MINISTRY OF EDUCATION
TE TĀHUHU O TE MĀTAURANGA

# Briefing Note: Further advice on measures of system progress

| | | | |
|---|---|---|---|
| **To:** | Hon. Chris Hipkins, Minister of Education | | |
| **Copy:** | Hon. Jan Tinetti, Associate Minister of Education<br>Hon. Kelvin Davis, Associate Minister of Education<br>Hon. Aupito William Sio, Associate Minister of Education | | |
| **Date:** | 28 September 2021 | **Priority:** | Medium |
| **Security Level:** | In Confidence | **METIS No:** | 1272093 |
| **Drafter:** | Andrew Webber | **DDI:** | 9(2)(a) |
| **Key Contact:** | Alexander Brunt | **DDI:** | 9(2)(a) |
| **Messaging seen by Communications team:** | No | **Round Robin:** | No |

## Purpose of Report

Following the Education Work Programme strategy session on 15 September, this briefing note provides some contextual information on our current capacity to monitor and analyse outcomes relating to literacy, numeracy, attendance and wellbeing. It also summarises the development work underway on measuring these outcomes, and opportunities to further improve our ability to identify system shifts in the future.

The purpose of this paper is for you to:

a. **Note** that the Ministry has a work programme underway focusing on the improved collection of attendance, wellbeing and engagement data, in order to support and underpin the operational and policy work in the Education Work Programme.

b. **Note** that using new datasets for system monitoring purposes requires involvement with and clear communication to schools and kura. If the purpose is perceived to be related to accountability of providers or teachers (rather than about measuring improvement to the system), there is a danger that this will damage the integrity of the underlying tools.

## Proactive Release

c. **agree** that the Ministry of Education release this briefing in full once it has been considered by you.

[✓ **Agree**] **Disagree.**

Alexander Brunt
**Deputy Secretary**
**Evidence, Data and Knowledge**

28/09/2021

Hon. Chris Hipkins,
**Minister of Education**

__/__/2021

## Background

1. This paper relates to discussion at the recent strategy session focused on the Education Work Programme. In that conversation, you asked questions about how and when we will be able to measure the impact the Education Work Programme is making on learner outcomes across the system, in particular in relation to progress in literacy and numeracy, and attendance and wellbeing. A summary of the activity discussed in this briefing is attached (Annex 1).

2. An Education Report on an Education System Monitoring Framework for the NELP and TES is being provided to your office alongside this paper (METIS 1271484 refers). The framework identifies headline indicators that can show progress against the NELP and TES, as well as components of Ka Hikitia, the Action Plan for Pacific Education and other strategies. We have aimed for a small set of headline indicators. These have been selected based on having high quality and robust data, for which we have an ongoing time-series. The report identifies where new work is currently being done to develop indicators (for example by ERO's Education Now surveys and the Ministry's work on Student Wellbeing measures) and identifies the gaps in measures which will require consideration of options for development.

3. 9(2)(f)(iv) [redacted]

## Progress and achievement in literacy and numeracy

**Medium- and long-term measures**

4. Our system monitoring of literacy and numeracy learning outcomes is underpinned by a set of studies undertaken on a representative sample of school students at particular intervals (see below table). These each provide snapshots of student achievement at key points of schooling. Each collection also includes student surveys that gather information about teaching practices, classroom climate, engagement and wellbeing.

| Study | Year level | Learning areas | Māori medium? | Interval | Next reporting |
|---|---|---|---|---|---|
| National Monitoring Study of Student Achievement (NMSSA) | Year 4/8 | All learning areas in the NZC (over a 5-year cycle) | No | Annual | Maths will be collected in 2022 and reported in 2023 |
| Progress in International Reading Literacy Study (PIRLS) | Year 5 | Reading | Yes | Every 5 years | Dec 2022 (Collected T4 2020) |
| Trends in International Mathematics and Science Study (TIMSS) | Year 5/9 | Mathematics, science | No | Every 4 years | Dec 2024 (collected 2022 for Y5; 2023 for Y9) |

| | | | | | |
|---|---|---|---|---|---|
| Programme of International Student Assessment (PISA) | Age 15 | Reading, mathematics, science | No | Every 3 years | Dec 2023 (Collected 2022) |

5. These studies all incorporate a large amount of assessment development, validation, and a careful selection of sample to ensure the results are nationally representative and internationally comparable. The data resulting from these studies is highly reliable and include a very long time series, allowing us to articulate long-term shifts in system outcomes. These studies also include the collection of additional information on the drivers of student achievement, allowing us to better understand the context of literacy and numeracy learning over time.

6. However, the consequence of this depth and reliability is that these studies do not provide us with very responsive measures. Each study measures literacy and numeracy outcomes only every three to five years. The studies also measure achievement at a specified point (for example, age 15), which is the product of all years of learning up until this point. This can mean that even substantial system shifts can take some time to translate into meaningful changes in these results.

**Shorter-term measures**

7. An alternative data source that might provide an indication of more short-term changes is the e-asTTle student assessment tool that is administered by the Ministry. e-asTTle assesses reading, writing and mathematics over Years 4 to 10, and is freely available for teachers' optional use, to inform approaches to teaching, communication to whānau, and school planning. Data from this tool is available for Ministry statistical and research purposes that do not identify individual students, teachers, or schools. Many teachers also use Progressive Achievement Tests (PAT), administered by NZCER (for which the Ministry does not receive data), and usage of the Ministry's Progress and Consistency Tool (PaCT) is also beginning to increase.

8. Historically, this information has not been extensively relied upon for system-level uses, because the tools are optional, and teachers who do participate may not be representative of the system as a whole. However, e-asTTle has very high coverage, with 59% of schools and 40% of students represented in the 2019 data (these proportions are greater when focused on students in Years 4-10). Recent Ministry analysis of e-asTTle data (linked to PISA) data found that not only do e-asTTle assessments of reading and mathematics display high degrees of validity and reliability, but the students assessed in this data also appear to be representative of the overall English-medium schooling population, at least when examining progress (see Annex 2).

9. As a consequence of these findings, we are beginning to use e-asTTle more for system-level research and evaluation purposes. For example, e-asTTle reading data was used in the Ministry's evaluation of Reading Recovery, by looking at the impact of a student having Reading Recovery available to them when they were in Year 2 on later reading outcomes over Years 4 to 10. We are currently using e-asTTle data as part of our evaluation of the impact on learners of provision of connectivity and devices to households in response to COVID-19 during 2020. We also recently published a research paper using e-asTTle as a responsive monitoring tool, estimating the impacts of COVID-19 on literacy and numeracy scores (Webber, 2021). That study found little evidence of substantial drops in learning progress in 2020, except potentially in writing.

10. The bulk of e-asTTle assessments are undertaken in Term 4, meaning the most robust measure of progress is available annually. However, large numbers of assessments are also undertaken in Terms 1 to 3. We have experimented with using this data for more responsive monitoring, on a termly basis. Our experience from 2020 was that the results in the end-of-

year comparisons of progress (little evidence of negative impacts on learning) were consistent with data that was available at the end of Term 2, seven weeks after the end of the national lockdown. This might imply e-asTTle data holds great promise for monitoring even very short-term changes in learning progress.

11. However, it would be important to proceed cautiously, and involve sector peak bodies, if we were to explore using data from these assessment tools for regular reporting purposes (as opposed to solely research and evaluation uses, as is our current practice). Failing to adequately involve those from the sector in any discussions about monitoring may lead to destruction of the integrity of these critical tools for learning.

12. While teachers and school leaders might be amenable to measures that were aligned to what they already use in regular teaching practice and impose no additional burden, it is important to be clear that the purpose of any such reporting would be in articulating progress for the system, rather than accountability for providers, and it will continue to be the case that no student, teacher or school will be identified in any data. Our previous experience with assessment tools such as PaCT is that perceived connection to previous accountability requirements led many teachers and school leaders to avoid using them, despite their utility for teaching practice. In the case of PaCT, use of the tool is only now beginning to grow from very low levels, despite being available since 2016.

**Future opportunities**

13. As part of the changes to the NCEA qualification, the Ministry and NZQA are currently piloting a new set of standards that represent literacy and numeracy corequisites. As part of the evaluation of this pilot, the assessment of these standards are being assessed for accuracy and reliability (including a comparison to e-asTTle literacy and numeracy data). Once these standards are implemented across the system, NCEA data will provide us a comprehensive view of literacy and numeracy skills by the time students arrive at senior secondary year levels. Depending on how these standards are assessed (for example, through digital assessment methods), these standards may provide more detailed information on aspects of literacy and numeracy than many other NCEA standards.

14. The changes to the national curricula and the development of a record of learning (to be securely managed and shared – where appropriate – by the Te Rito platform) also present powerful future opportunities to better articulate improved learning across the system, including (but not limited to) in literacy and numeracy.

15. One gap of current assessment data is examining learning outcomes in early primary years, where most of our literacy and numeracy data relates to Years 4 to 10. The Ministry is currently implementing a new early literacy approach, which is aimed at encouraging a more consistent evidence-based approach to all aspects of teaching literacy over Years 1 to 3, including assessment. This early literacy approach has an ongoing evaluation over a three-year span. The insights from this evaluation (both from assessment data and from student and teacher voice) will allow us to articulate system shifts as they are happening.

16. While we have a range of data sources relating to literacy and numeracy outcomes in English-medium schooling, our current ability to articulate how the Education Work Programme is improving learning outcomes for ākonga in Māori medium (outside of NCEA attainment) is more limited. The Te Rito platform provides an infrastructure to consistently collect and securely share (where authorised) aromatawai and learning data across the system, but there is an opportunity to further support kura with tools and support to make aromatawai easier. This is also related to a broader conversation about Māori data governance, and how to best ensure that Māori determine the narrative of their own success.

### Medium- and long-term measures

17. The four comparative studies (NMSSA, PIRLS, TIMSS and PISA) discussed above also each contain surveys of students, asking them about aspects of their lives that are related to engagement, attendance and wellbeing. Many of these questions have been asked of students over a long period of time, and are asked internationally, allowing us to put any future changes in context. The questions often relate to contemporaneous events (for example, bullying over the previous 12 months) as opposed to the achievement measures in these studies, which measure cumulative learning over all education to this point. This might mean they are more able to detect shorter-term changes in system conditions.

### Shorter-term measures

18. The most responsive measure we have on this topic is the rate of school attendance. Detailed data is recorded by schools every day (using one of 26 different codes relating to different types of attendance or absence, for each student in each period of the day), and across most schools, collected and reported by the Ministry every week. This collection now captures data across most schools each week, with more than 95% of schools reporting data at least on a termly basis. This attendance data allows us to examine shifts in attendance (whether on-site or off-site) on a very granular basis. A similar granularity, coverage and timeliness is available for participation at early learning services (although this data does not currently include nga kōhanga reo).

19. This influx of comprehensive and timely attendance data is relatively new, and we are continuing to make improvements to how we report the most important trend information at a system level, as well as how regional Ministry offices can be provided the data they need to examine local conditions and support schools to respond to emerging issues.

20. In terms of responsive measures of other aspects of student wellbeing and/or engagement, the largest current system-level dataset is the Wellbeing@School student survey. This is a free psychometrically-validated survey toolkit available for schools to use at any point during the year and includes student, teacher and school systems surveys. The survey tools are administered online by NZCER on behalf of the Ministry, and support schools to undertake an in-depth self-review of the aspects of school life that contribute to creating a safe and caring school climate that deters bullying. While the Ministry does not have access to the data resulting from the survey, NZCER are able to provide analysis of aggregate data trends that could be used for regular monitoring.

21. There are three key limitations of using Wellbeing@School for system monitoring. Firstly, a minority of schools opt to regularly use it, and we do not know if these schools are representative of the system as a whole. Secondly, while reporting on key findings to the Ministry can be completed on request, there may not be a large enough take-up of the survey to report more frequently than on an annual basis. Thirdly, this survey is primarily intended to support self-review processes to help schools create climates that promote student wellbeing, and using this data for system monitoring purposes carries similar risks as using assessment tools: that some schools might opt out of the tool, due to concerns about system-level uses of the data.

### Future opportunities

22. There are several collections that are likely to provide us with system-level data relating to learner wellbeing at regular snapshots into the future. The first is the What About Me student survey commissioned by the Ministry of Social Development. This is a wellbeing survey of a nationally representative sample of secondary school-aged young people. The first collection of this survey was earlier in 2021 (prior to the national lockdown) with reporting due in early 2022. Follow-ups in future years are likely, although the exact regularity has not been confirmed.

23. The second set of emerging data is from Education Review Office’s new work in schools which includes schools participating in leader, teacher, student and board surveys, as well as a survey of a sample of parents and whānau. This work is underway but has been impacted by the current heightened alert levels so results are not due till next year. These are likely to provide useful short-term data on how the system is promoting wellbeing for learners.

