3) there is weak evidence that the contextualized test reduces the number of correct answers, but this effect is only statistically significant at the 10 percent level in the sample with full controls, which is missing observations for 72 students. In columns (4)-(6) there is no consistent evidence that the contextualized test increases the number of incorrect answers. The coefficients in columns (7)-(8) suggest that there is a 0.15 standard deviation increase in the number of questions not attempted, significant at the 10 percent level with Wild cluster bootstrapping, which corresponds to 0.83 fewer questions attempted by students in the treatment

Since the 2019 regional-urban gap was (514.5-479.1) = 35.4 NAPLAN points, this would close the gap by more than 33%.

[15]NAPLAN Bands correspond to ranges of numeric scores and go from Band 1 to Band 6. For Year 5, Bands 3 and lower are below the national minimum standard, Band 4 is at the national minimum standard, and Bands 5 and higher are above the national minimum standard.

[16]The mean Indigenous control score is 14.2, with a standard deviation of 5.1. Since the ATE is 0.3, which would increase score to 15.7 This closes the sample Indigenous to non-Indigenous gap from 3 correct answers to 1.5 correct answers (50%).

group. However, in the specification with full controls (and missing observations) the magnitude of this effect is diminished and no longer statistically significant.

Next, we estimate the effect for the group of Indigenous students. For each outcome, we find no evidence of an effect of the contextualized test on numeracy outcomes. The coefficients on the number of correct answers are positive, but small and statistically insignificant, while the coefficients on incorrect answers and questions not attempted are negative, small, and statistically insignificant, suggesting no impact of the contextualized numeracy test on Indigenous student outcomes.

Finally, we estimate the ATE for non-Indigenous students. In the specification with partial controls there is weak evidence that the contextualized numeracy test reduces the number of correct answers by 0.17 standard deviations, which becomes larger in magnitude in the sample with full controls, but with missing observations for 44 students. There is no evidence that the contextualized test increases the number of incorrect answers, but there is strong and robust evidence that it increases the number of questions not attempted. In the worst case scenario, the coefficients in columns (7) and (8) suggest a 0.29 standard deviation increase in the number of questions not attempted, statistically significant at the 1 percent level in all specifications, which corresponds to 2.3 fewer questions attempted by non-Indigenous students who received the contextualized numeracy test.

This effect has qualitative implications if the reduction in number of questions attempted leads to lower overall scores for non-Indigenous students. The mean number of correct answers for non-Indigenous students in the control group was 20.9, which is above the 2019 minimum standard for Year 5 numeracy and above the mean for other comparable regional areas of NSW. The upper-bound ATE that we observe in column (3) suggests that the contextualized numeracy test reduces the number of correct answers to 19.12, which corresponds to a 14 point reduction in the NAPLAN scale. The overall score is still above the minimum standard and remains in the same assessment Band ("Band 5") as the control group, but the reduction in points on the NAPLAN scale would place these students below the mean for other comparable regional areas of New South Wales.

4.2 Year 8 student performance

4.2.1 Numeracy

Table 7 presents the results for Year 8 numeracy. We begin with the pooled sample of Indigenous and non-Indigenous students, which shows no evidence that the contextualized test impacts numeracy test outcomes. In all specifications, for all outcomes, the coefficients are small in magnitude with large standard errors. For Indigenous students, there is some evidence that the contextualized test increases the number of correct answers, but the magnitude and statistical significance declines as we include controls. Further, there is no evidence that the test impacts the number of incorrect answers or questions not attempted by Indigenous students. Finally, we find no evidence that the test impacts outcomes for non-Indigenous students. In all specifications, for all outcomes, the standard errors are large relative to the coefficients, which are small in magnitude. Together, these results suggest, qualitatively, minimal benefit of the contextualized Year 8 numeracy test, but little harm, as well.

4.3 Possible mechanisms: effort, recognition, and relevance

In this section we examine potential mechanisms by utilizing the survey responses on perceived effort and relevance of the exam. In Table 8 we explore Year 6 responses, which are standardized to mean zero, standard deviation one. The first panel suggests that for reading, in the pooled sample of Indigenous and non-Indigenous students, there is weak evidence that effort is slightly lower, and some evidence that recognition and relevance of local context are higher, in the treatment group. The coefficients on recognition and relevance correspond to 5.6 and 3.6 point increases, respectively, on the subjective "speedometer" scale, which is qualitatively meaningful relative to control means of 55 (out of 100) for both responses. When we separate the sample by Indigenous and non-Indigenous students, we find consistent responses, but the standard errors are much larger. There is weak evidence that Indigenous students in the treatment group reduce effort slightly; the coefficient corresponds to a 3.2 point reduction in subjective effort, which relative to a control mean of 82.4 is qualitatively small.

In the second panel of Table 8 we explore Year 6 numeracy responses. In the pooled, as well as the separate Indigenous and non-Indigenous samples, students who received the contextualized test report greater recognition of local context. Indigenous students also report greater relevance. The pooled sample coefficients suggest an 18 and 7 point increase in recognition and relevance, which relative to control means of 53.4 and 56.3, respectively, are qualitatively meaningful. In the non-Indigenous sample, there is weak evidence that students in the treatment group exert less effort, which is consistent with the finding that they attempted fewer questions.

In Table 9 we explore Year 8 numeracy responses. Again, we find large increases in recognition and relevance across all samples. In the pooled sample for Indigenous and non-Indigenous students, the coefficients suggest a 17.5 and 6.5 point increase in recognition and relevance, which correspond to 33 and 11.5 percent increases from the control mean, respectively. There is no evidence that the contextualized test reduces effort. In the non-Indigenous group, the coefficient is negative, but with large standard errors.

