Using junior assessments and statistics to ‘prove’ impact on achievement and learning

Darcy Fawcett, darcy@sounddata.co.nz

Director, Sound Data.

Today’s resources:

- Preliminary survey bit.ly/4sdxgEF
- Online dashboards sounddata.co.nz/dashboard/

SOUND DATA

Removing Noise From Decision Making

How can you ‘prove’ the impact of initiatives on learning?

- You must compare assessment data of those who experienced the initiative with those who didn’t
- The initiative group must make *significantly* more progress than the comparison group *and* this effect must be large enough to be worth attending to.

Experimental method

- Compare the performance of two equivalent groups
  - control and experiment group
- Often very difficult to arrange and can be unethical

Quasi-experimental method

- Compare naturally occurring groups
  - Compare the performance of latest cohort with historical performance
- Assumptions
  - latest cohort is typical
  - The historical record is large enough to describe the population

Summary statistics

Measures of centre (averages)

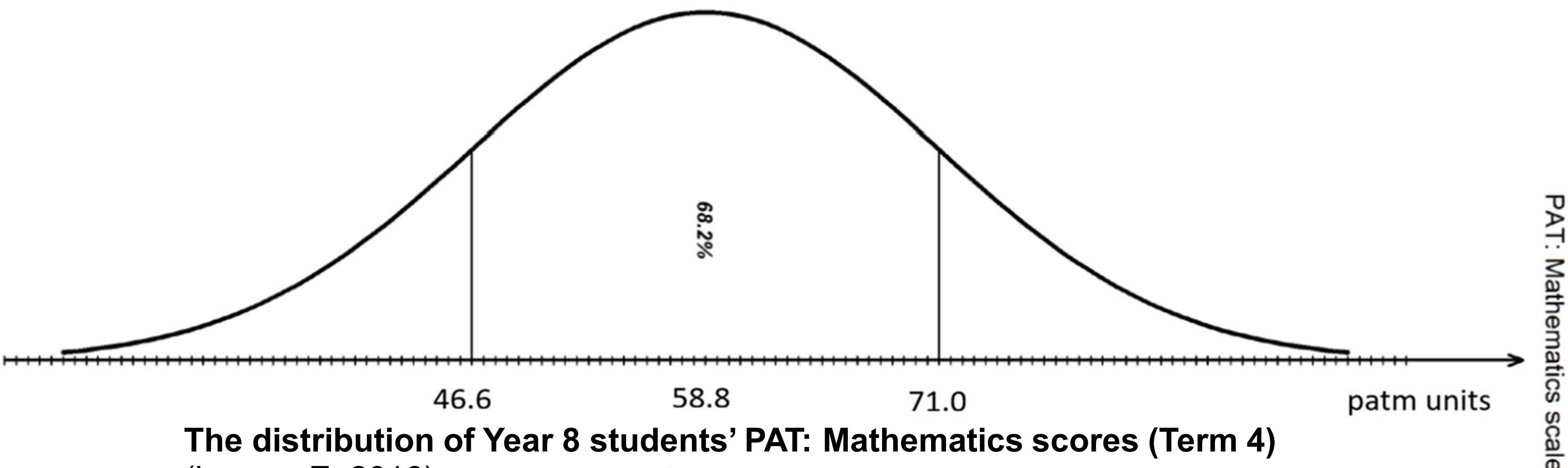
- mode
  - most frequently occurring value
- Medium
  - the middle of ranked data
- mean
  - $\mu = \bar{x} = \frac{x_1+x_2+x_3+\cdots}{n}$

Measures of spread (distribution)

- range
  - $R = x_{max} - x_{min}$
- interquartile range
  - $IQR = x_{Q3} - x_{Q1}$
- standard deviation
  - $\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots}{n}}$