24. The third (and more long-term) opportunity is the set of measures of student wellbeing currently being co-designed between the Ministry and students (METIS 1252367 refers). This project aims to result in a common understanding of the aspects of student wellbeing that students most strongly identify with and deem most important to measure, and the most practical means of measuring these aspects. The primary objective of these measures is to inform and empower schools, kura, providers and communities to respond to and improve learner and ākonga wellbeing, with potential applications for system-level uses a secondary consideration 9(2)(f)(iv) [redacted]

## Annexes

Annex 1: How our measurement is evolving

Annex 2: e-asTTle: Validity, reliability, and representativeness

## Annex 1 How our measurement is evolving

MINISTRY OF EDUCATION
TE TĀHUHU O TE MĀTAURANGA

# HOW OUR MEASUREMENT IS EVOLVING 2021

Over the last several years, we have hugely expanded our ability to understand and monitor the student outcomes that matter the most, and pinpoint how our actions are affecting these outcomes.

| TOPIC | PAST STATE (1-2 YEARS AGO) | RECENT AND CURRENT ACTIVITY | | FUTURE ACTIVITY |
|---|---|---|---|---|
| ATTENDANCE: DATA COLLECTION | COLLECT TERM 2 ATTENDANCE ONLY | COLLECT DATA FOR ALL TERMS [DONE] | AUTOMATICALLY COLLECT EVERY WEEK; NO BURDEN ON SCHOOLS [DONE] | SIMPLIFY CODES TO MAKE IT EASIER TO RECORD ATTENDANCE DATA |
| ATTENDANCE: SYSTEM REPORTING | REPORT ONCE ANNUALLY TERM 2 ONLY | REPORT DATA FOR ALL TERMS [DONE] | REPORT DAILY HEADCOUNTS EACH WEEK [DONE] | REDUCE WEEKLY TURN-AROUND TIME TO THREE DAYS AND ANALYSE ABSENCES [THIS WEEK] |
| ATTENDANCE: DATA TO SUPPORT SCHOOLS | REGIONAL ANALYSTS GET SOME DATA BY SCHOOL; SCHOOLS GET ONE REPORT A YEAR | SCHOOLS/ REGIONAL ANALYSTS GET DATA AT THE END OF EVERY TERM [DONE] | REGIONAL ANALYST GET DETAILED DATA EACH WEEK [DONE] | REGIONAL ANALYSTS GET DYNAMIC DASHBOARDS THAT LOOK AT LOOK AT STUDENT TRAJECTORIES |
| STUDENT WELLBEING | SAMPLE-BASED STUDIES LOOKING AT SOME ASPECTS OF WELLBEING EVERY 3-5 YEARS | A NUMBER OF DIFFERENT NEW SYSTEM-LEVEL SURVEYS, OVER PRIMARY AND SECONDARY, ALLOWING US TO TRIANGULATE AND EXAMINE DIFFERENT ASPECTS OF WELLBEING [DATA COLLECTION IN THE FIELD] | | A MEANS OF CONSISTENT, EASY TO USE MEASUREMENT AVAILABLE TO ALL STUDENTS |
| LITERACY AND NUMERACY | SAMPLE-BASED STUDIES EVERY 3-5 YEARS | INCREASED USE OF E-ASTTLE PROGRESS FOR RESEARCH/ EVALUATION [DONE] | NCEA ASSESSMENTS OF LITERACY AND NUMERACY [CURRENTLY PILOTING] | MEASURES ARE INCLUSIVE OF TE REO MATATINI ME TE PĀNGARAU |
| USE OF IDI | DESCRIBE THE RELATIONSHIP BETWEEN SOME CHARACTERISTICS OF STUDENT BACKGROUND AND EDUCATIONAL SUCCESS | DESCRIBE THE RELATIONSHIP BETWEEN ALL CHARACTERISTICS (FOR WHICH DATA IS AVAILABLE) AND EDUCATIONAL SUCCESS [DONE] | USE THESE RELATIONSHIPS TO TARGET SCHOOLING RESOURCES WHERE THERE IS THE LARGEST NEED (EQUITY INDEX) [CURRENTLY IMPLEMENTING] | USE THE EQUITY INDEX TO SUPPORT EVALUATION BY COMPARING OUTCOMES WITH STUDENTS FROM SIMILAR SOCIO-ECONOMIC BACKGROUNDS |
| SYSTEM MONITORING AND EVALUATION | NO OUTCOME STATEMENT | KŌRERO MĀTAURANGA AND THE CREATION OF KEY EDUCATION STRATEGIES. [DONE] | MONITORING AGAINST NELP \| TES; KA HIKITIA; PACIFIC EDUCATION AND EARLY LEARNING ACTION PLANS [IN PROCESS OF REPORTING TO YOU] | SYSTEMATIC EVALUATION OF ACTIVITIES ACROSS THE EDUCATION WORK PROGRAMME |

New Zealand Government

MINISTRY OF EDUCATION
TE TĀHUHU O TE MĀTAURANGA

# e-asTTle: Validity, reliability, and representativeness

## Preliminary results, June 2020

Andrew Webber

# For us to put a lot of stock in data, it must be valid, reliable, and representative

| | |
|---|---|
| Face validity | Whether a test appears to be a good measure of reading |
| Content validity | Whether a test covers the aspects of reading we care about |
| Convergent validity | Whether a test agrees with other tests of reading |
| Discriminant validity | Whether a test gives different results than a test of, say, maths |
| Reliability | Whether a test produces the same results for the same students |
| Predictive validity | Whether results of the test have relationships with later outcomes we think are important |
| Represent-ativeness | Whether the students who take the test are representative of the population as a whole |

# For some aspects of validity, you don’t need data to assess

| | |
|---|---|
| Face validity | Whether a test appears to be a good measure of reading |
| Content validity | Whether a test covers the aspects of reading we care about |

This is a subjective measure anyone can make, including (especially) the people actually taking the test.

For this, you need a good understanding of pedagogy and the curriculum.

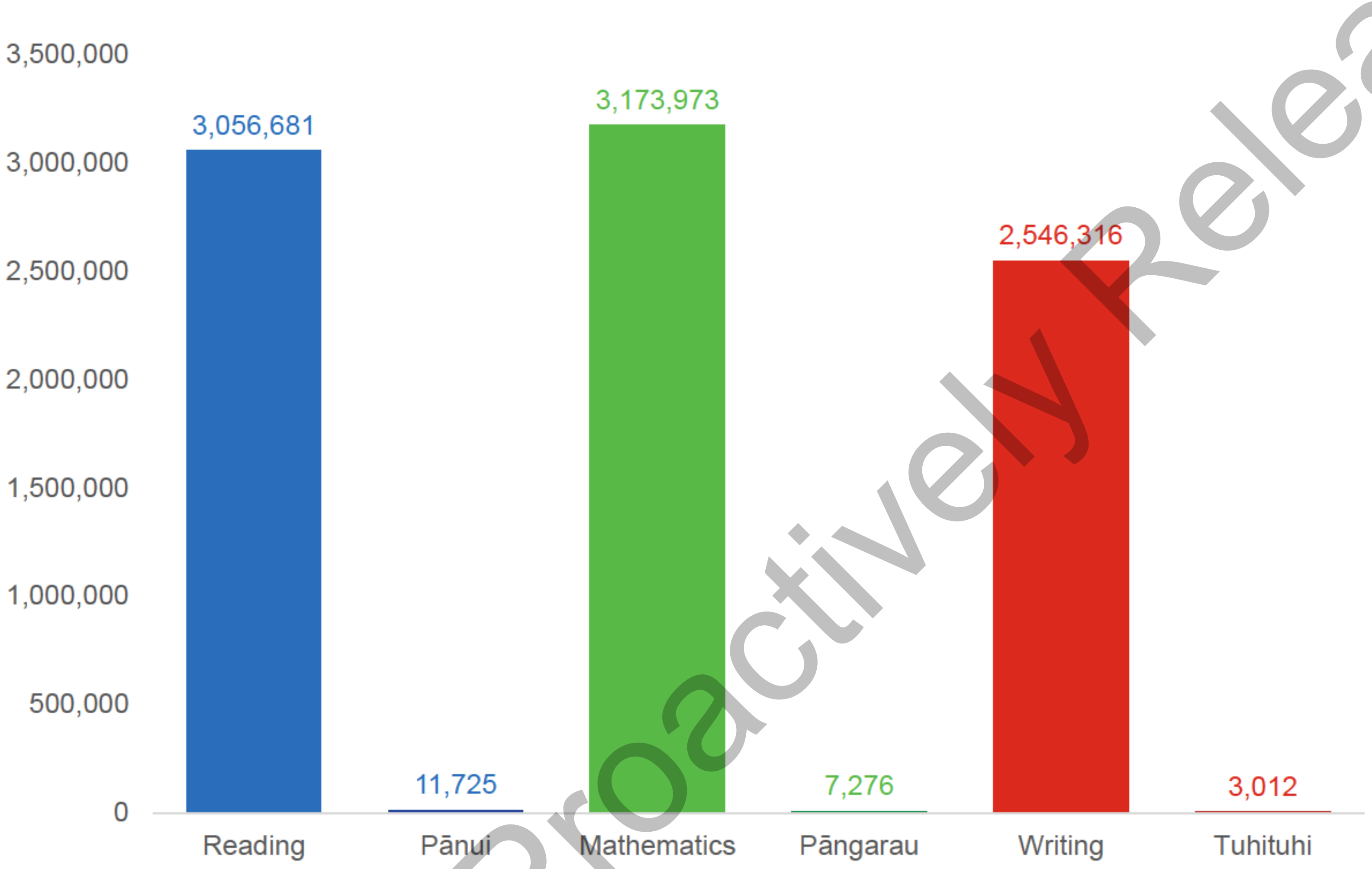
My research does not assess these aspects of validity (or other aspects of validity not mentioned here).

# How can we use data to assess these other aspects of validity?

I compare e-asTTle scores for the same students at different points during the year to get at this

| | |
|---|---|
| Reliability | Whether a test produces the same results for the same students |

# How do we assess reliability?



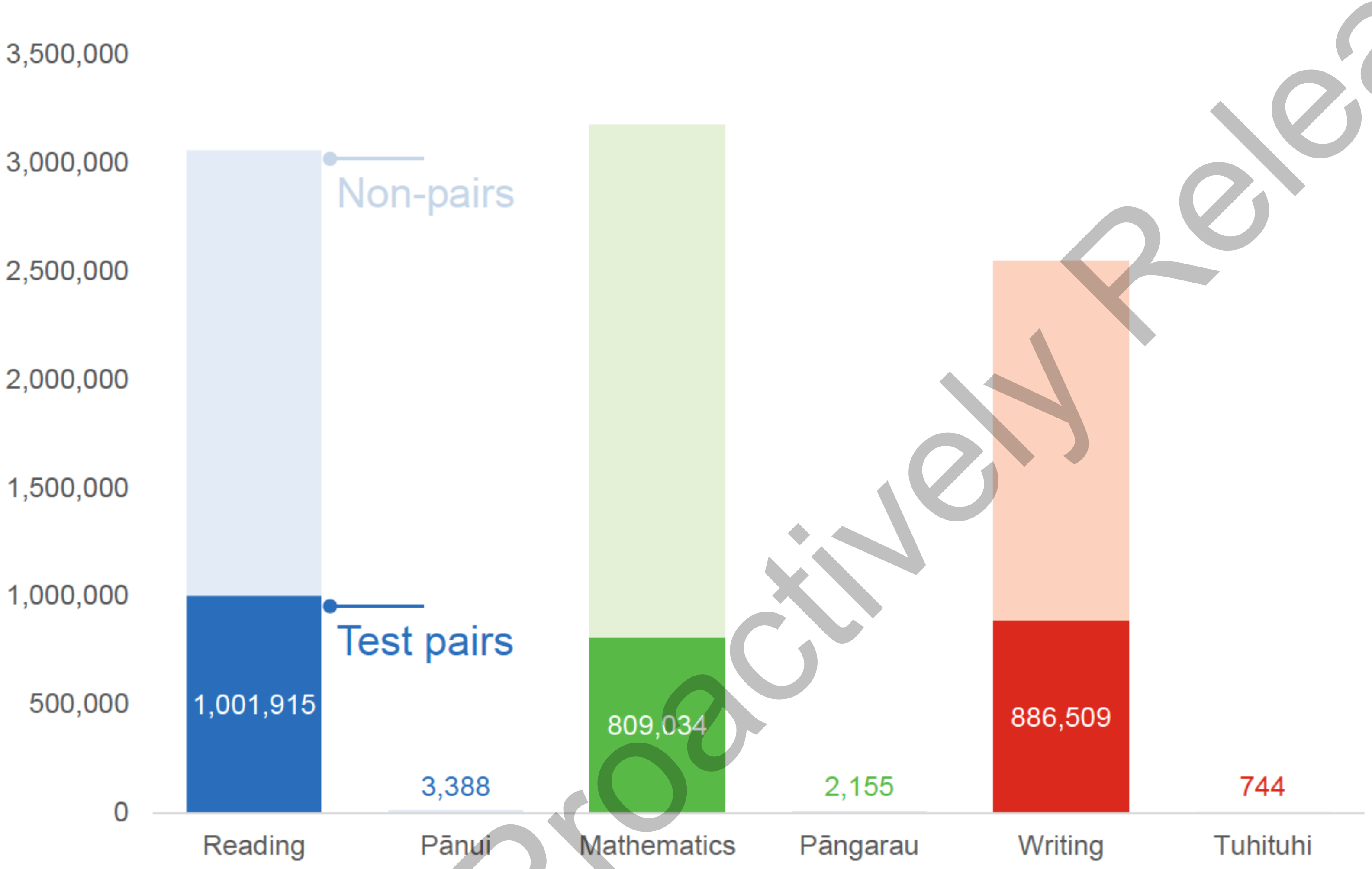
In total, there are about 8.8m individual assessments in the e-asTTle database (from 2011-2018).

Note that e-asTTle has assessments in English and te reo Māori, but these latter assessments aren't commonly used:

- Pānui represent 0.4% of reading tests
- Pāngarau represent 0.2% of maths tests
- Tuhituhi represent 0.1% of writing tests

(2.6% of students are in Māori medium education.)