Taken together these results imply that recognition improves in the treatment group for all samples, and relevance improves in most samples. But the results offer no clear rationale for why the treatment "works" for reading, but not numeracy. It seems that whatever underlying mechanism drives the treatment effect is not well captured by self-reported measures, suggesting true mechanism may not be readily apparent to students themselves.

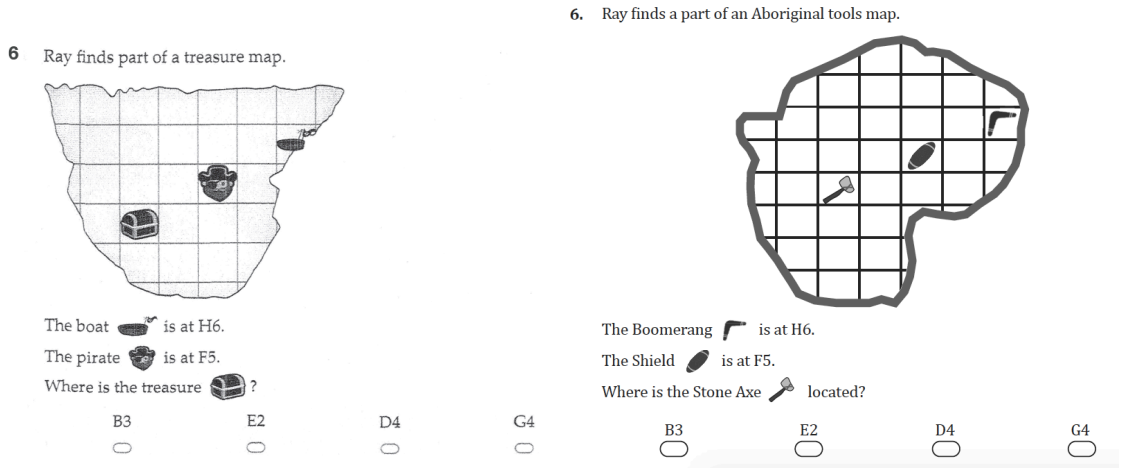
5 Conclusion

Our experiment shows that cultural context in standardized test can matter. Specifically, in our treatments there is an economically large and statistically significant impact of contextualized test on reading test scores, and no robust evidence of an effect on numeracy test score.

Because the culturally contextualized test we administered was specifically designed to be identical in all respects other than those pertaining to item bias, our results provide causal evidence on the extent of such bias in our educational setting.

It has not escaped our attention that our work has implications for the design of educational materials, in addition to standardized tests. Given the large impact of cultural context on assessment, we conjecture that adapting educational materials such as textbooks, slides, and multi-media content, to students' cultural context will have a meaningful impact on educational outcomes, at least in some subjects.

A proper understanding of the effect of culturally-contextualized educational materials on educational outcomes would be highly desirable. A further RCT focused on precisely this issue is an enticing prospect for future work.



(a) NAPLAN (b) Culturally relevant

Figure 1 – Example questions: Year 5 Numeracy

Table 1 – Sampling design

	Indigenous		non-Indigenous	
	Control	Treatment	Control	Treatment
Year 6 Reading:	134	129	193	188
Year 6 Numeracy:	134	129	193	188
Year 8 Numeracy:	150	146	220	216

Table 2 – Summary Statistics: Year 6

	Numeracy Year 6					Reading Year 6				
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
Age	570	11.562	0.367	10.750	12.833	565	11.563	0.368	10.750	12.833
Male	570	0.523	0.500	0	1	565	0.522	0.500	0	1
Indigenous background	570	0.391	0.488	0	1	565	0.391	0.488	0	1
English-speaking background	554	0.955	0.208	0	1	546	0.954	0.209	0	1
High parent education	570	0.163	0.370	0	1	565	0.163	0.370	0	1
Low parent education	570	0.240	0.428	0	1	565	0.244	0.430	0	1
Student SES quartile	554	1.917	0.968	1	4	546	1.918	0.969	1	4
Prior Naplan Numeracy presence	525	0.960	0.196	0	1	517	0.959	0.198	0	1
Prior Naplan Numeracy score band	504	3.304	1.252	1	6	496	3.280	1.249	1	6
Prior Naplan Reading presence	525	0.970	0.172	0	1	517	0.969	0.173	0	1
Prior Naplan Reading score band	509	3.631	1.523	1	6	501	3.613	1.521	1	6
Number of correct answers	570	18.153	8.079	1	39	565	16.717	5.209	3	31
Number of incorrect answers	570	17.658	8.369	0	38	565	13.630	5.457	1	28
Number of questions not attempted	570	3.189	5.561	0	37	565	1.653	3.804	0	25
Amount of effort exerted	490	84.092	19.299	10	100	485	83.340	18.033	10	100
Amount of effort potentially exerted	483	89.358	19.426	10	100	483	90.455	17.682	0	100
Level of context recognition in test	477	62.264	30.457	0	100	479	58.956	29.097	0	100
Level of example relevance in test	486	59.805	28.855	0	100	480	56.771	27.135	0	100
Level of appreciation for local context	484	71.694	29.482	0	100	479	72.902	27.862	0	100

Note. There were 39 total questions in the numeracy test and 33 total questions in the reading test.