Continuous scale data - parametric

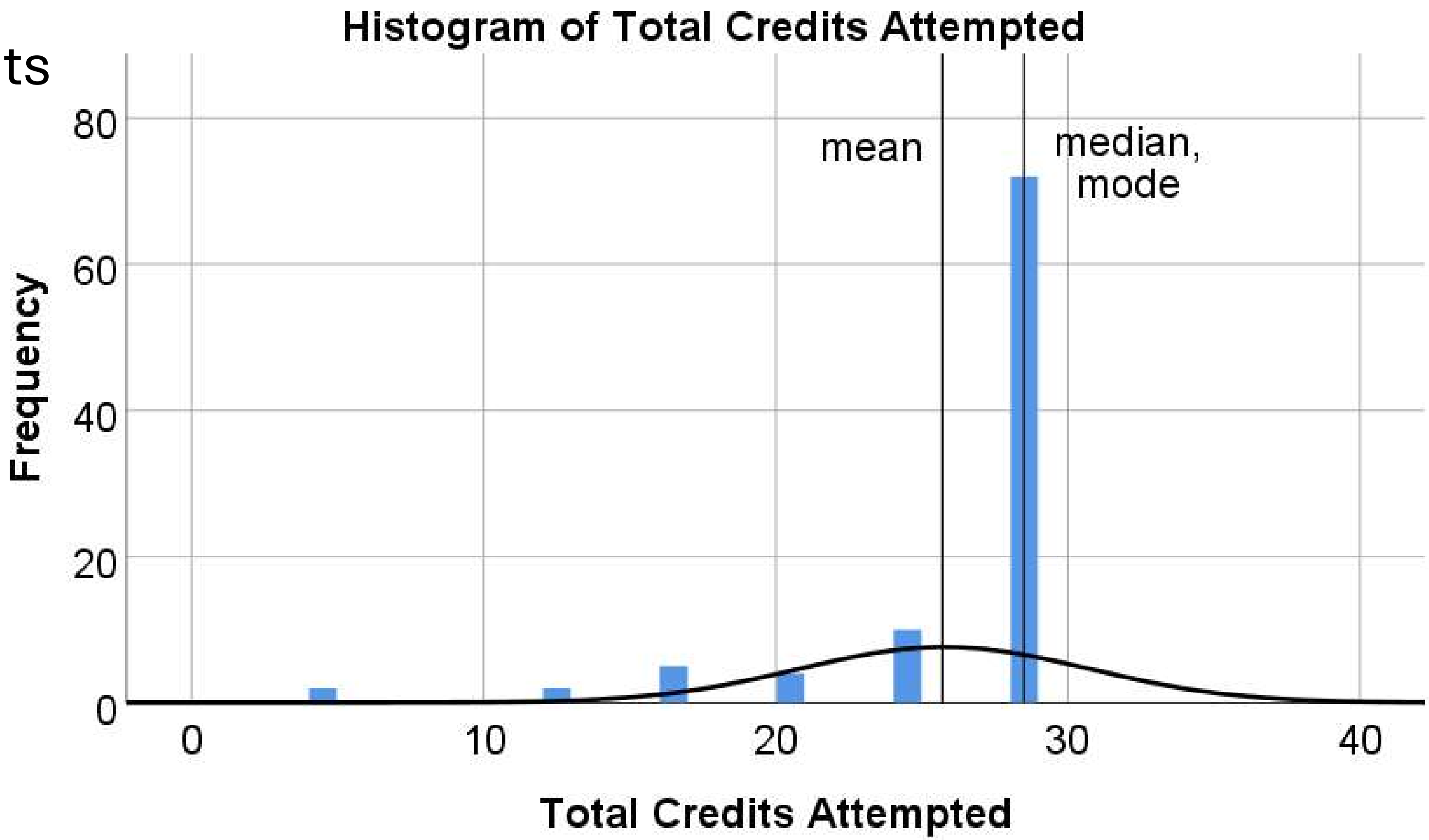
- **Continuous scale:** +1 has consistent meaning, so progress from all start points has same meaning
- **Parametric:** normally distributed (bell curve), mean = median = mode
- e.g. PAT and e-asTTle scale scores, IQ, heights



The distribution of Year 8 students' PAT: Mathematics scores (Term 4)
(Lawes, E, 2016) mean = 58.8, sd = 12.2 patm