# How do we assess reliability?



I went through all assessments and identified pairs where:

a. The same **student**

b. Was assessed in the same **subject**

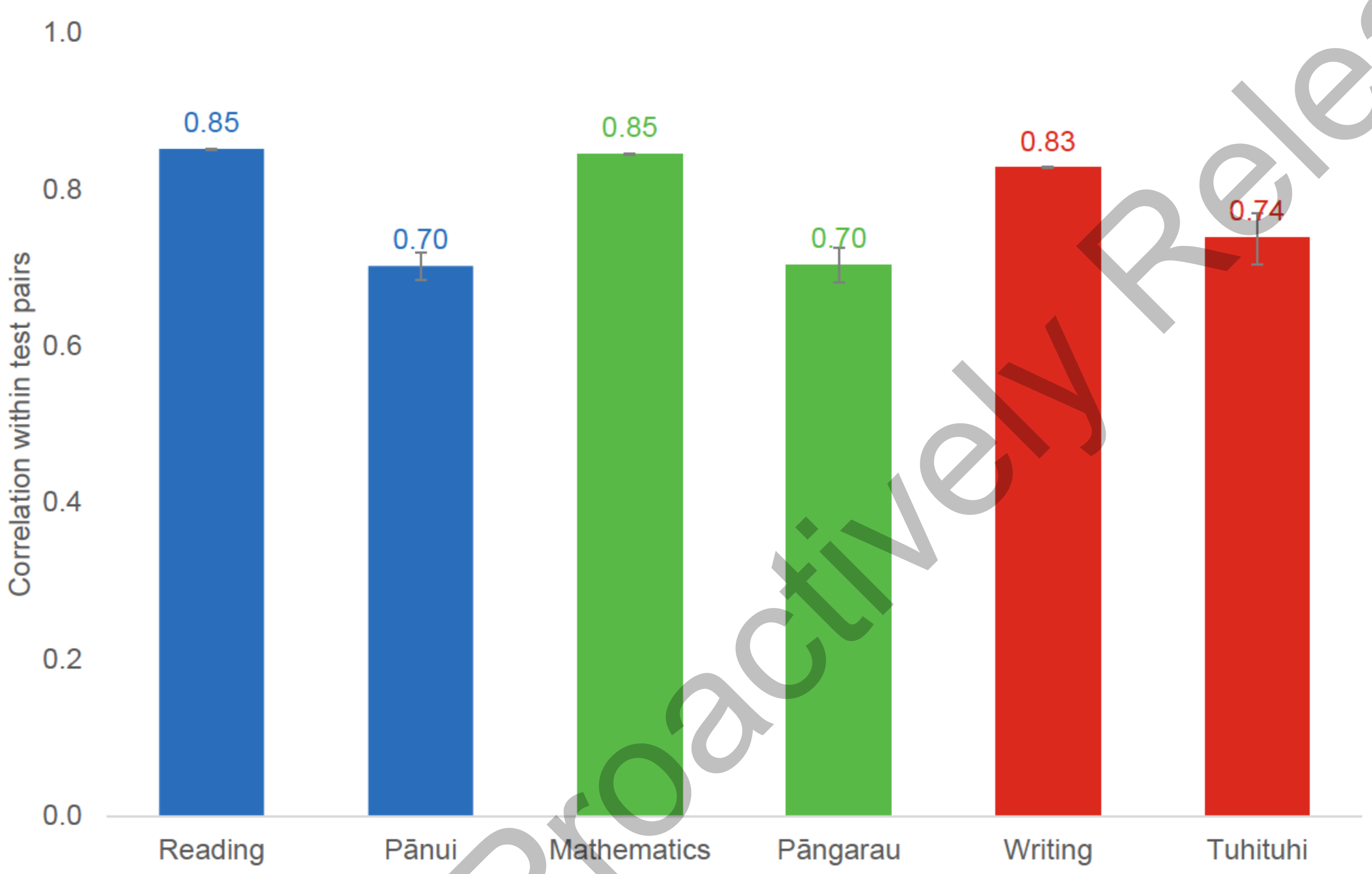
c. More than once in the same **year**.

I then looked at the correlation between the scores in each pair.

The theory behind this is if e-asTTle is reliable, taking the same test multiple times should produce roughly the same result.

(In reality, this is complicated by the fact that students are learning in between each test!)

# Results indicate that all domains are highly reliable



All correlations are pretty close to 1 (which indicates where one score perfectly predicts the next score).

Differences from 1 can be quite easily explained by:

- Learning taking place throughout the year
- Different aspects being tested (eg trig in Term 1; algebra in term 2)

Assessments in te reo Māori are less reliable according to this measure.

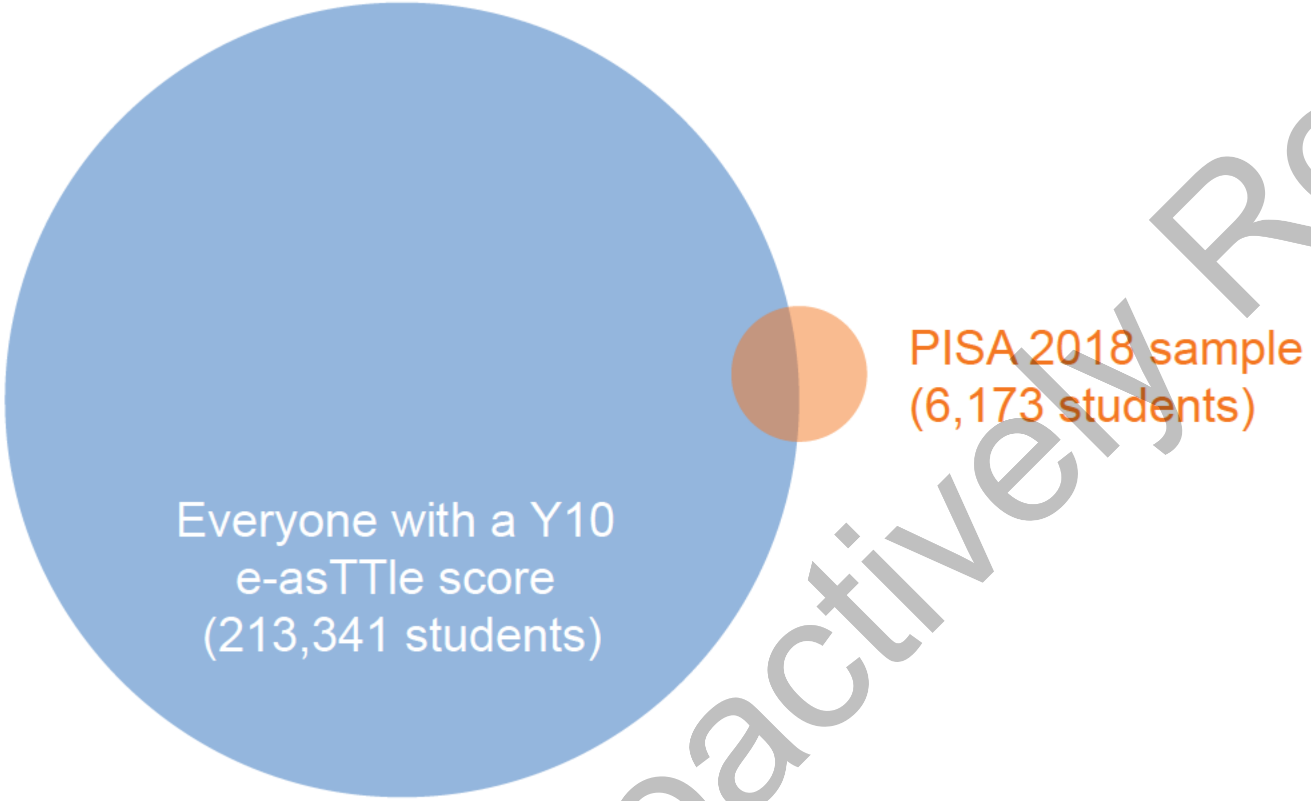
Note: Reliability is a necessary but not sufficient condition for validity. (A scale that is 5kg off is reliable but not valid.)

# How can we use data to assess these other aspects of validity?

| | |
|---|---|
| Convergent validity | Whether a test agrees with other tests of reading |
| Discriminant validity | Whether a test gives different results than a test of, say, maths |

To get at these aspects, I join e-asTTle data to PISA data (2018 wave)

# We can use PISA as a gold standard to calibrate e-asTTle



PISA is an OECD-led assessment of reading, maths and science of 15 year old students in more than 70 countries. A huge amount of work goes into designing PISA to be valid and reliable – it is the closest we have to a gold standard assessment of reading and maths.

About 2,000 students participated in PISA in 2018 **and** had a Y10 e-asTTle test. This Y10 test should have happened within 3-9 months of PISA.

If students who do well in e-asTTle also tended to do well in PISA, then this tells us something about e-asTTle validity.

Note: PISA does not assess writing or Māori medium.

# Psychometricians use a framework called the Multitrait-Multimethod Matrix (MTMM)

| | | e-asTTle | | PISA | |
|---|---|---|---|---|---|
| | | Reading | Maths | Reading | Maths |
| e-asTTle | Reading | 0.851 | | | |
| | Maths | 0.660 | 0.845 | | |
| PISA | Reading | 0.719 | 0.638 | 0.85 | |
| | Maths | 0.631 | 0.679 | 0.794 | 0.86 |

This matrix shows the relationships between different tools that assess the same subjects (e-asTTle and PISA), and different subjects (reading and maths). All of the figures in the table are correlations.

The diagonals in blue are the reliability estimates shown earlier for e-asTTle, and are taken from the PISA technical report in 2015. All values are extremely similar, and all very high.

# e-asTTle has high convergent validity…

| | | e-asTTle | | PISA | |
|---|---|---|---|---|---|
| | | Reading | Maths | Reading | Maths |
| e-asTTle | Reading | 0.851 | | | |
| | Maths | 0.660 | 0.845 | | |
| PISA | Reading | 0.719 | 0.638 | 0.85 | |
| | Maths | 0.631 | 0.679 | 0.794 | 0.86 |

These shaded cells in blue show us the similarity between e-asTTle's estimate of reading ability and PISA's estimate of the same thing.

We want these numbers to be high. This is called **convergent validity**.

These figures are relatively high, with reading being higher than maths.

# …and also high discriminant validity

| | | e-asTTle | | PISA | |
|---|---|---|---|---|---|
| | | Reading | Maths | Reading | Maths |
| e-asTTle | Reading | 0.851 | | | |
| | Maths | 0.660 | 0.845 | | |
| PISA | Reading | 0.719 | 0.638 | 0.85 | |
| | Maths | 0.631 | 0.679 | 0.794 | 0.86 |

The orange shaded cells show us the relationships between e-asTTle scores in one subject (eg reading) and PISA scores in the other subject.

If e-asTTle is valid, reading e-asTTle scores should be more similar to PISA scores in reading than they are to PISA scores in maths. (Blue cells should be greater than orange cells.) This appears to be the case.

This is called **discriminant validity**.

# E-asTTle does not appear to have a strong ‘methods factor’ (which is good)

| | | e-asTTle | | PISA | |
|---|---|---|---|---|---|
| | | Reading | Maths | Reading | Maths |
| e-asTTle | Reading | 0.851 | | | |
| | Maths | 0.660 | 0.845 | | |
| PISA | Reading | 0.719 | 0.638 | 0.85 | |
| | Maths | 0.631 | 0.679 | 0.794 | 0.86 |

The green shaded cells show us the relationships between scores in a different subject but using the same method (e-asTTle or PISA).

We ideally want the green cells to be lower than the relationships between two different methods of measuring the same subject (the blue shaded cells).

One of the green cells is higher. However, it is the within-PISA relationship, indicating that PISA (not e-asTTle) has a strong ‘methods factor’.