Table 3 – Summary Statistics: Year 8

	Numeracy Year 8				
Variable	Obs	Mean	Std. Dev.	Min	Max
Age	536	13.576	0.375	12.750	14.917
Male	536	0.554	0.498	0	1
Indigenous background	536	0.382	0.486	0	1
English-speaking background	530	0.949	0.220	0	1
High parent education	536	0.183	0.387	0	1
Low parent education	536	0.233	0.423	0	1
Student SES quartile	528	1.962	0.981	1	4
Prior Naplan Numeracy presence	460	0.957	0.204	0	1
Prior Naplan Numeracy score band	440	5.214	1.190	3	8
Prior Naplan Reading presence	460	0.976	0.153	0	1
Prior Naplan Reading score band	449	5.272	1.480	3	8
Number of correct answers	536	17.067	6.935	1	36
Number of incorrect answers	536	17.929	6.801	0	34
Number of questions not attempted	536	2.004	3.839	0	36
Amount of effort exerted	481	76.143	20.416	0	100
Amount of effort potentially exerted	482	88.143	18.231	10	100
Level of context recognition in test	475	56.200	28.395	0	100
Level of example relevance in test	482	51.120	26.959	0	100
Level of appreciation for local context	483	66.077	26.614	0	100

Note. There were 38 total questions in the numeracy test.

Table 4 – Balancing Tests: Year 6

	PANEL A - Numeracy Year 6						PANEL B - Reading Year 6					
	Treatment		Control		Difference		Treatment		Control		Difference	
Variable	Mean	SD	Mean	SD	C-T	SE	Mean	SD	Mean	SD	C-T	SE
All students												
Age	11.556	0.378	11.569	0.357	0.014	0.031	11.557	0.376	11.570	0.360	0.013	0.031
Male	0.498	0.501	0.547	0.499	0.049	0.042	0.495	0.501	0.549	0.499	0.054	0.042
Indigenous background	0.384	0.487	0.398	0.490	0.014	0.041	0.391	0.489	0.392	0.489	0.001	0.041
English speaking	0.945	0.228	0.964	0.186	0.019	0.018	0.945	0.229	0.964	0.188	0.019	0.018
High parent education	0.164	0.371	0.163	0.370	−0.001	0.031	0.158	0.365	0.168	0.374	0.010	0.031
Low parent education	0.246	0.431	0.235	0.425	−0.010	0.036	0.258	0.438	0.231	0.422	−0.027	0.036
Student SES quartile	1.909	0.992	1.925	0.946	0.016	0.082	1.889	0.983	1.946	0.956	0.056	0.083
Prior Naplan Numeracy presence	0.973	0.163	0.947	0.224	−0.026	0.017	0.973	0.163	0.946	0.226	−0.026	0.017
Prior Naplan Numeracy score band	3.337	1.250	3.270	1.256	−0.068	0.112	3.317	1.267	3.243	1.232	−0.074	0.112
Prior Naplan Reading presence	0.969	0.173	0.970	0.171	0.001	0.015	0.969	0.174	0.969	0.173	0.001	0.015
Prior Naplan Reading score band	3.665	1.570	3.597	1.479	−0.068	0.135	3.629	1.574	3.597	1.470	−0.032	0.136
Indigenous students, only												
Age	11.529	0.376	11.529	0.375	0.000	0.050	11.538	0.379	11.522	0.379	−0.015	0.051
Male	0.491	0.502	0.574	0.497	0.083	0.067	0.495	0.502	0.580	0.496	0.085	0.067
English speaking	0.962	0.192	0.991	0.095	0.029	0.020	0.962	0.192	0.991	0.096	0.029	0.021
High parent education	0.065	0.247	0.078	0.270	0.013	0.035	0.055	0.229	0.080	0.273	0.025	0.034
Low parent education	0.407	0.494	0.383	0.488	−0.025	0.066	0.413	0.495	0.375	0.486	−0.038	0.066
Student SES quartile	1.486	0.774	1.607	0.798	0.121	0.107	1.462	0.758	1.620	0.806	0.158	0.107
Prior Naplan Numeracy presence	0.970	0.171	0.918	0.275	**-0.0521**	**0.0319**	0.961	0.195	0.925	0.265	−0.036	0.032
Prior Naplan Numeracy score band	2.949	1.143	2.832	1.105	−0.117	0.159	2.959	1.148	2.806	1.081	−0.153	0.159
Prior Naplan Reading presence	0.970	0.171	0.964	0.188	−0.007	0.025	0.961	0.195	0.962	0.192	0.002	0.027
Prior Naplan Reading score band	3.163	1.497	3.217	1.359	0.054	0.200	3.153	1.502	3.177	1.367	0.023	0.203
Non-Indigenous students, only												
Age	11.572	0.379	11.595	0.342	0.023	0.039	11.570	0.375	11.600	0.345	0.031	0.039
Male	0.503	0.501	0.529	0.501	0.026	0.054	0.494	0.501	0.529	0.501	0.035	0.054
English speaking	0.935	0.247	0.946	0.226	0.012	0.026	0.933	0.250	0.946	0.227	0.013	0.026
High parent education	0.225	0.419	0.218	0.414	−0.007	0.045	0.224	0.418	0.224	0.418	0.001	0.045
Low parent education	0.145	0.353	0.138	0.346	−0.007	0.038	0.159	0.367	0.138	0.346	−0.021	0.038
Student SES quartile	2.172	1.024	2.137	0.978	−0.035	0.109	2.164	1.014	2.156	0.988	−0.008	0.110
Prior Naplan Numeracy presence	0.975	0.158	0.968	0.177	−0.007	0.019	0.981	0.139	0.961	0.194	−0.019	0.019
Prior Naplan Numeracy score band	3.584	1.256	3.563	1.268	−0.022	0.145	3.550	1.289	3.530	1.244	−0.020	0.146
Prior Naplan Reading presence	0.968	0.176	0.974	0.159	0.006	0.019	0.974	0.160	0.974	0.159	0.000	0.018
Prior Naplan Reading score band	3.987	1.535	3.862	1.505	−0.125	0.174	3.940	1.547	3.881	1.474	−0.059	0.174