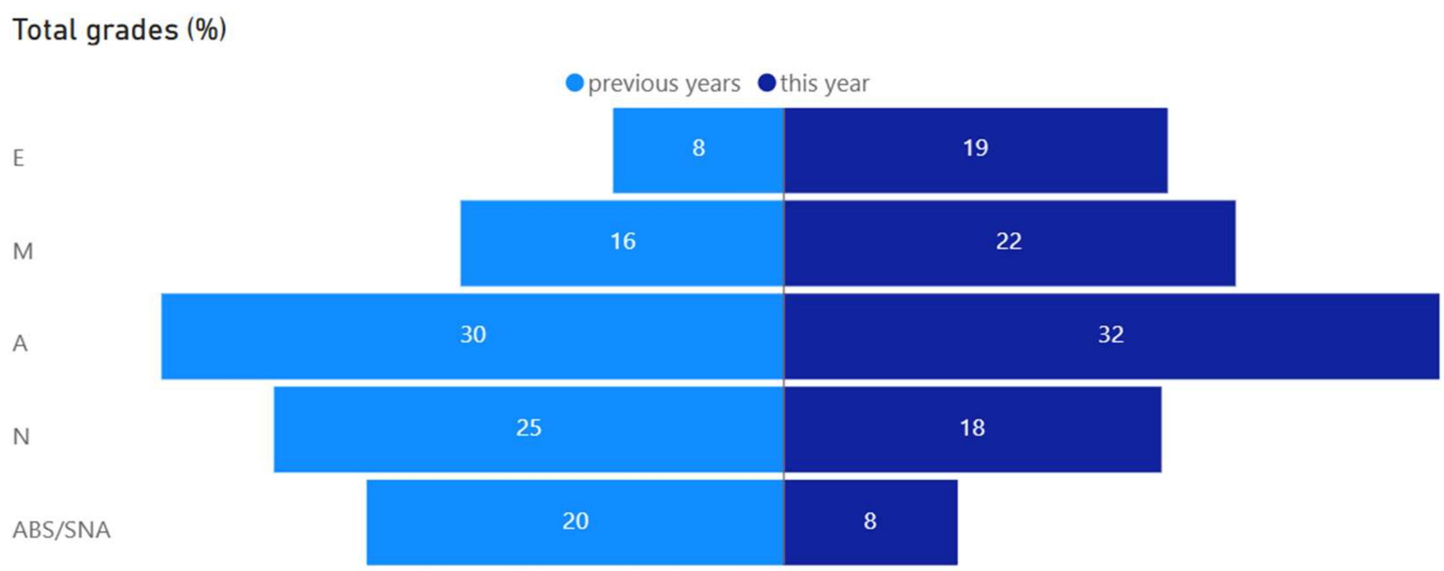
Continuous scale data – non-parametric

- **not** normally distributed
- E.g. NCEA credits



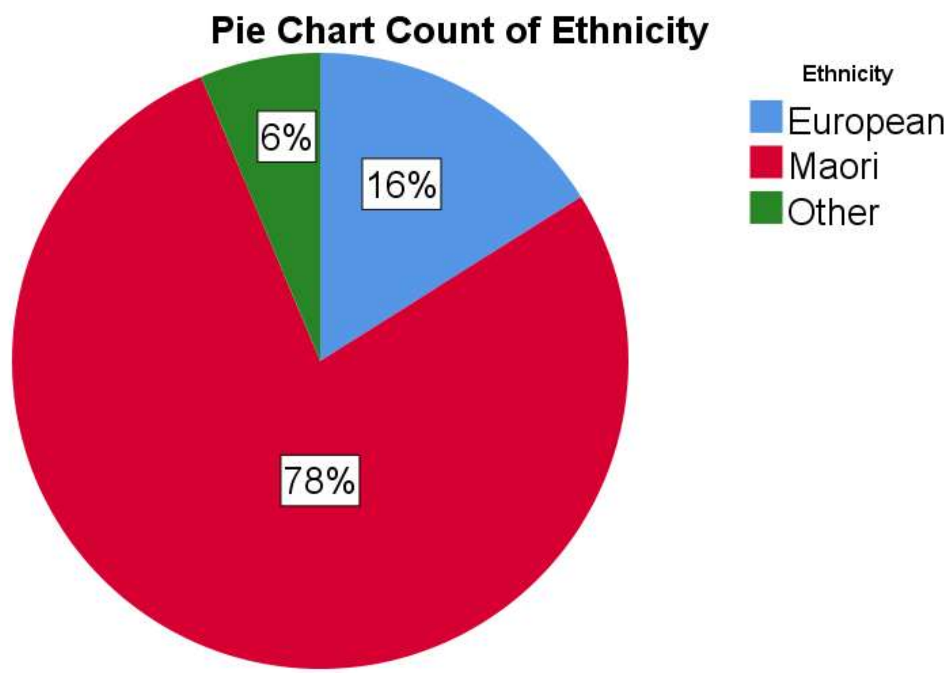
Categorical Ordinal data

- Categories that have order
- e.g. curriculum levels, achievement standards …



Categorical nominal data

- Categories that have no order
  - Count/frequency or percentage
- Summary: mode
- e.g. Gender, ethnicity, cohor



Statistical tests for significance

p: the probability that your observations are a random occurrence

Test result	probability
similar to / no significant difference	$p > 0.10$
notable difference	$0.10 \geq p > 0.05$
significant difference	$0.05 \geq p > 0.01$
highly significant difference	$0.01 \geq p$

The p-value acts as a filter for changes that are statistically similar to the normal vaiations that occur from year to year

Associated measure for strength of association or effect size

Effect size	ES
None/minimal	$0.0 \le |ES| < 0.1$
Weak/small	$0.1 \le |ES| < 0.3$
Moderate	$0.3 \le |ES| < 0.5$
Strong/large	$0.5 \le |ES|$

- Large populations can show significant relationships for very small shifts. Use related tests to show the strength of the association or effect size
  - The relative size of the skew on the graph
- Which measure you use depends on the nature of the test and variables E.g.,
  - Scale by nominal independent sample or scale by expected value (e.g., Cohen’s D)
- These trigger points are context dependent and should be varied depending on data properties and theoretical assumptions (Cohen 1969; Kraft 2020).

Tests for significant differences and associated effect sizes

<table>
<tr><td colspan="2">**Variables**</td><td colspan="2">Tests for significance (and effect size)</td></tr>
<tr><td>Independent</td><td>Dependent = 1</td><td>Independent samples</td><td>Related samples</td></tr>
<tr><td>**0 variables**
- Theoretical</td><td>**Categorical, Ordinal**
National rates, averages, distributions</td><td colspan="2">**Cat2: One sample binomial test (Rosenthal's *R*)**
- e.g., Year Level Achieved, School vs National %
Cat3+: One sample chi-sq test (Phi)
- e.g., Level Qualifications, School vs National distribution
Non-parametric scale: One sample median (Rosenthal's *R*)
- e.g., Credits gained, School vs National average ;
Parametric scale: One sample t-test (Rosenthal's *R*)
- e.g., PAT Scale Scores, School vs National average ;</td></tr>
<tr><td rowspan="3">**1 variable**
2 levels
- Cohort
- non/Māori
- Gender

3+ levels
- Prioritised Ethnicity
- Year
- Room</td><td>**Ordinal, Ord ≤ 3**
Course Endorsement, UE, Below/At+</td><td>**Chi-Square Linear-by-Linear Association test (Gamma)**
e.g., UE; this year vs previous years.</td><td>**McNemar's test (Phi, AR)**
- e.g., Below/At+; T1 vs T4</td></tr>
<tr><td>**Ordinal, Ord ≥ 4**
Year **L**evel **Q**ual.
Achievement **S**tds
Curriculum **L**evels
Non-Parametric Scale
Credits, SMART</td><td>1IV2L: **Wilcoxon-Mann-Whitney U rank test (R)**
- e.g., YLQ; male vs female.
1IV3+L: **Kruskal-Wallis Anova by ranks (Eta)**
- e.g., AS; Prioritised Ethnicity.
1IV3+L: **Anom of transformed ranks.**
- e.g., progress on CL; School vs all rooms.</td><td>1IV2L: **Wilcoxon signed ranks test (Cohen's D)**
- e.g., Curriculum Levels; T1 vs T4</td></tr>
<tr><td>**Parametric scale**
PAT scale scores
e-asTTle s. scores
PaCT s. scores</td><td>1IV2L: **Independent samples t-test (R)**
- e.g., Scale score, Māori vs non-Māori.
1IV3+L: **Analysis of Variance by means (Eta).**
- e.g., progress on s.sc; School vs Room1.
1IV3+L: **Analysis of means, Anom.**
- e.g., progress on s.sc; School vs all rooms.</td><td>1IV2L: **Paired samples t-test (D)**
- e.g., Scale score; T1 vs T4</td></tr>
</table>

https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/

The independent samples t-test

- 1 independent variable with 2 levels: ty v py
- 1 parametric scale dependent variables: scale scores, Achievement or Progress

Cohort longitudinal comparison

Are this year's results similar to previous years? Do any subgroups stand out?