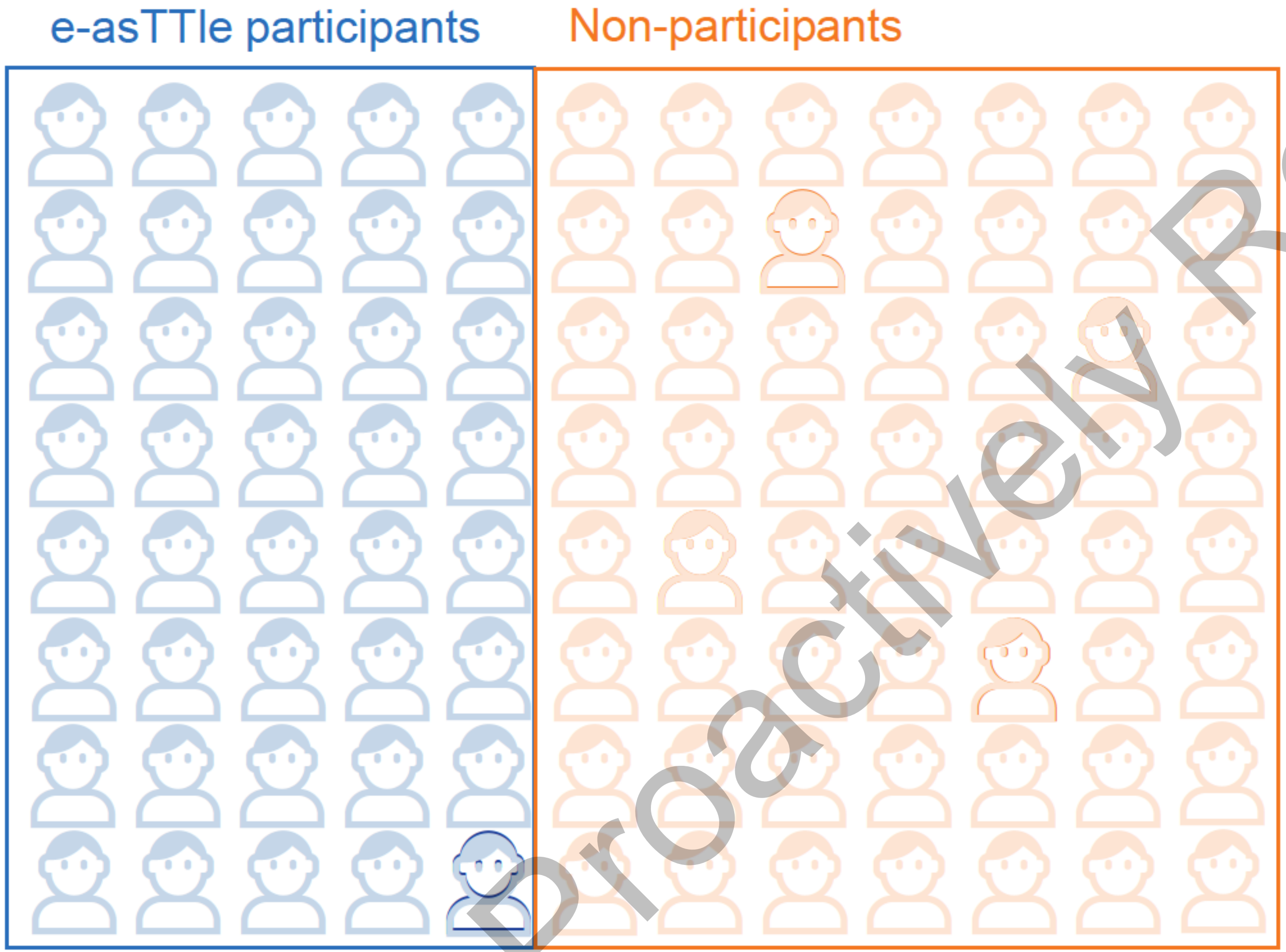
Taken together, this matrix is strong evidence for the validity of e-asTTle.

# For us to put a lot of stock in data, it must be valid, reliable, and representative

**Representativeness** — Whether the students who take the test are representative of the population as a whole

To assess this, I also use data that has been linked with PISA, but in a slightly different way.

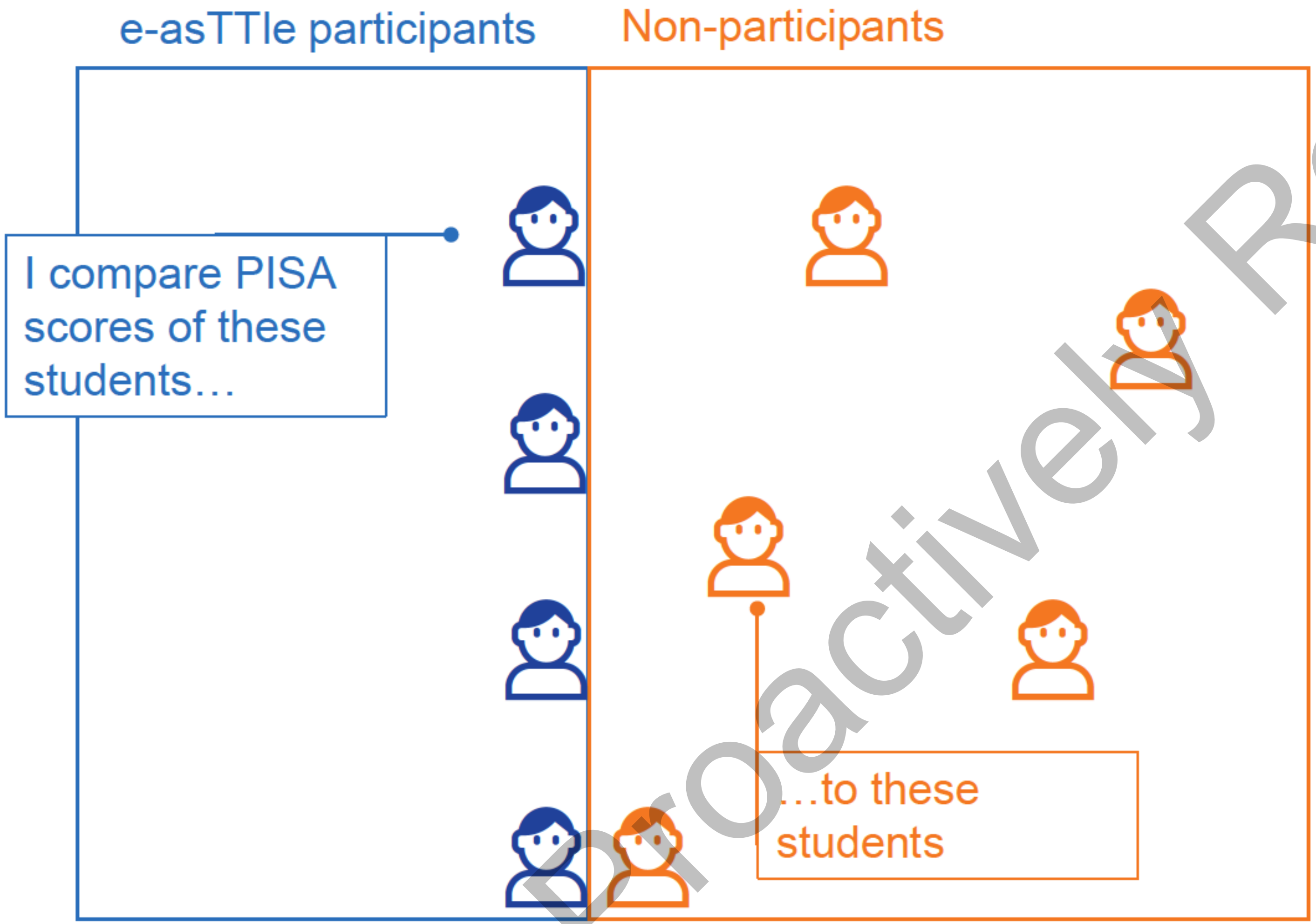
# We can use PISA to compare the ability of e-asTTle participants with non-participants



The problem with e-asTTle is participation isn't determined randomly. Because we don't get scores for those who don't participate, we don't know to what extent participants are similar to non-participants.

PISA is a representative slice of the 15 year old English medium population. Slightly more than 10% of the relevant cohort ends up in the PISA sample.

# We can use PISA to compare the ability of e-asTTle participants with non-participants



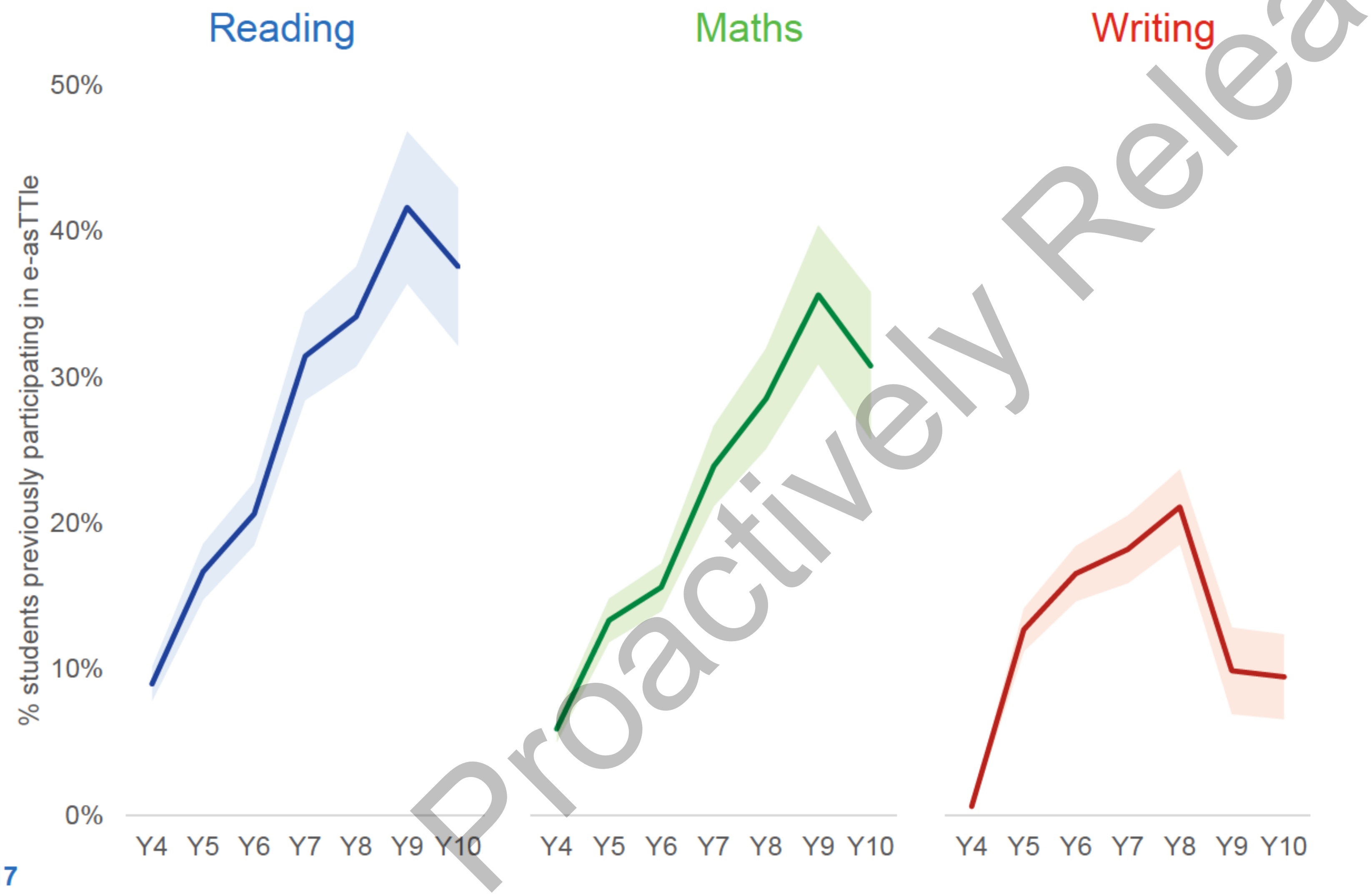
I compare PISA scores of these students…

…to these students

We want to know, is the ability of e-asTTle participants (as measured by PISA scores) significantly different to non-participants? If it is, this is evidence that the e-asTTle dataset is not representative.

(Note that this requires the assumption that participating in e-asTTle doesn't *cause* a student later go on to get better/worse PISA results.)

# PISA 2018 students were most likely to have participated in e-asTTle in Years 9 and 10

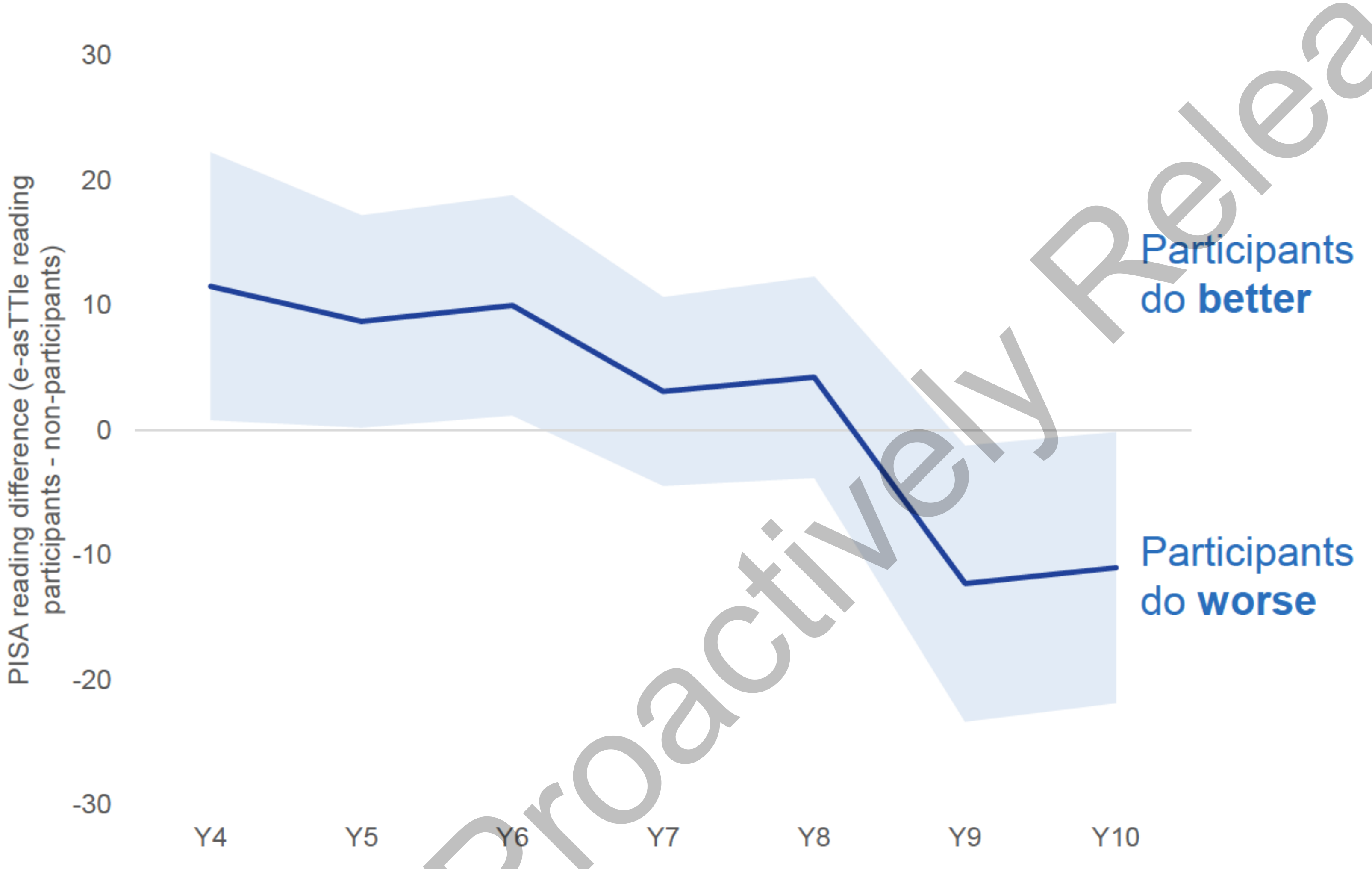


Intermediate and secondary schools are more likely to use e-asTTle than primary schools. This means that usage jumps up as students move into these types of schools.