Table 5 – Balancing Tests: Year 8

	PANEL C - Numeracy Year 8					
	Treatment		Control		Difference	
Variable	Mean	SD	Mean	SD	C-T	SE
All students						
Age	13.578	0.388	13.574	0.364	−0.004	0.032
Male	0.573	0.496	0.537	0.500	−0.035	0.043
Indigenous background	0.400	0.491	0.367	0.483	−0.034	0.042
English speaking	0.956	0.205	0.943	0.233	−0.014	0.019
High parent education	0.169	0.375	0.196	0.398	0.027	0.034
Low parent education	0.235	0.425	0.231	0.422	−0.004	0.037
Student SES quartile	1.952	0.987	1.971	0.978	0.019	0.086
Prior Naplan Numeracy presence	0.951	0.217	0.962	0.192	0.011	0.019
Prior Naplan Numeracy score band	5.211	1.173	5.216	1.209	0.005	0.114
Prior Naplan Reading presence	0.978	0.148	0.975	0.158	−0.003	0.014
Prior Naplan Reading score band	5.247	1.494	5.296	1.469	0.049	0.140
Indigenous students, only						
Age	13.578	0.396	13.525	0.347	−0.053	0.052
Male	0.578	0.496	0.466	0.501	−0.112	0.070
English speaking	0.990	0.100	0.961	0.194	−0.029	0.022
High parent education	0.108	0.312	0.078	0.269	−0.030	0.041
Low parent education	0.294	0.458	0.359	0.482	0.065	0.066
Student SES quartile	1.690	0.961	1.631	0.792	−0.059	0.123
Prior Naplan Numeracy presence	0.921	0.272	0.967	0.180	0.047	0.034
Prior Naplan Numeracy score band	4.901	1.136	4.773	1.014	−0.129	0.165
Prior Naplan Reading presence	0.977	0.150	0.978	0.147	0.001	0.022
Prior Naplan Reading score band	4.802	1.404	4.719	1.297	−0.083	0.204
Non-Indigenous students, only						
Age	13.577	0.383	13.602	0.371	0.025	0.042
Male	0.569	0.497	0.579	0.495	0.010	0.055
English speaking	0.934	0.250	0.932	0.253	−0.002	0.028
High parent education	0.209	0.408	0.264	0.442	0.055	0.047
Low parent education	0.196	0.398	0.157	0.365	−0.039	0.042
Student SES quartile	2.126	0.968	2.172	1.022	0.047	0.111
Prior Naplan Numeracy presence	0.971	0.170	0.959	0.200	−0.012	0.022
Prior Naplan Numeracy score band	5.402	1.158	5.496	1.242	0.095	0.146
Prior Naplan Reading presence	0.978	0.147	0.972	0.164	−0.006	0.019
Prior Naplan Reading score band	5.534	1.485	5.660	1.458	0.126	0.178

Note. There were 38 total questions in the numeracy test.

Table 6 – Results: Year 6

	Std # Correct			Std # Incorrect			Std # Not-attempted		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reading Year 6									
All students									
Treatment	0.271***	0.274***	0.298***	-0.259***	-0.265***	-0.232***	0.001	0.006	-0.074
	(0.061)	(0.057)	(0.049)	(0.055)	(0.053)	(0.062)	(0.079)	(0.076)	(0.064)
Wild *p*-val	0.00	0.00	0.00	0.00	0.00	0.01	0.99	0.94	0.27
N	565	565	495	565	565	495	565	565	495
Indigenous students									
Treatment	0.314**	0.305**	0.383***	-0.291**	-0.285***	-0.235	-0.011	-0.010	-0.187
	(0.137)	(0.127)	(0.115)	(0.108)	(0.089)	(0.134)	(0.198)	(0.194)	(0.184)
Wild *p*-val	0.06	0.05	0.00	0.02	0.01	0.12	0.95	0.96	0.35
N	221	221	196	221	221	196	221	221	196
Non-Indigenous students									
Treatment	0.235***	0.239**	0.236**	-0.234***	-0.239***	-0.204***	0.014	0.016	-0.030
	(0.075)	(0.078)	(0.077)	(0.072)	(0.074)	(0.060)	(0.053)	(0.044)	(0.065)
Wild *p*-val	0.03	0.04	0.03	0.03	0.02	0.02	0.79	0.71	0.66
N	344	344	299	344	344	299	344	344	299
Numeracy Year 6									
All students									
Treatment	-0.078	-0.076	-0.127*	-0.024	-0.029	0.075*	0.150**	0.153**	0.072
	(0.060)	(0.044)	(0.061)	(0.062)	(0.049)	(0.039)	(0.064)	(0.065)	(0.061)
Wild *p*-val	0.24	0.13	0.08	0.74	0.66	0.07	0.06	0.07	0.25
N	570	570	498	570	570	498	570	570	498
Indigenous students									
Treatment	0.041	0.048	0.025	-0.020	-0.011	0.080	-0.029	-0.053	-0.157
	(0.113)	(0.105)	(0.096)	(0.101)	(0.096)	(0.079)	(0.166)	(0.172)	(0.153)
Wild *p*-val	0.75	0.68	0.79	0.86	0.91	0.31	0.87	0.77	0.38
N	223	223	195	223	223	195	223	223	195
Non-Indigenous students									
Treatment	-0.177	-0.174*	-0.216***	-0.019	-0.025	0.080	0.286***	0.289***	0.194**
	(0.115)	(0.091)	(0.064)	(0.115)	(0.098)	(0.070)	(0.056)	(0.061)	(0.074)
Wild *p*-val	0.17	0.10	0.03	0.89	0.83	0.28	0.00	0.00	0.01
N	347	347	303	347	347	303	347	347	303
Controls:	No	Partial	Full	No	Partial	Full	No	Partial	Full

* p< 0.10, ** p<0.05, *** p < 0.01. Unit of observation is a student in the indicated subject and grade. Standard errors clustered at the school level in parentheses. Wild cluster bootstrap *p*-values calculated over 999 repetitions using the Stata *boottest* command with Webb weights. All estimates control for school fixed effects. Partial controls include age, Indigenous background, and parents' education level. Full controlls add socio-economic status, whether the student comes from an English-speaking background, and previous NAPLAN score.