PAT Comprehension	T1

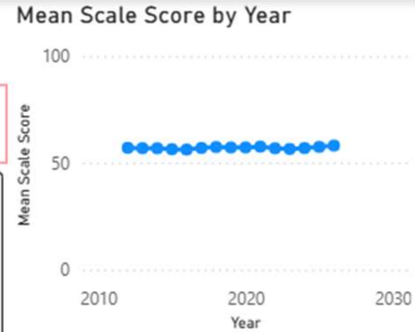
Gender: All
Ethnicity: All
Year Level: All
Year: All
ESOL: All

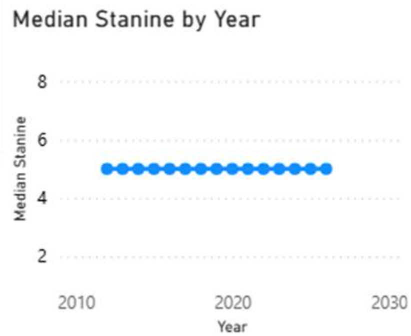
Cohort	n	Mean SS	Med St
previous years	9905	56.9	5
this year	726	58.0	5
Total	**10631**	**57.0**	**5**

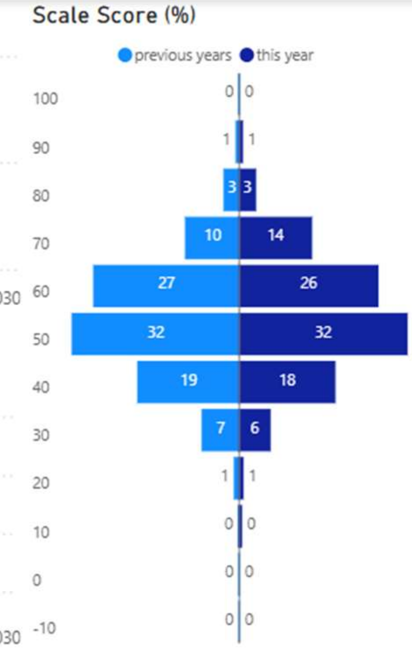
Search Student Name or NSN

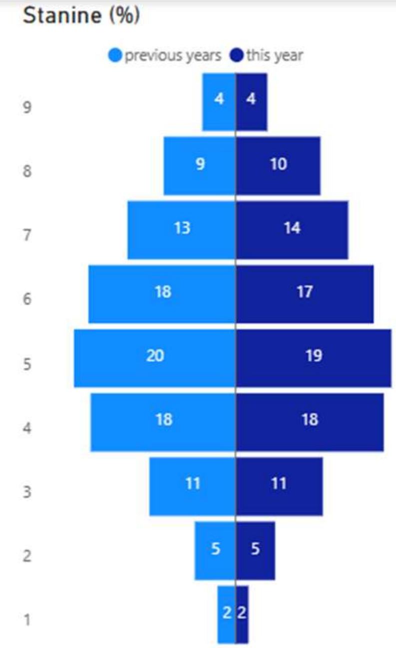
Search

Year	NSN	Student Name	SS	ST
2015	100003	Student, 100003	51.9	4
2018	100004	Student, 100004	69.9	8
2021	100005	Student, 100005	47.9	4
2018	100006	Student, 100006	63.3	7
2012	100009	Student, 100009	71.4	7
2012	100010	Student, 100010	58.5	6
2025	100011	Student, 100011	62.5	7









YL	Dem	Statement	n py	n ty	mean SS py	mean SS ty	med St py	med St ty	p	ES
All	All	minimal significant improvement	9905	726	56.9	58.0	5.0	5.0	0.018	0.09
Y07	All	NO SIGNIFICANT DIFFERENCE, improvement	5363	417	54.5	55.6	5.0	5.0	0.077	0.09
Y08	All	small significant improvement	4542	309	59.8	61.3	5.0	5.0	0.023	0.13
All	Asian	NO SIGNIFICANT DIFFERENCE, improvement	2003	155	57.2	57.9	5.0	5.0	0.489	0.06
Y07	Asian	NO SIGNIFICANT DIFFERENCE, improvement	1094	98	54.9	57.1	5.0	5.0	0.086	0.18
Y08	Asian	NO SIGNIFICANT DIFFERENCE, reduction	909	57	60.1	59.3	5.0	5.0	0.631	-0.07
All	ESOL	NO SIGNIFICANT DIFFERENCE, improvement	510	35	57.0	57.0	5.0	5.0	0.995	0.00
Y07	ESOL	NO SIGNIFICANT DIFFERENCE, improvement	286	21	54.8	57.0	5.0	6.0	0.475	0.16
Y08	ESOL	INVALID TEST, n < 20, reduction	224	14	59.8	57.0	5.0	4.0	0.381	-0.24
All	Female	NO SIGNIFICANT DIFFERENCE, improvement	5010	362	57.4	58.4	5.0	5.0	0.131	0.08