Writing is used far less than other subjects, especially in secondary school.

Because PISA is a representative survey we can use it to make inferences about the population. However, this is measured with error, as shown by confidence intervals (shaded areas).

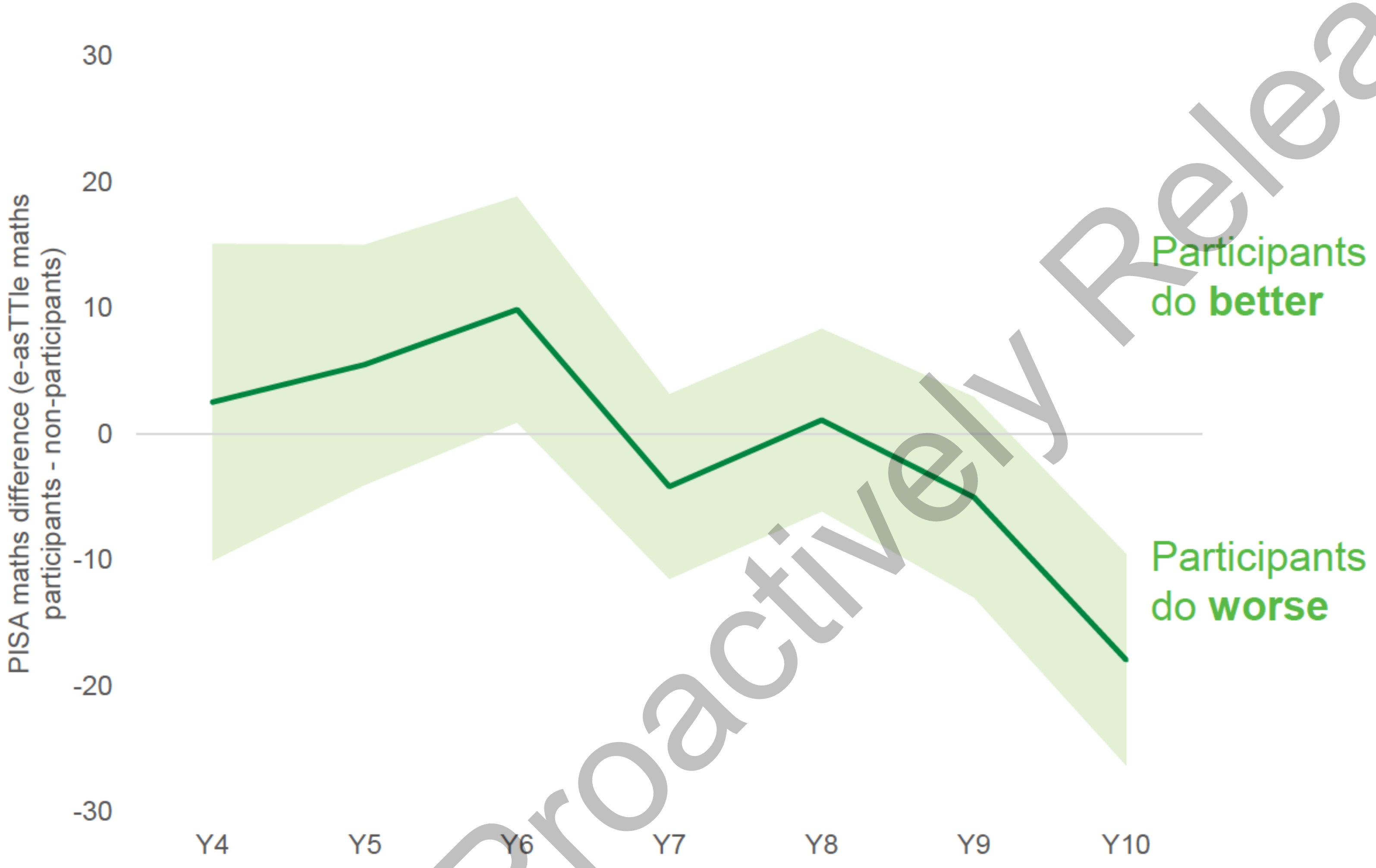
# e-asTTle students are not representative of the population in Years 4, 5, 9 or 10



This graph shows the difference in PISA reading scores between people who participated in a e-asTTle reading test in a particular year and those who did not participate.

This indicates the e-asTTle sample in primary school years is disproportionately higher ability students, and in secondary school years is disproportionately lower ability students.

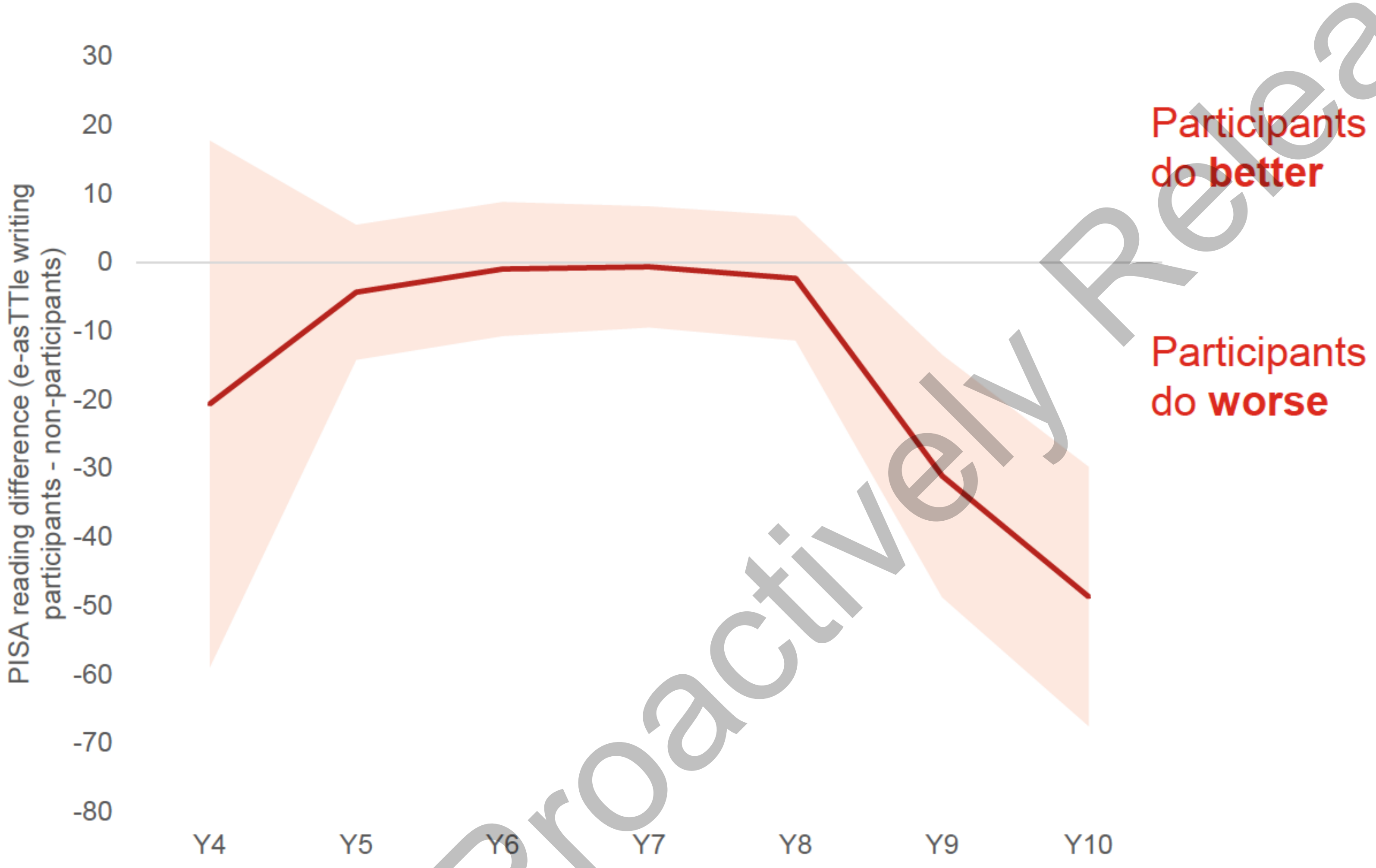
# There is a similar pattern for maths



This is the same graph but looking at differences in PISA maths scores for participants and non-participants of e-asTTle maths tests.

There is less of selection effect in primary school here, but e-asTTle maths participants in secondary school are still significantly lower performers. (A difference in PISA score of 18 points is quite substantial.)

# Huge selection effects in e-asTTle writing

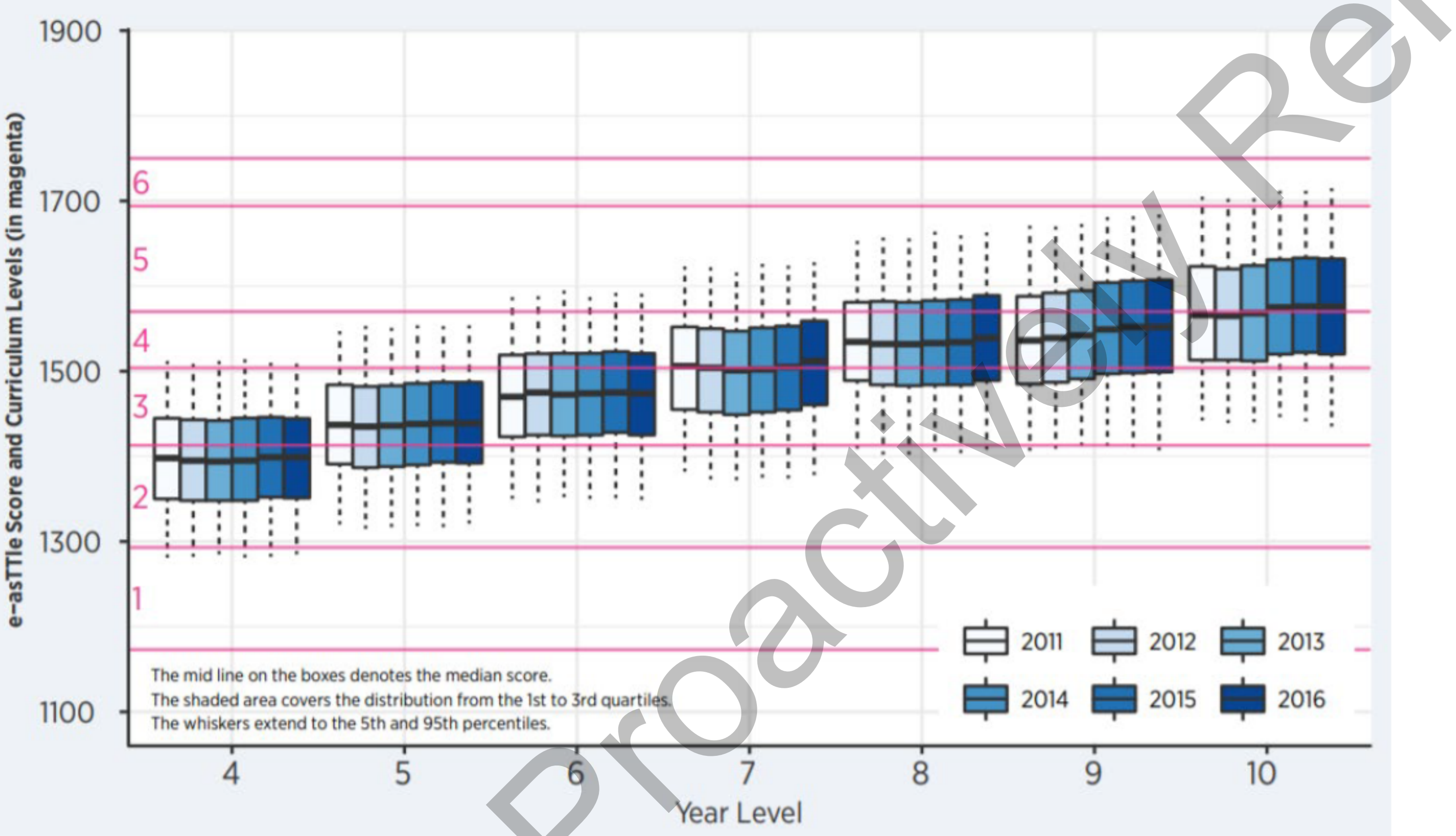


This looks at reading ability differences of participants vs non-participants in e-asTTle writing. (PISA doesn't have a writing assessment.)