Table 7 – Results: Year 8

	Std # Correct			Std # Incorrect			Std # Not-attempted		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Numeracy Year 8									
All students									
Treatment	0.039	0.058	-0.016	-0.040	-0.058	-0.004	0.000	-0.002	0.036
	(0.077)	(0.070)	(0.069)	(0.077)	(0.070)	(0.043)	(0.058)	(0.060)	(0.079)
Wild p-val	0.60	0.40	0.82	0.60	0.40	0.93	1.00	0.97	0.66
N	536	536	439	536	536	439	536	536	439
Indigenous students									
Treatment	0.154**	0.099	0.051	-0.082	-0.025	-0.026	-0.133	-0.136	-0.047
	(0.062)	(0.066)	(0.068)	(0.101)	(0.089)	(0.077)	(0.135)	(0.131)	(0.151)
Wild p-val	0.12	0.25	0.47	0.46	0.82	0.72	0.42	0.47	0.74
N	205	205	169	205	205	169	205	205	169
Non-Indigenous students									
Treatment	0.006	0.036	-0.053	-0.044	-0.065	-0.006	0.067	0.050	0.107
	(0.106)	(0.089)	(0.050)	(0.098)	(0.090)	(0.024)	(0.073)	(0.063)	(0.088)
Wild p-val	0.95	0.65	0.45	0.64	0.46	0.79	0.50	0.56	0.37
N	331	331	270	331	331	270	331	331	270
Controls:	No	Partial	Full	No	Partial	Full	No	Partial	Full

* p< 0.10, ** p<0.05, *** p < 0.01. Unit of observation is a student in the indicated subject and grade. Standard errors clustered at the school level in parentheses. Wild cluster bootstrap p-values calculated over 999 repetitions using the Stata *boottest* command with Webb weights. All estimates control for school fixed effects. Partial controls include age, Indigenous background, and parents' education level. Full controlls add socio-economic status, whether the student comes from an English-speaking background, and previous NAPLAN score.

Table 8 – Survey questions, beyond academic performance: Year 6

	Actual effort (1)	Potential effort (2)	Recognition (3)	Relevance (4)	Appreciation (5)
Reading Year 6					
All students					
Treatment	-0.082*	0.027	0.193**	0.142*	0.120
	(0.044)	(0.076)	(0.083)	(0.076)	(0.080)
Wild p-val	0.08	0.72	0.04	0.09	0.20
N	485	483	479	480	479
Indigenous students					
Treatment	-0.166*	-0.027	0.119	0.165	0.167
	(0.077)	(0.151)	(0.161)	(0.097)	(0.135)
Wild p-val	0.05	0.87	0.48	0.13	0.25
N	178	176	177	178	178
Non-Indigenous students					
Treatment	-0.053	0.041	0.226	0.113	0.091
	(0.076)	(0.096)	(0.145)	(0.092)	(0.114)
Wild p-val	0.51	0.68	0.17	0.27	0.44
N	307	307	302	302	301
Numeracy Year 6					
All students					
Treatment	-0.086	0.031	0.599***	0.248**	0.121
	(0.082)	(0.092)	(0.127)	(0.097)	(0.118)
Wild p-val	0.30	0.74	0.00	0.04	0.35
N	490	483	477	486	484
Indigenous students					
Treatment	0.055	0.101	0.627***	0.484***	0.188
	(0.152)	(0.181)	(0.184)	(0.081)	(0.178)
Wild p-val	0.72	0.70	0.01	0.00	0.32
N	187	181	178	185	182
Non-Indigenous students					
Treatment	-0.187*	-0.013	0.571***	0.088	0.071
	(0.099)	(0.087)	(0.106)	(0.111)	(0.112)
Wild p-val	0.09	0.91	0.00	0.47	0.56
N	303	302	299	301	302

* p< 0.10, ** p<0.05, *** p < 0.01. Unit of observation is a student in the indicated subject and grade. Standard errors clustered at the school level in parentheses. Wild cluster bootstrap p-values calculated over 999 repetitions using the Stata *boottest* command with Webb weights. All estimates control for school fixed effects.

Table 9 – Survey questions, beyond academic performance: Year 8

	Actual effort (1)	Potential effort (2)	Recognition (3)	Relevance (4)	Appreciation (5)
Numeracy Year 8					
All students					
Treatment	-0.094	0.072	0.656***	0.242**	0.036
	(0.104)	(0.047)	(0.094)	(0.090)	(0.079)
Wild p-val	0.50	0.16	0.01	0.05	0.73
N	481	482	475	482	483
Indigenous students					
Treatment	-0.007	0.222	0.616***	0.279*	0.166
	(0.192)	(0.128)	(0.092)	(0.119)	(0.161)
Wild p-val	0.98	0.20	0.00	0.10	0.41
N	183	182	178	181	181
Non-Indigenous students					
Treatment	-0.139	0.018	0.685***	0.211**	-0.023
	(0.082)	(0.058)	(0.123)	(0.081)	(0.030)
Wild p-val	0.16	0.76	0.02	0.08	0.45
N	298	300	297	301	302

* p< 0.10, ** p<0.05, *** p < 0.01. Unit of observation is a student in the indicated subject and grade. Standard errors clustered at the school level in parentheses. Wild cluster bootstrap p-values calculated over 999 repetitions using the Stata *boottest* command with Webb weights. All estimates control for school fixed effects.