Year Level and Demographics
(All, Y07, Y08, Male, Female, Māori, Pacifica, Asian, MELAA, Other Ethnicities, Pakeha, ESOL)

number of students
(previous years or this year)

mean Scale Score and median Stanine
(previous years or this year)

YL	Dem	Statement	n py	n ty	mean SS py	mean SS ty	med St py	med St ty	p	ES
All	All	minimal significant improvement	9905	726	56.9	58.0	5.0	5.0	0.018	0.09
Y07	All	NO SIGNIFICANT DIFFERENCE, improvement	5363	417	54.5	55.6	5.0	5.0	0.077	0.09
Y08	All	small significant improvement	4542	309	59.8	61.3	5.0	5.0	0.023	0.13

Computer generated statistical conclusion
(all you need if you don’t like maths!)

the p-value

The effect size

Gender cross-sectional comparison

How does the achievement of Females compare to Males?

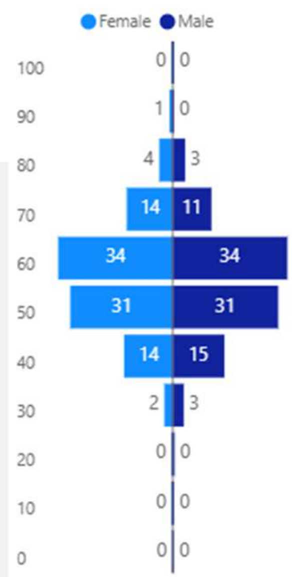
PAT Comprehension	T1

Year Level

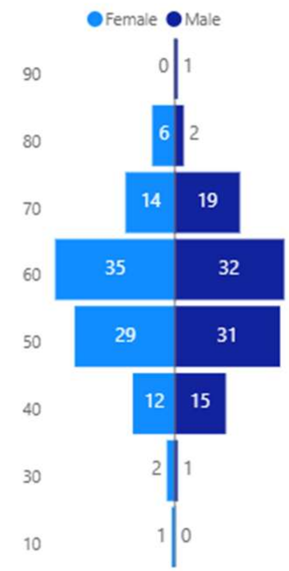
Y08

YL	Cohort	Statement	n F	n M	mean SS F	mean SS M	med St F	med St M	p	ES
All	this year	NO SIGNIFICANT DIFFERENCE, Female > Male	362	364	58.4	57.6	5.0	5.0	0.384	0.06
All	previous years	minimal highly significant difference, Female > Male	5010	4895	57.4	56.4	5.0	5.0	0.000	0.08
Y07	this year	NO SIGNIFICANT DIFFERENCE, Female > Male	222	195	56.7	54.4	5.0	5.0	0.064	0.18
Y07	previous years	minimal highly significant difference, Female > Male	2743	2620	55.0	54.0	5.0	5.0	0.004	0.08
Y08	this year	NO SIGNIFICANT DIFFERENCE, Male > Female	140	169	61.2	61.4	5.0	5.0	0.864	-0.02
Y08	previous years	small highly significant difference, Female > Male	2267	2275	60.4	59.2	5.0	5.0	0.000	0.11

Scale Score (%) previous years



Scale Score (%) this year

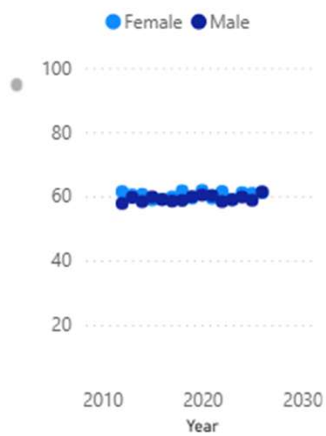


Search Student Name or NSN

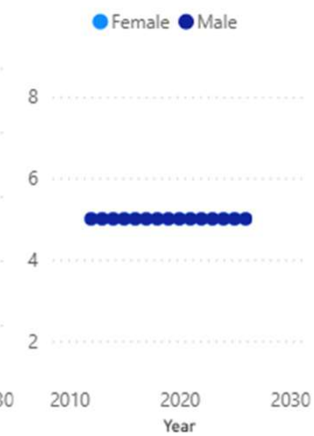
Search

Year	NSN	Student Name	Gender	SS	ST
2012	109333	Student, 109333	Male	2.3	1
2012	107471	Student, 107471	Male	7.3	1
2012	104930	Student, 104930	Male	18.5	1
2012	107218	Student, 107218	Male	18.5	1
2012	104776	Student, 104776	Female	36.0	1
2012	101636	Student, 101636	Male	36.0	1
2012	101271	Student, 101271	Female	37.7	1
2012	102295	Student, 102295	Male	37.7	1

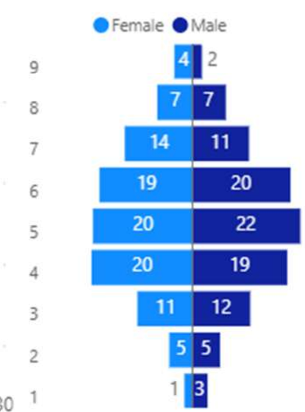
Mean Scale Score by Year



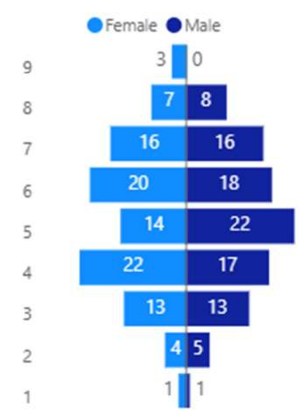
Median Stanine by Year



Stanine (%) previous years



Stanine (%) this year



As a continuous scale we can measure and compare progress

Cohort longitudinal comparison (all YL)

Did students make more progress this year? Do any subgroups stand out?