Confidence intervals are much wider here because e-asTTle writing is less commonly used, but it indicates huge selection effects in secondary school (about half a standard deviation), and potentially also in primary school.

# These selection effects mean comparing averages between years is misleading

**Figure 3.** Distribution of mathematics achievement by year level mapped to curriculum level (2011 to 2016)



The mid line on the boxes denotes the median score.
The shaded area covers the distribution from the 1st to 3rd quartiles.
The whiskers extend to the 5th and 95th percentiles.

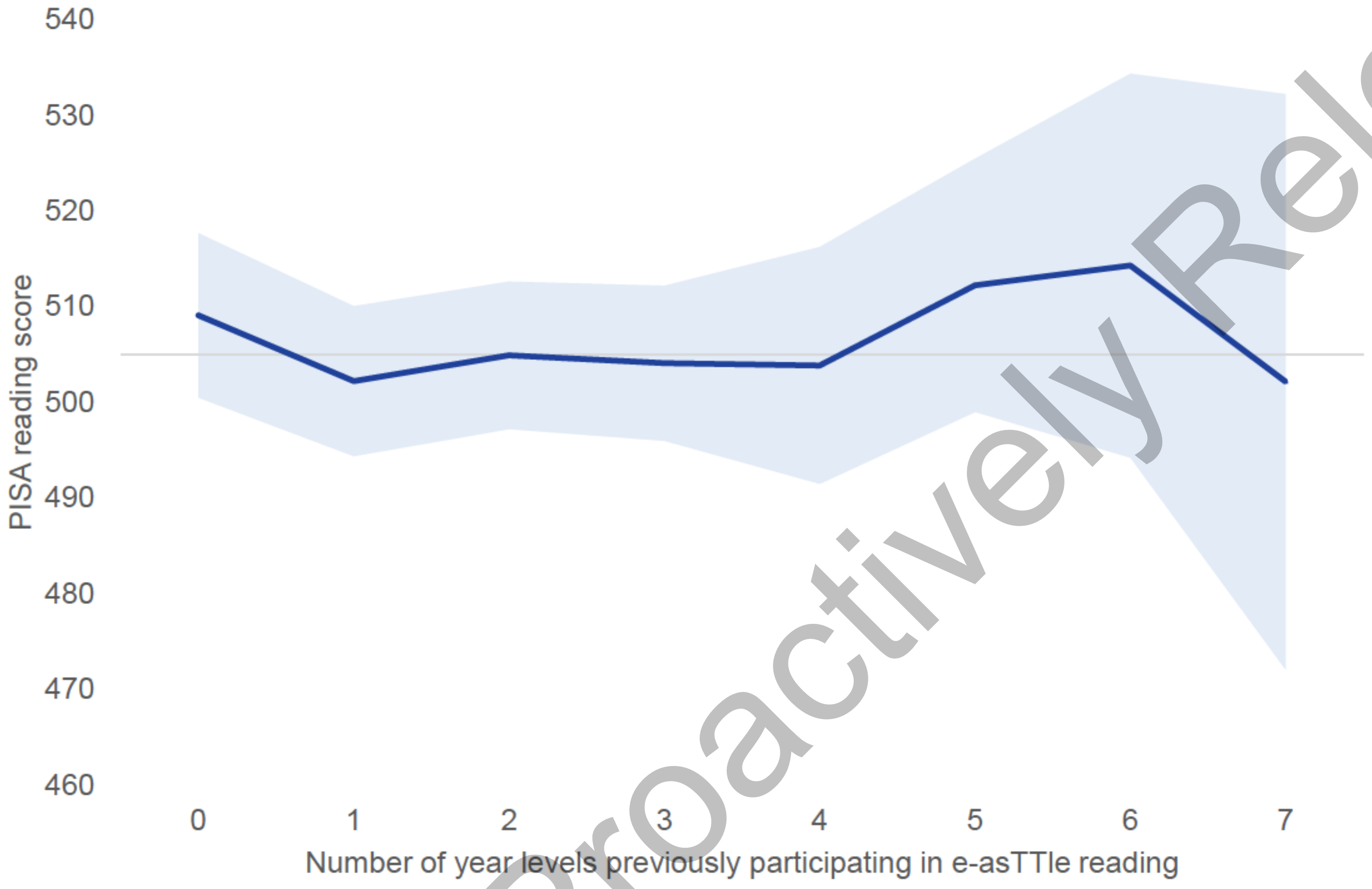
Because the composition of students participating in e-asTTle changes so much from Year 4-10, we should avoid comparing averages across year levels (as this graph does).

These comparisons understate the true progress students are making.

The report this graph was taken from went on to compare progress using linked data between years, which doesn't suffer the same problem.

**When using e-asTTle, we should always use longitudinally-linked data.**

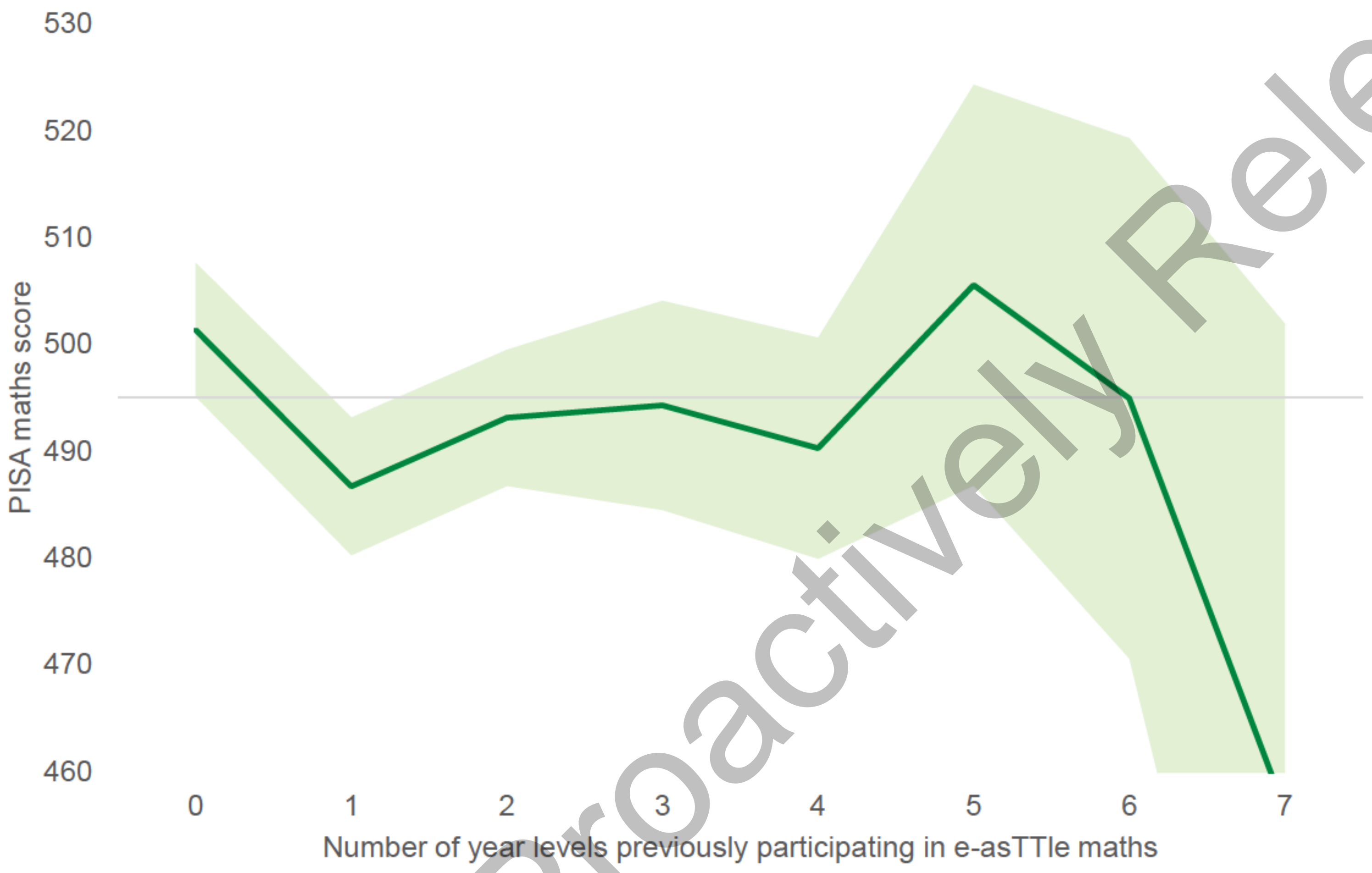
# Students with long histories of e-asTTle reading tests are representative of the broader population



Because some teachers use e-asTTle and some don't, not all students have a long history of consecutive e-asTTle assessments. You might be worried that the students who do have this long history are not representative of the population.

Happily, this graph shows this not to be the case: PISA students with histories of 3-7 previous e-asTTle reading assessments had about the same PISA reading scores as students who had never undertaken an e-asTTle reading test.

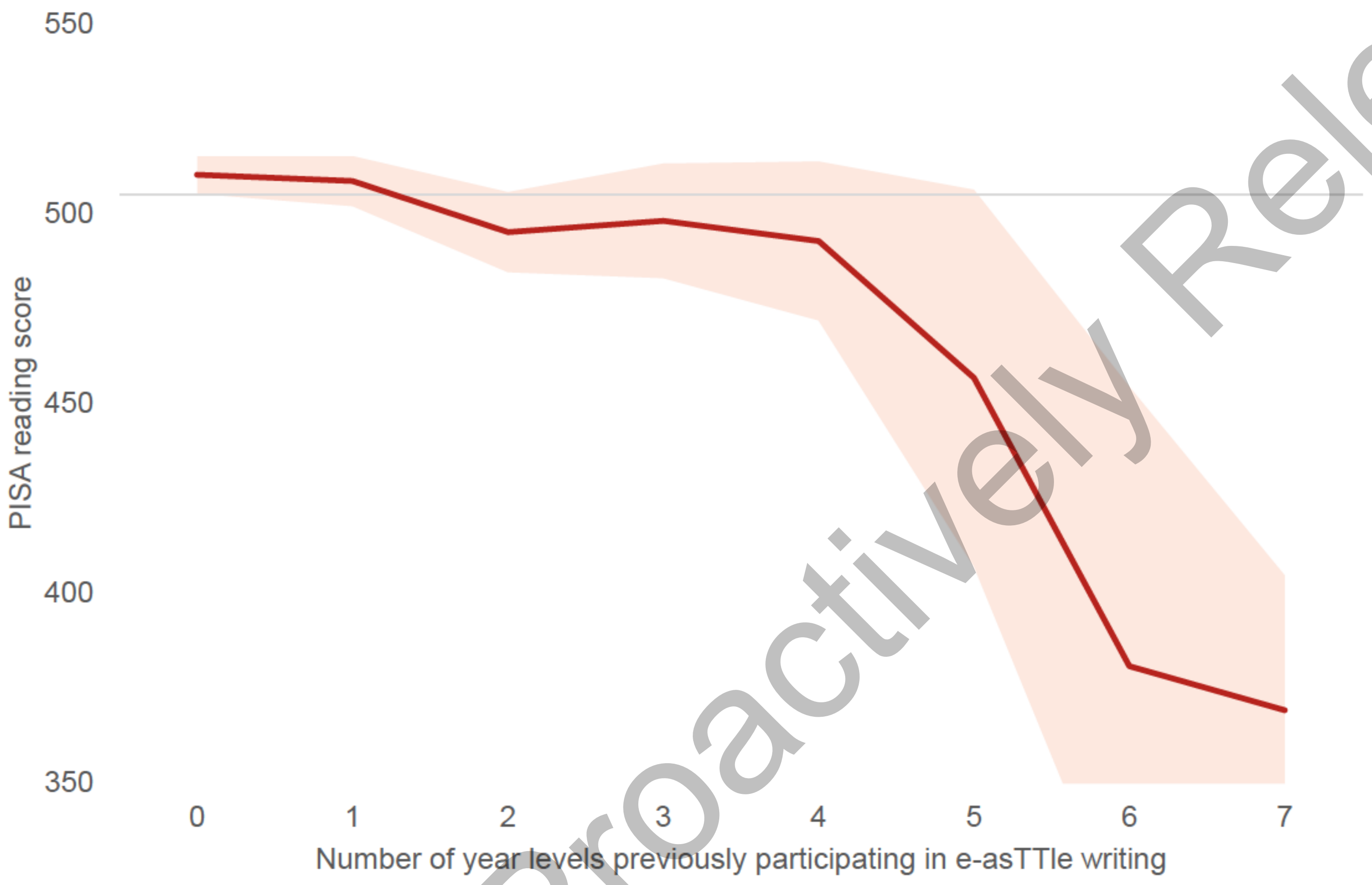
# The same is true in maths



A similar story in maths, although we might be a bit worried about students with only one e-asTTle assessment (significantly lower performers), or those with 5+ (although there aren't that many of these students).

This indicates that if we undertake analysis of longitudinal e-asTTle data where students are linked across 2-4 years, we can have some confidence that our results are broadly applicable to the rest of the population.