References

Ackerman, Terry A, "A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective," *Journal of educational measurement*, 1992, *29* (1), 67–91.

Benjamin, Daniel J, Sebastian A Brown, and Jesse M Shapiro, "Who is 'behavioral'? Cognitive ability and anomalous preferences," *Journal of the European Economic Association*, 2013, *11* (6), 1231–1255.

Binet, Alfred and Theophile Simon, *The development of intelligence in children.*, Baltimore: Williams & Wilkins Company, 1916.

Broesch, Tanya, Alyssa N Crittenden, Bret A Beheim, Aaron D Blackwell, John A Bunce, Heidi Colleran, Kristin Hagel, Michelle Kline, Richard McElreath, Robin G Nelson et al., "Navigating cross-cultural research: methodological and ethical considerations," *Proceedings of the Royal Society B*, 2020, *287* (1935), 20201245.

Carlana, Michela, "Implicit stereotypes: Evidence from teachers' gender bias," *The Quarterly Journal of Economics*, 2019, *134* (3), 1163–1224.

Cunha, Flavio, James J Heckman, and Susanne M Schennach, "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 2010, *78* (3), 883–931.

de Vijver, Fons Van and Norbert K Tanzer, "Bias and equivalence in cross-cultural assessment: An overview," *European Review of Applied Psychology*, 2004, *54* (2), 119–135.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde, "On the relationship between cognitive ability and risk preference," *Journal of Economic Perspectives*, 2018, *32* (2), 115–34.

Faggen-Steckler, Jane, Karen A McCarthy, and Carol K Tittle, "A Quantitative Method For Measuring Sex "Bias" in Standardized Tests," *Journal of Educational Measurement*, 1974, *11* (3), 151–161.

Gallagher, Carole J, "Reconciling a tradition of testing with a new learning paradigm," *Educational Psychology Review*, 2003, *15* (1), 83–99.

Grodsky, Eric, John Robert Warren, and Erika Felts, "Testing and social stratification in American education," *Annual Review Sociology*, 2008, *34*, 385–404.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann, "Culture and student achievement: The intertwined roles of patience and risk-taking," Working Paper, National Bureau of Economic Research 2020.

Holmlund, Helena, Helmut Rainer, Patrick Reich et al., "All geared towards success? Cultural origins of gender gaps in student achievement," Working Paper, IFAU - Institute for Evaluation of Labour Market and Education Policy 2021.

Lechner, Michael, "A note on the common support problem in applied evaluation studies," *Annales d'Économie et de Statistique*, 2008, pp. 217–235.

Mellenbergh, Gideon J, "Item bias and item response theory," *International journal of educational research*, 1989, *13* (2), 127–143.

Mercer, Jane R, "Test validity, bias, and fairness: An analysis from the perspective of the sociology of knowledge.," *Interchange*, 1978.

Rabbitt, Matthew P, "Causal inference with latent variables from the Rasch model as outcomes," *Measurement*, 2018, *120*, 193–205.

Rasch, Georg, *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*, Nielsen & Lydiche, 1960.

Raspberry, William, "Standardized tests and cultural bias," *The Washington Post*, 1974, p. A15. September 4.

Reynolds, Cecil R. and Lisa A. Suzuki, *Bias in Psychological Assessment*, American Cancer Society, 2012.

Saboe, Matt and Sabrina Terrizzi, "SAT optional policies: Do they influence graduate quality, selectivity or diversity?," *Economics Letters*, 2019, *174*, 13–17.

Stenner, A Jackson, William P Fisher Jr, Mark Stone, and Donald Burdick, "Causal Rasch models," *Frontiers in psychology*, 2013, *4*, 536.

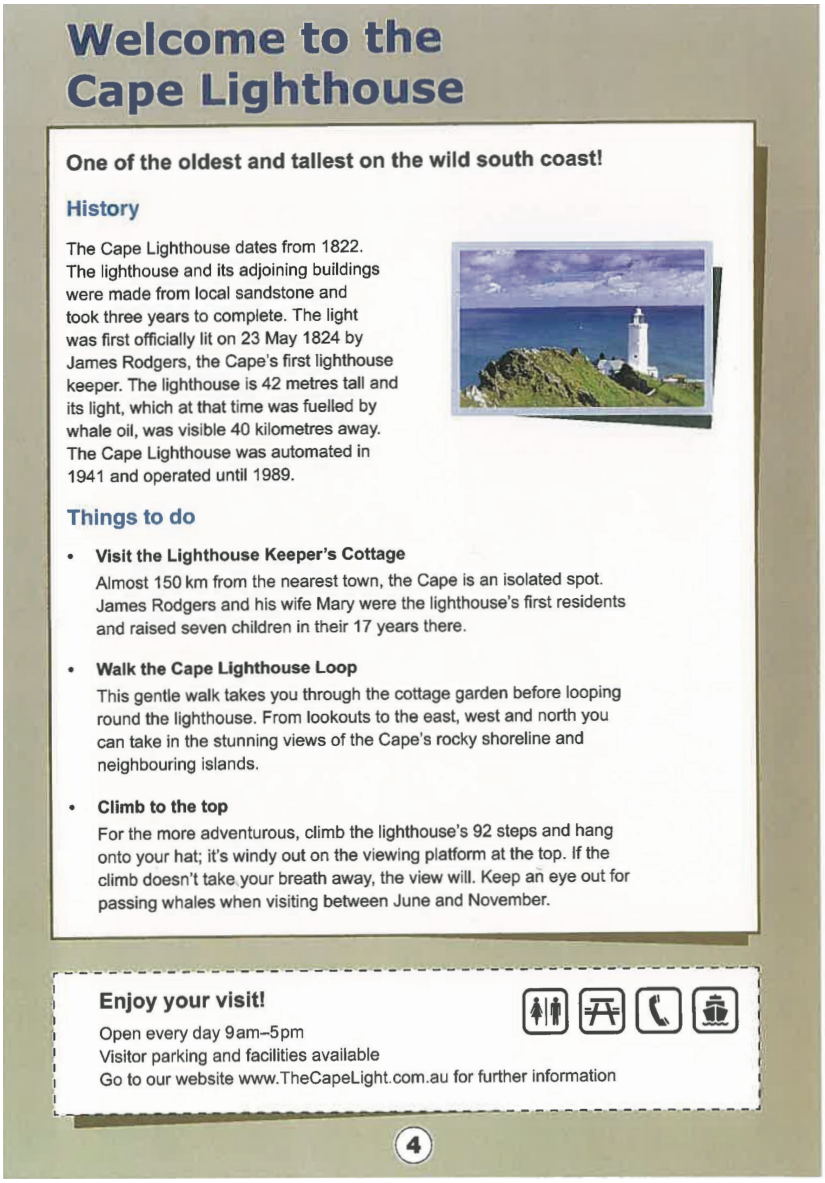
Thissen, David, "Psychometrics: Item Response Theory," in James D. Wright, ed., *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, second edition ed., Oxford: Elsevier, 2015, pp. 436–439.