PAT Mathematics | T1 to T4

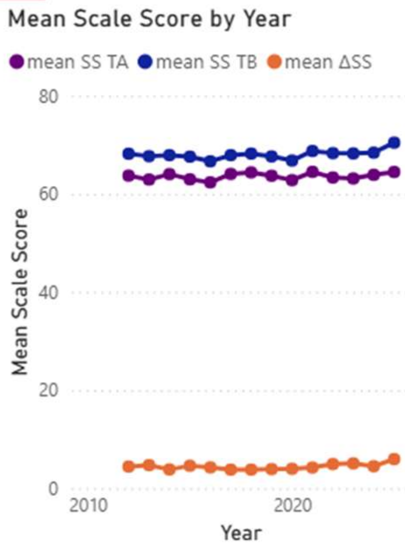
Gender: All
Ethnicity: All
Year Level: All
Year: All
ESOL: All

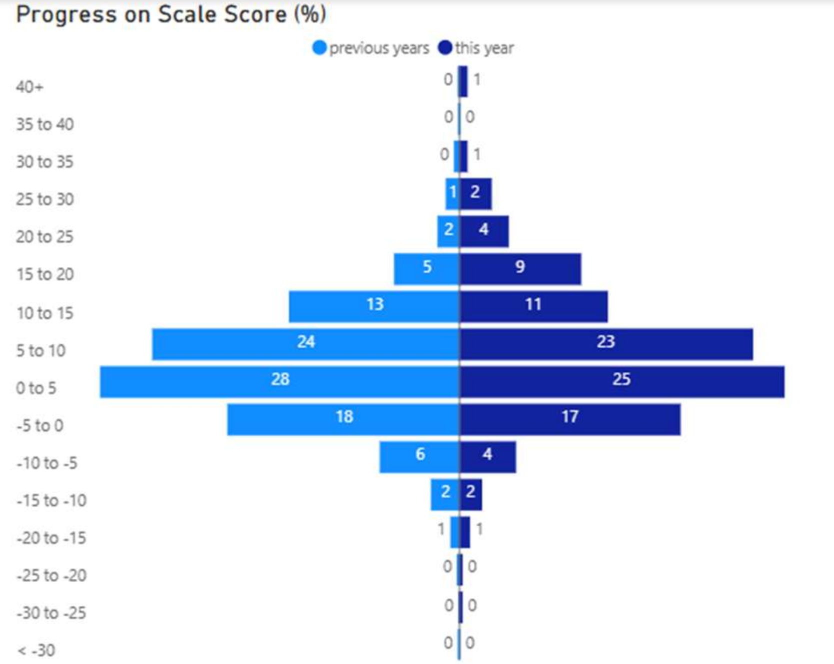
Cohort	N	Mean SS TA	Mean SS TB	Mean ΔSS
this year	534	64.5	70.4	6.0
previous years	6369	63.5	67.9	4.3
Total	**6903**	**63.6**	**68.1**	**4.5**

Search Student Name or NSN

Search

Year	NSN	Student Name	SS TA	SS TB	ΔSS
2025	100011	Student, 100011	71.9	65.2	-6.7
2025	100022	Student, 100022	56.5	60.0	3.5
2025	100055	Student, 100055	59.7	84.2	24.5
2025	100077	Student, 100077	41.5	41.2	-0.3
2025	100087	Student, 100087	75.1	95.9	20.8
2025	100094	Student, 100094	84.2	103.4	19.2
2025	100166	Student, 100166	53.1	72.0	18.9





YL	Dem	Statement	n py	n ty	mean ΔSS py	mean ΔSS ty	p	ES
All	All	small highly significant improvement	6369	534	4.3	6.0	0.000	0.20
All	Asian	moderate highly significant improvement	1353	106	4.4	7.4	0.000	0.41
All	ESOL	NO SIGNIFICANT DIFFERENCE, improvement	331	25	5.0	5.0	0.978	0.01
All	Female	small highly significant improvement	3236	261	4.9	6.3	0.005	0.18
All	Male	small highly significant improvement	3133	273	3.8	5.7	0.001	0.22
All	Maori	NO SIGNIFICANT DIFFERENCE, improvement	878	68	5.4	6.8	0.192	0.16
All	MELAA	NO SIGNIFICANT DIFFERENCE, improvement	331	25	5.0	5.0	0.978	0.01
All	Other	INVALID TEST, n < 20, reduction	110	10	4.9	3.6	0.601	-0.17
All	Pacific	small highly significant improvement	1128	89	4.6	6.8	0.009	0.29
All	Pakeha	small highly significant improvement	4240	346	4.6	6.6	0.000	0.24
Y07	All	small significant improvement	2844	232	4.8	6.0	0.029	0.15

Analysis of means, Anom

- 1 independent variable with 3+ levels: Many rooms
- 1 parametric scale dependent variable: scale score or change in scale score