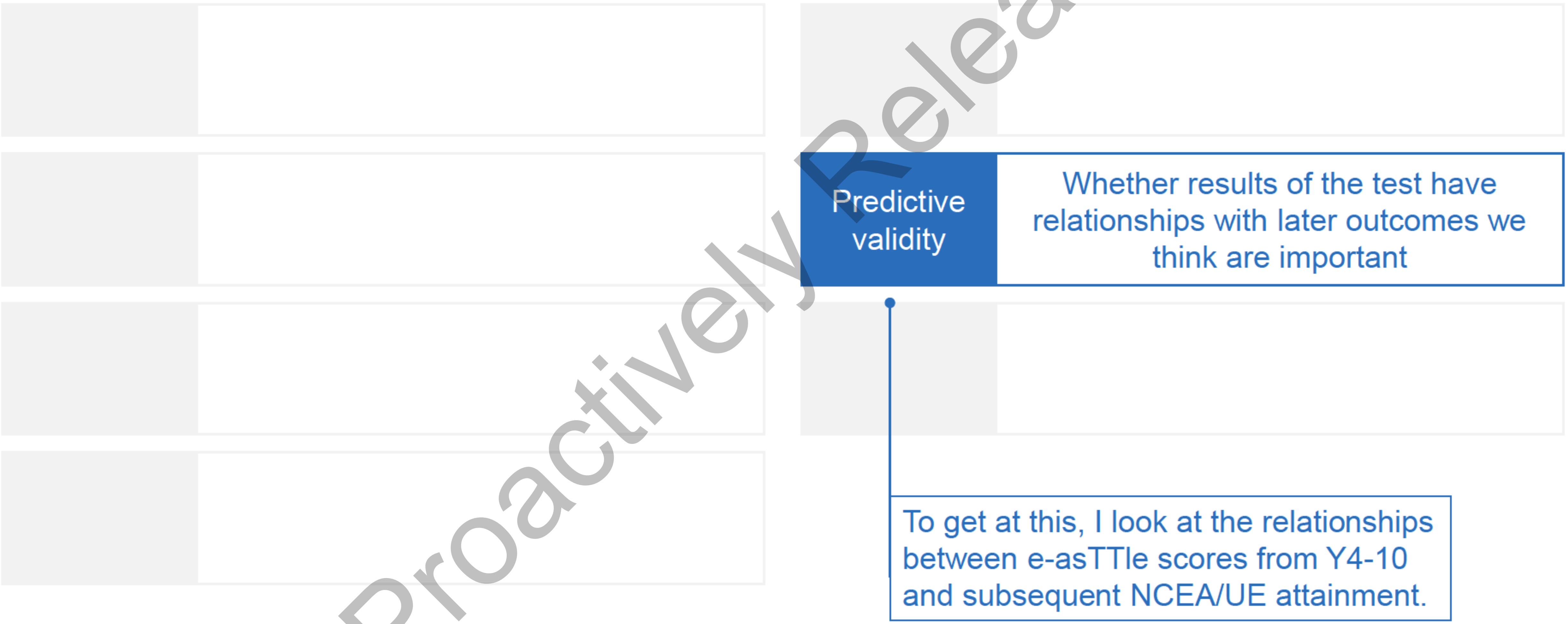
# More problems with writing, though



Students with long (5+ years) histories of e-asTTle writing are clearly not representative, with performance 1-2 standard deviations below those never participating in e-asTTle.

However, we should also worry about the representativeness of students with histories of 2-4 years. Although not statistically significant (due to low numbers of students), this group performed about 20 points (~0.2 standard deviations) below those who never had an e-asTTle writing test.

# For us to put a lot of stock in data, it must be valid, reliable, and representative

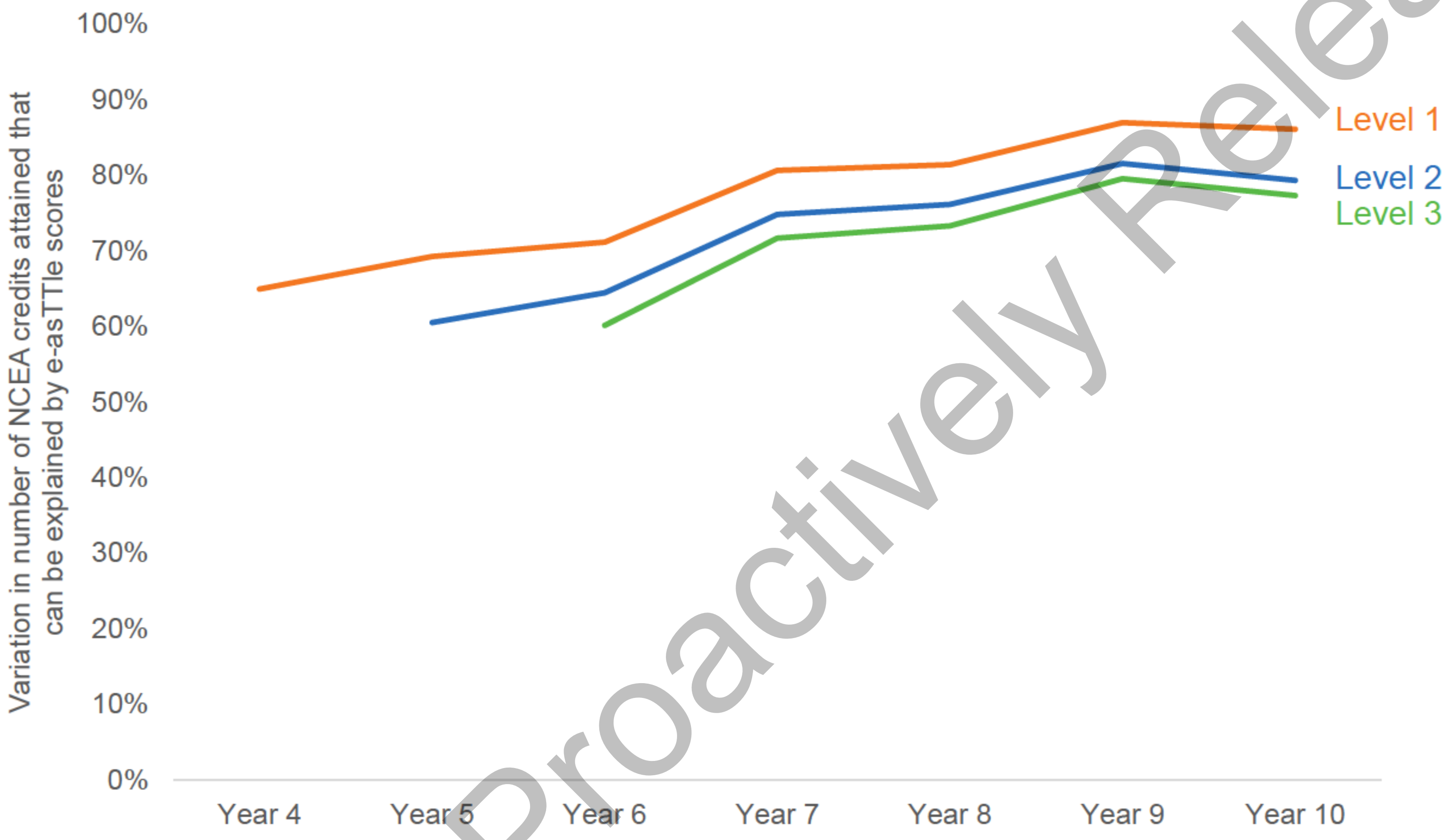


Predictive validity

Whether results of the test have relationships with later outcomes we think are important

To get at this, I look at the relationships between e-asTTle scores from Y4-10 and subsequent NCEA/UE attainment.

# e-asTTle test scores explain almost all variation in future NCEA performance

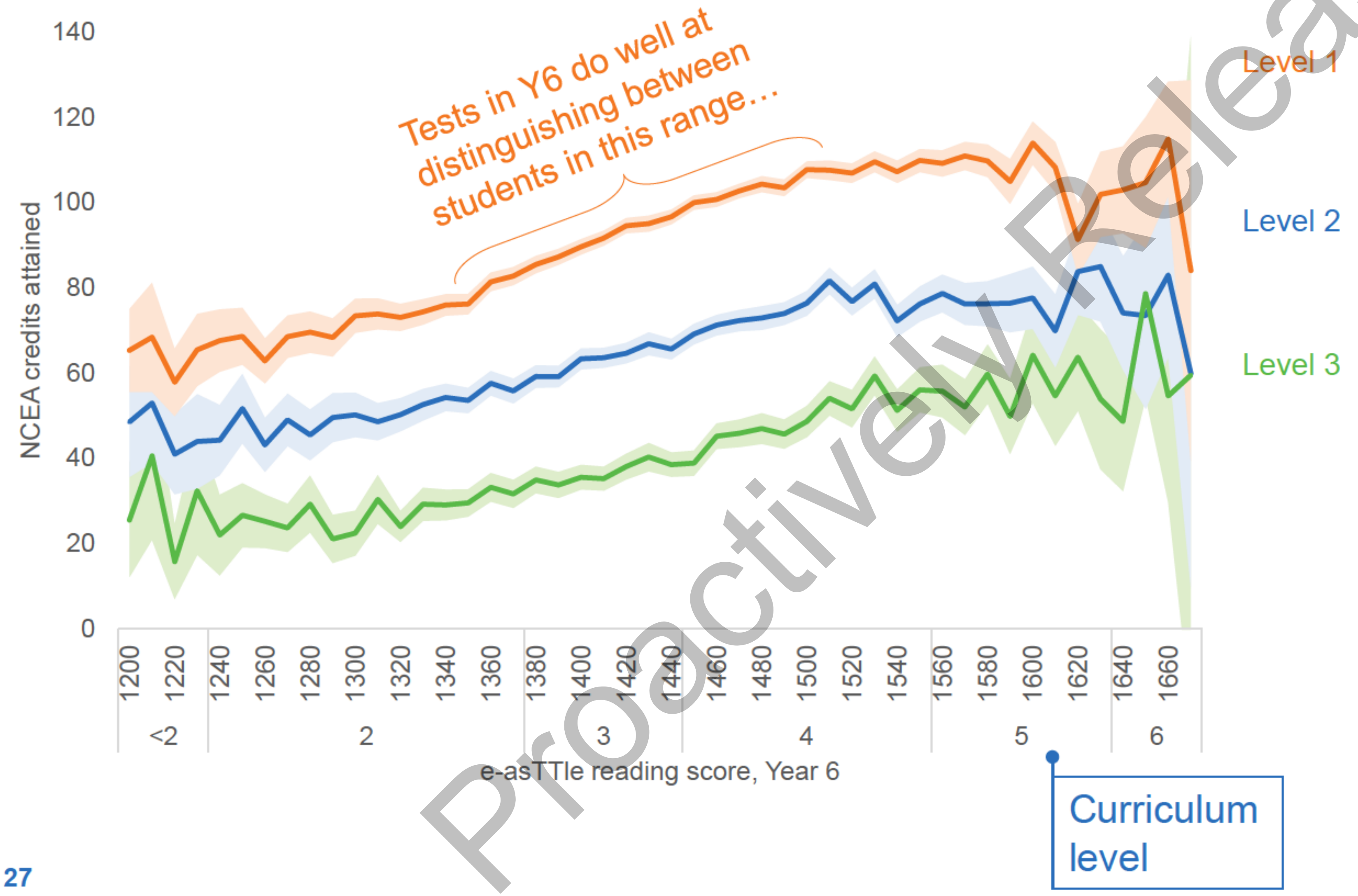


E-asTTle scores are highly predictive of later attainment.

Variation in Year 4 e-asTTle reading scores explains about two-thirds of the variation we later see between those same students in number of NCEA Level 1 credits attained. e-asTTle scores in Year 10 explain 86% of the variation in Level 1 credits.

The model underlying the Equity Index (representing cumulative effect of SES) explains about 35-40%.