Triandis, Harry C, "Introduction to diversity in clinical psychology," in AS Bellack and M Hersen, eds., *Comprehensive Clinical Psychology [Volume 10]*, first edition ed., Oxford: Elsevier, 2000, pp. 1–33.

Zeinoun, Pia, Dragos Iliescu, and Rhawann El Hakim, "Psychological tests in Arabic: A review of methodological practices and recommendations for future use," *Neuropsychology Review*, 2021, pp. 1–19.

Appendices

A Additional Figures and Tables



Welcome to the Cape Lighthouse

One of the oldest and tallest on the wild south coast!

History

The Cape Lighthouse dates from 1822. The lighthouse and its adjoining buildings were made from local sandstone and took three years to complete. The light was first officially lit on 23 May 1824 by James Rodgers, the Cape's first lighthouse keeper. The lighthouse is 42 metres tall and its light, which at that time was fuelled by whale oil, was visible 40 kilometres away. The Cape Lighthouse was automated in 1941 and operated until 1989.

Things to do

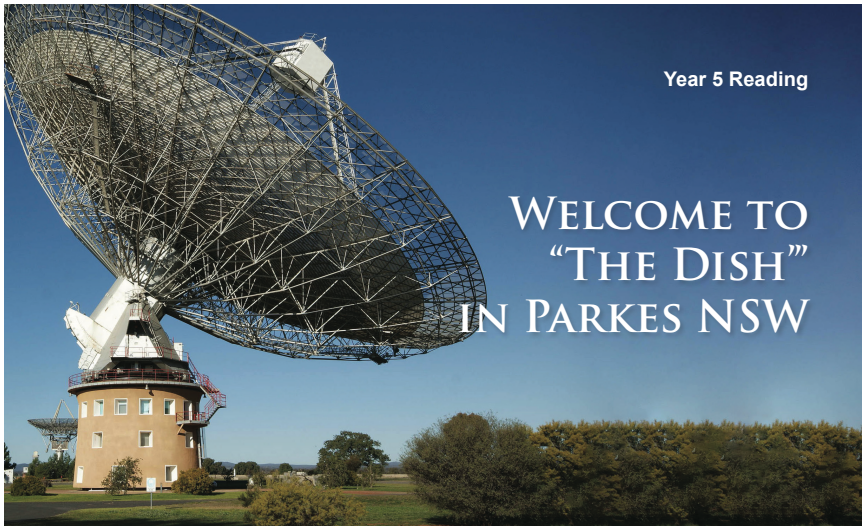
- **Visit the Lighthouse Keeper's Cottage**
  Almost 150 km from the nearest town, the Cape is an isolated spot. James Rodgers and his wife Mary were the lighthouse's first residents and raised seven children in their 17 years there.
- **Walk the Cape Lighthouse Loop**
  This gentle walk takes you through the cottage garden before looping round the lighthouse. From lookouts to the east, west and north you can take in the stunning views of the Cape's rocky shoreline and neighbouring islands.
- **Climb to the top**
  For the more adventurous, climb the lighthouse's 92 steps and hang onto your hat; it's windy out on the viewing platform at the top. If the climb doesn't take your breath away, the view will. Keep an eye out for passing whales when visiting between June and November.

Enjoy your visit!

Open every day 9am–5pm
Visitor parking and facilities available
Go to our website www.TheCapeLight.com.au for further information

4

Appendix Figure 1 – Year 5 Reading: NAPLAN



History

The CSRIO Observatory (The Dish) dates from 1961 and is still in use today.

The Dish has a diameter of 64 metres and is one of the largest single-dish telescopes in the southern hemisphere. It took three years to design and two years to build.

First opened on the 31st October 1961, The Dish is now 10 000 times more responsive than when it was first built. In 2012 The Dish helped to track NASA's Curiosity Rover during its landing on the surface of Mars.

The Dish is located in Parkes on an isolated spot.