Comparative Group

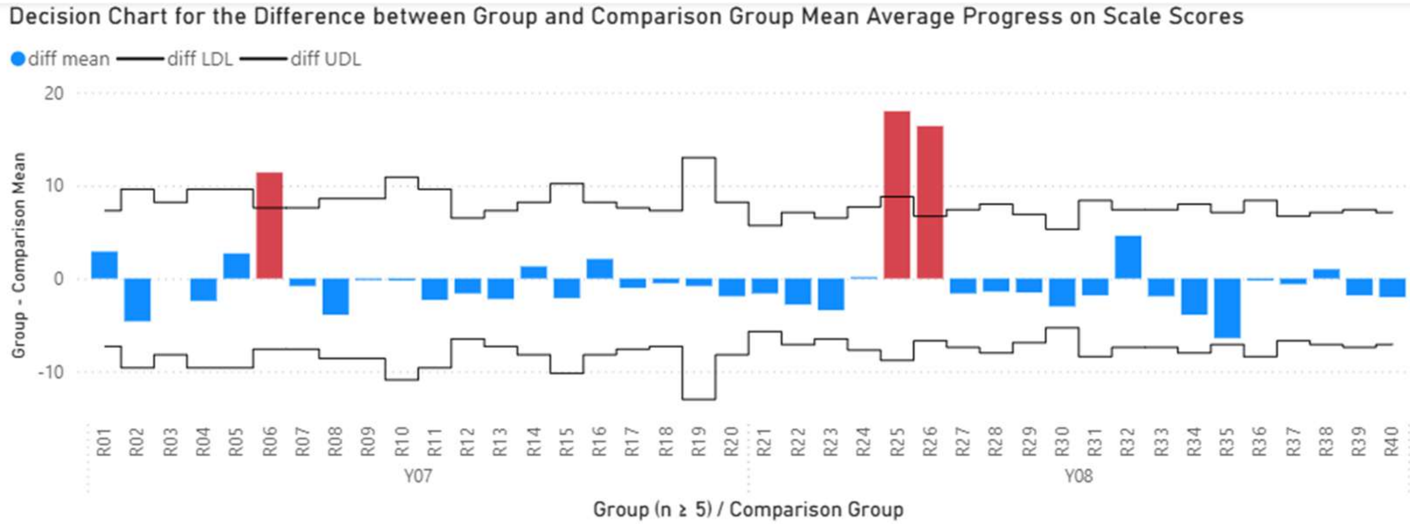
Do any Groups stand outside the normal range for the Comparison Group?

PAT Mathematics

T1 to T4

Group Type
- Room
- Teacher
- YearLevel

Cohort
- previous years
- this year



The Decision Limits are set to a confidence interval of **99%** i.e., **99%** of randomly selected groups of size n (G) should fall inside the decision limits

Cohort	Comparison Group	Group	Statement	n CG	n G	mean ΔSS CG	mean ΔSS G	LDL	UDL	diff mean	diff LDL	diff UDL
this year	Y07	R01	Group mean similar to Comparison mean	232	15	6.0	8.9	-1.4	13.3	2.9	-7.3	7.3
this year	Y07	R02	Group mean similar to Comparison mean	232	9	6.0	1.4	-3.6	15.6	-4.6	-9.6	9.6
this year	Y07	R03	Group mean similar to Comparison mean	232	12	6.0	6.0	-2.3	14.2	0.0	-8.2	8.2
this year	Y07	R04	Group mean similar to Comparison mean	232	9	6.0	3.5	-3.6	15.6	-2.4	-9.6	9.6
this year	Y07	R05	Group mean similar to Comparison mean	232	9	6.0	8.7	-3.6	15.6	2.7	-9.6	9.6
this year	Y07	R06	Group mean significantly greater than Comparison mean	232	14	6.0	17.3	-1.6	13.6	11.4	-7.6	7.6
this year	Y07	R07	Group mean similar to Comparison mean	232	14	6.0	5.2	-1.6	13.6	-0.8	-7.6	7.6
this year	Y07	R08	Group mean similar to Comparison mean	232	11	6.0	2.1	-2.7	14.6	-3.9	-8.6	8.6
this year	Y07	R09	Group mean similar to Comparison mean	232	11	6.0	5.9	-2.7	14.6	-0.1	-8.6	8.6
this year	Y07	R10	Group mean similar to Comparison mean	232	7	6.0	5.7	-4.9	16.9	-0.2	-10.9	10.9
this year	Y07	R11	Group mean similar to Comparison mean	232	9	6.0	3.7	-3.6	15.6	-2.3	-9.6	9.6
this year	Y07	R12	Group mean similar to Comparison mean	232	19	6.0	4.4	-0.5	12.4	-1.6	-6.5	6.5
this year	Y07	R13	Group mean similar to Comparison mean	232	15	6.0	3.8	-1.4	13.3	-2.2	-7.3	7.3
this year	Y07	R14	Group mean similar to Comparison mean	232	12	6.0	7.3	-2.3	14.2	1.3	-8.2	8.2
this year	Y07	R15	Group mean similar to Comparison mean	232	8	6.0	3.9	-4.2	16.2	-2.1	-10.2	10.2

One sample t-test

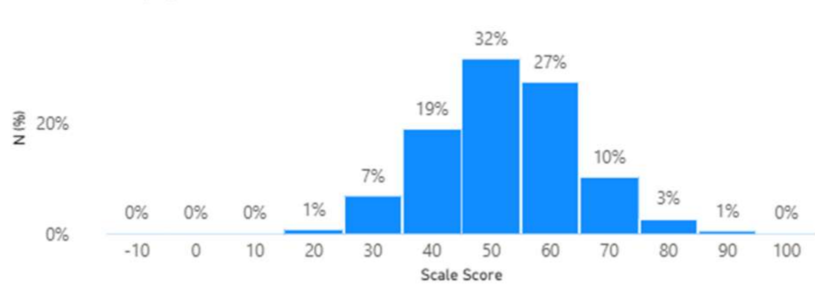
- 0 independent variable: theoretical comparison
- 1 ordinal dependent variable: scale score or change in scale score