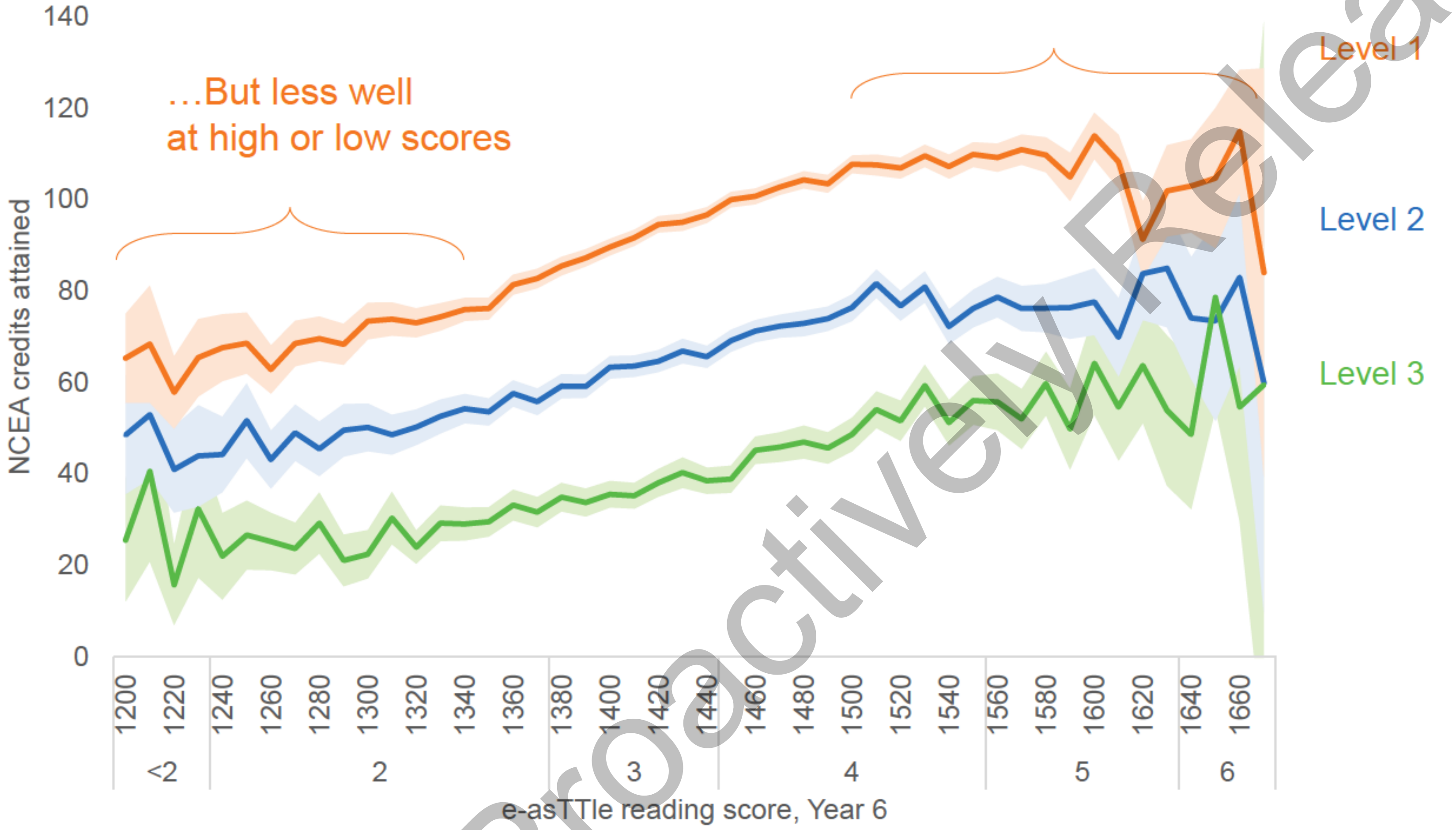
# e-asTTle can finely distinguish between students, provided they are within about a Curriculum Level



For e-asTTle to distinguish between students well, we want to see a sloped line here – students with slightly higher e-asTTle scores end up getting noticeably better NCEA results.

This does happen, but only in the middle of the range…

# e-asTTle can finely distinguish between students, provided they are within about a Curriculum Level

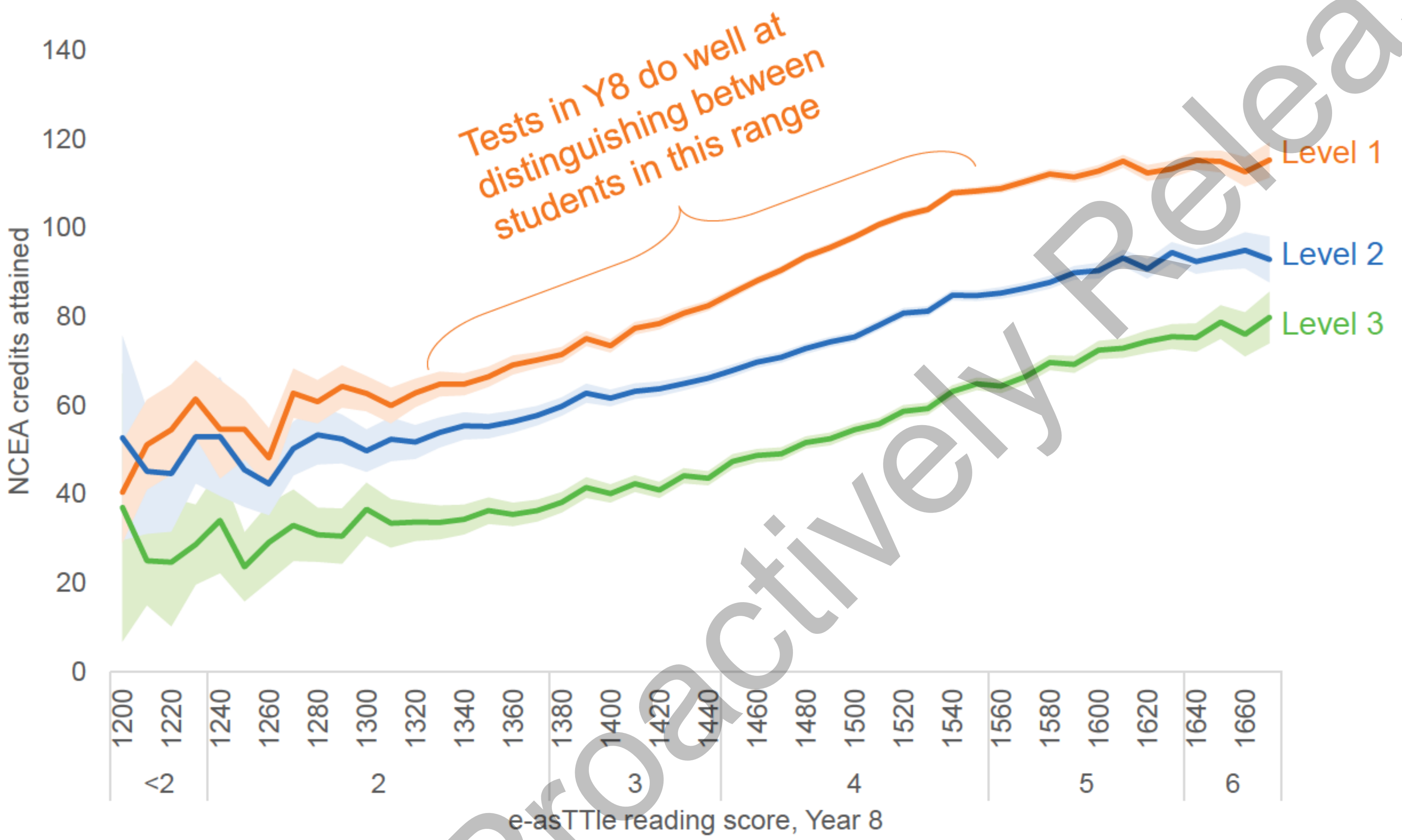


For students who get very high or low Y6 e-asTTle scores, the relationship with NCEA performance is more like a flat line.

This is potentially evidence that e-asTTle is not very good at finely distinguishing between these students.

The median student in Y6 should be operating at the upper end of Curriculum Level 3, and e-asTTle is giving us meaningful data for these students.

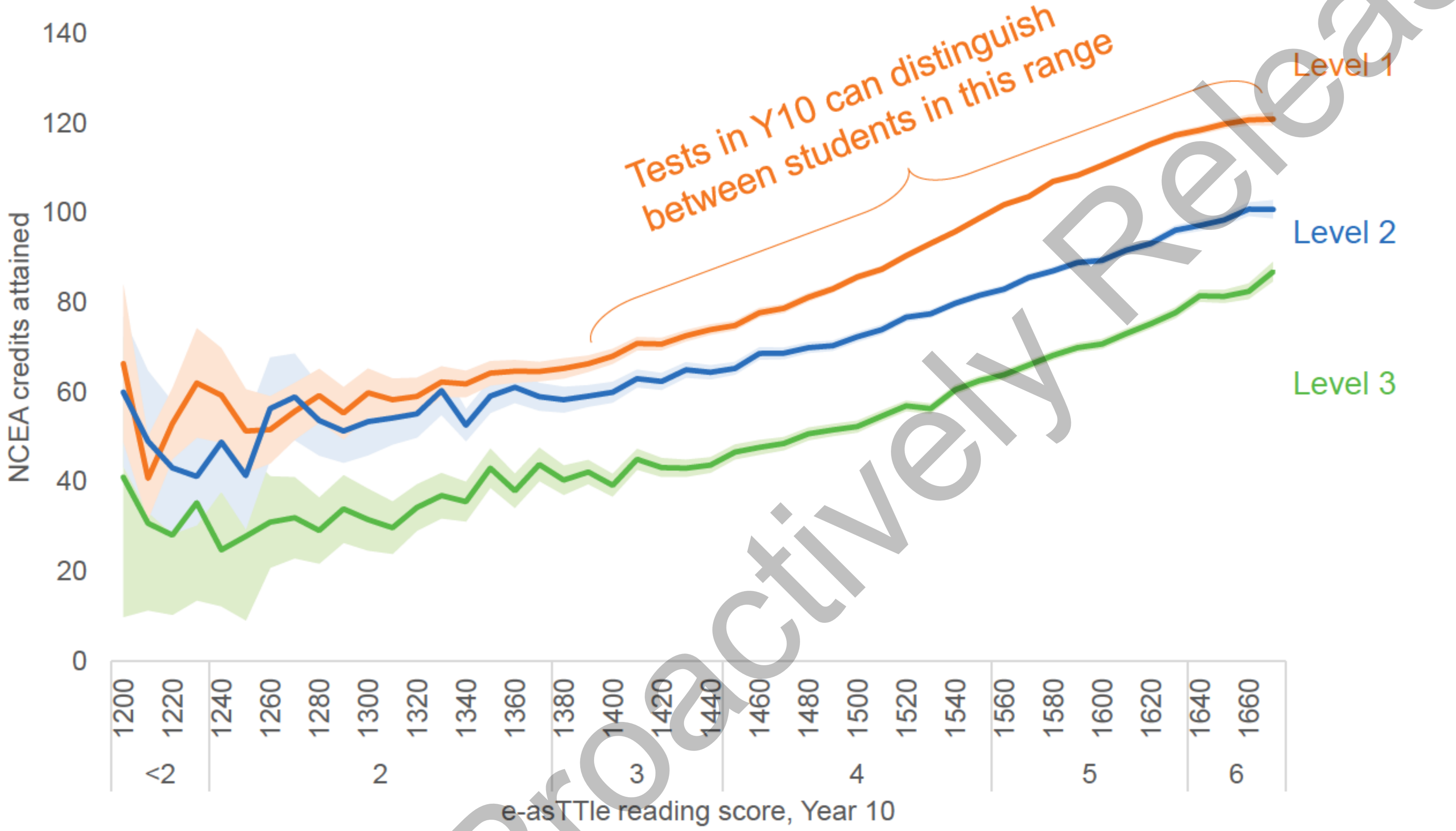
# The range at which e-asTTle can distinguish changes with the year level of the test



As we go up in year levels, e-asTTle is more able to distinguish between students at the upper end of the performance range (and less able to distinguish between students at the lower end).

In Year 8, it can finely distinguish between people from about the middle of CL 2 to the start of CL 5.

# At Year 10, e-asTTle can distinguish between students above Curriculum Level 2



In Year 10, e-asTTle can distinguish between almost all students in Curriculum Levels 3-6.

There are similar patterns for e-asTTle maths, and when predicting the probability of attaining NCEA (rather than number of credits), although probability of attainment is less precise of an outcome (because most people attain NCEA).

# Summary/Implications

# This work finds substantial evidence that e-asTTle scores are valid and reliable

- Broadly consistent with other psychometric work on e-asTTle previously undertaken by NZCER
- Some caveats/findings of concern:
  - e-asTTle is not representative of the ability of the student population in Years 4-6 (where it overrepresents higher performers) or Years 9-10 (where it overrepresents lower performers)
  - The Māori medium assessments in e-asTTle are very infrequently used and are substantially less reliable than their English medium counterparts (the lack of other available assessment data to benchmark them against means it is difficult to assess their overall validity)
  - Writing scores generally show weaker validity and far larger selection effects than reading or maths
  - There is evidence that e-asTTle is not very good at finely distinguishing the performance of students who are operating well above or below the expected curriculum level

# Implications for EDK analysts

- e-asTTle is a strong proxy for later educational attainment and has solid measurement properties – we should be using it more as an outcome measure.
  - We used it to some success in the recent Reading Recovery evaluation
  - It is worth exploring moving into the IDI, given the current lack of outcome measures prior to senior secondary school (both within and outside of education)
- All analysis needs to account for the selection effects shown here – **never** compare unadjusted attainment between year levels, for example
- Results support analysing e-asTTle data longitudinally across a 2-4 year horizon – this is likely to give us a representative view of all (English-medium) students
- Previous e-asTTle analysis used business rules that threw out scores more than two curriculum levels away. There is some evidence this was justified, but might be worth exploring top/bottom-coding scores instead?
- Strong need for exploratory analysis on the way teachers use e-asTTle to assess students (formative assessment behaviours)

# Implications for policy

- To the extent that teachers/schools are not using e-asTTle because of concerns around accuracy/usefulness of results, these findings could be reassurance.
- Measurement solutions relating to Curriculum Progress and Achievement should build off, rather than substitute for, the demonstrated measurement strength of e-asTTle.
- Potential application in using relationships with NCEA to report predicted future outcomes for teachers alongside scores (if student stays on this path…)? Especially regarding literacy/numeracy requirements.
- Need to stress that although there are strong relationships with future outcomes, plenty of scope for quality teaching to influence trajectories.
- e-asTTle has tremendous potential as a dataset used by researchers (within and outside of the Ministry). Challenge is how to set up governance to maximise utility but restrict uses that are unethical or threaten the integrity of the data (eg school/teacher league tables).
  - The IDI could help here.
  - And/or an established formal process to assess applications to use data for specific research purposes (eg Growing Up in New Zealand data)