Things to do

- Discover the Universe in the high definition 3D Theatre
  The theatre shows a variety of short 3D films. These animated features give a glimpse of the wonders of our Universe.
- Challenge yourself with the Astrokids Scavenger Hunt
  This fun activity takes you on a tour of the visitors centre looking for clues to solve a puzzle. Find the secret word and you can collect the official Astrokids stamp.
- Taste the delights of the region in the award winning Dish Café.
  The Dish cafe is open for breakfast and lunch every day. It also serves great coffee and hot scones.

Enjoy your visit!

Open every day 8:30am to 4:15pm
Visitor parking and facilities available
Go to our website https://www.csiro.au/en/Locations/NSW/Parkes for further information.

4

Appendix Figure 2 – Year 5 Reading: Culturally relevant

Appendix Table 1 – Non-compliers across Year, Subject & Cultural Background

	Year 6		Year 8
	Numeracy	Reading	Numeracy
Indigenous students	40	42	93
Non-Indigenous students	36	41	106

Note. Figures represent the number of students in each year, subject, and cultural background who did not attempt the tests on the day of examination.

Appendix Table 2 – Balancing Tests: Non-compliers - Year 6

	PANEL A - Numeracy Year 6						PANEL B - Reading Year 6					
	Treatment		Control		Difference		Treatment		Control		Difference	
Variable	Mean	SD	Mean	SD	C-T	SE	Mean	SD	Mean	SD	C-T	SE
All students												
Age	11.755	0.521	11.615	0.459	−0.140	0.113	11.723	0.516	11.609	0.435	−0.114	0.105
Male	0.611	0.494	0.575	0.501	−0.036	0.114	0.625	0.490	0.581	0.499	−0.044	0.109
Indigenous background	0.583	0.500	0.475	0.506	−0.108	0.116	0.500	0.506	0.512	0.506	0.012	0.111
English speaking	0.939	0.242	1.000	0.000	0.061	0.040	0.947	0.226	1.000	0.000	0.053	0.035
High parent education	0.056	0.232	0.050	0.221	−0.006	0.052	0.100	0.304	0.023	0.153	−0.077	0.052
Low parent education	0.500	0.507	0.425	0.501	−0.075	0.116	0.400	0.496	0.442	0.503	0.042	0.110
Student SES quartile	1.364	0.822	1.703	0.845	0.339	0.200*	1.553	0.978	1.585	0.741	0.033	0.194
Indigenous students, only												
Age	11.774	0.566	11.645	0.602	−0.129	0.185	11.738	0.581	11.663	0.556	−0.075	0.175
Male	0.619	0.498	0.526	0.513	−0.093	0.160	0.600	0.503	0.500	0.512	−0.100	0.157
English speaking	1.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
High parent education	0.000	0.000	0.053	0.229	0.053	0.050	0.050	0.224	0.046	0.213	−0.005	0.067
Low parent education	0.714	0.463	0.526	0.513	−0.188	0.154	0.700	0.470	0.546	0.510	−0.155	0.152
Student SES quartile	1.191	0.602	1.375	0.806	0.185	0.231	1.300	0.733	1.350	0.745	0.050	0.234
Non-Indigenous students, only												
Age	11.728	0.470	11.587	0.289	−0.141	0.127	11.708	0.458	11.552	0.256	−0.157	0.115
Male	0.600	0.507	0.619	0.498	0.019	0.170	0.650	0.489	0.667	0.483	0.017	0.152
English speaking	0.833	0.389	1.000	0.000	0.167	0.084*	0.889	0.323	1.000	0.000	0.111	0.070
High parent education	0.133	0.352	0.048	0.218	−0.086	0.095	0.150	0.366	0.000	0.000	−0.150	0.080*
Low parent education	0.200	0.414	0.333	0.483	0.133	0.154	0.100	0.308	0.333	0.483	0.233	0.127*
Student SES quartile	1.667	1.073	1.952	0.805	0.286	0.329	1.833	1.150	1.810	0.680	−0.024	0.298

Note. Balance tests conducted on the sample of students who did not attempt the tests on the day of examination.

Appendix Table 3 – Balancing Tests: Non-compliers - Year 8

	PANEL C - Numeracy Year 8					
	Treatment		Control		Difference	
Variable	Mean	SD	Mean	SD	C-T	SE
All students						
Age	13.636	0.345	13.662	0.398	0.026	0.053
Male	0.440	0.499	0.500	0.503	0.060	0.071
Indigenous background	0.422	0.496	0.522	0.502	0.100	0.071
English-speaking background	0.963	0.191	0.932	0.254	−0.031	0.032
High parent education	0.165	0.373	0.089	0.286	−0.076	0.048
Low parent education	0.257	0.439	0.433	0.498	0.177	0.067***
Student SES quartile	1.720	0.960	1.546	0.801	−0.174	0.128
Indigenous students, only						
Age	13.634	0.344	13.624	0.421	−0.010	0.080
Male	0.500	0.506	0.575	0.500	0.075	0.104
English-speaking background	0.978	0.147	0.957	0.206	−0.022	0.037
High parent education	0.109	0.315	0.043	0.204	−0.066	0.055
Low parent education	0.435	0.501	0.553	0.503	0.118	0.104
Student SES quartile	1.370	0.771	1.391	0.802	0.022	0.164
Non-Indigenous students, only						
Age	13.638	0.349	13.704	0.371	0.066	0.071
Male	0.397	0.493	0.419	0.499	0.022	0.098
English-speaking background	0.951	0.218	0.905	0.297	−0.046	0.051
High parent education	0.206	0.408	0.140	0.351	−0.067	0.076
Low parent education	0.127	0.336	0.302	0.465	0.175	0.078**
Student SES quartile	1.984	1.008	1.714	0.774	−0.269	0.185

Note. Balance tests conducted on the sample of students who did not attempt the tests on the day of examination.