National Comparison

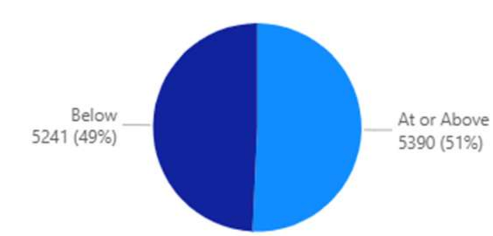
Are students achieving at the expected level?
Do any subgroups stand out?

PAT Comprehension | T1

Scale Score (%)



At or Above Expected Scale Score?



Gender: All
Ethnicity: All
Year Level: All
Cohort: All
ESOL: All

Cohort	N	Mean SS Sc	Mean SS Nat
previous years	9905	56.9	56.5
this year	726	58.0	56.3
Total	**10631**	**57.0**	**56.5**

YL	Cohort	Dem	Statement	n Sc	mean SS Sc	mean SS Nat	mean diff	p	ES
All	previous years	All	minimal highly significant difference, Sch > Nat	9905	56.9	56.5	0.4	0.001	0.03
All	this year	All	small highly significant difference, Sch > Nat	726	58.0	56.3	1.8	0.000	0.15
All	previous years	Asian	minimal highly significant difference, Sch > Nat	2003	57.2	56.5	0.8	0.004	0.06
All	this year	Asian	small significant difference, Sch > Nat	155	57.9	55.8	2.1	0.033	0.17
All	previous years	ESOL	NO SIGNIFICANT DIFFERENCE, Sch > Nat	510	57.0	56.4	0.6	0.244	0.05
All	this year	ESOL	NO SIGNIFICANT DIFFERENCE, Sch > Nat	35	57.0	56.1	0.9	0.700	0.07
All	previous years	Female	minimal highly significant difference, Sch > Nat	5010	57.4	56.5	1.0	0.000	0.08
All	this year	Female	small highly significant difference, Sch > Nat	362	58.4	56.0	2.5	0.000	0.20
All	previous years	Male	NO SIGNIFICANT DIFFERENCE, Sch < Nat	4895	56.4	56.5	-0.2	0.383	-0.01
All	this year	Male	NO SIGNIFICANT DIFFERENCE, Sch > Nat	364	57.6	56.5	1.1	0.077	0.09
All	previous years	Maori	small highly significant difference, Sch > Nat	1286	58.2	56.6	1.6	0.000	0.14
All	this year	Maori	small significant difference, Sch > Nat	88	59.7	56.5	3.2	0.020	0.25
All	previous years	MELAA	NO SIGNIFICANT DIFFERENCE, Sch > Nat	510	57.0	56.4	0.6	0.244	0.05
All	this year	MELAA	NO SIGNIFICANT DIFFERENCE, Sch > Nat	35	57.0	56.1	0.9	0.700	0.07
All	this year	Other	INVALID TEST, N < 20, Sch > Nat	15	61.5	58.5	3.0	0.371	0.24
All	previous years	Other	NO SIGNIFICANT DIFFERENCE, Sch < Nat	188	56.3	56.3	0.0	0.957	0.00
All	previous years	Pacific	minimal highly significant difference, Sch > Nat	1708	57.3	56.4	0.9	0.003	0.07
All	this year	Pacific	small significant difference, Sch > Nat	125	58.4	56.2	2.2	0.042	0.18
All	previous years	Pakeha	minimal highly significant difference, Sch > Nat	6675	57.1	56.5	0.6	0.000	0.05
All	this year	Pakeha	small highly significant difference, Sch > Nat	476	58.4	56.1	2.3	0.000	0.19
Y07	previous years	All	small highly significant difference, Sch > Nat	5363	54.5	53.2	1.3	0.000	0.10
Y07	this year	All	small highly significant difference, Sch > Nat	417	55.6	53.2	2.4	0.000	0.19
Y07	this year	Asian	moderate highly significant difference, Sch > Nat	98	57.1	53.2	3.9	0.001	0.33
Y07	previous years	Asian	small highly significant difference, Sch > Nat	1094	54.9	53.2	1.7	0.000	0.13

Feedback data collected from 23 middle leaders

Dashboards are useful and easy to use!

How confident are you in your ability to use and interpret the DASHBOARD:	Confident or very confident
Independently?	61%
With coaching?	100%

How useful is the data in the DASHBOARD for:	Useful or very useful
understanding student achievement and progress?	91%
evaluating teaching/learning initiatives?	78%
informing future actions?	83%

QUESTION: Does a Wānanga day enhance student outcomes?

HERES WHAT HAPPENED IN 2019:

Wānanga Day and the MCAT
91027 Apply algebraic procedures in solving problems

Wānanga?

In essence students were given a day of supervised preparation for the 2019 MCAT. The MCAT has had a lot of media coverage over its difficultly through students and teachers being underprepared for the external.

Teachers and NCEA students outraged over difficult Level 1 MCAT algebra exam – ***Stuff, 2016***
7 hacks for surviving NZ's toughest NCEA exam – ***Scoop, 2017***

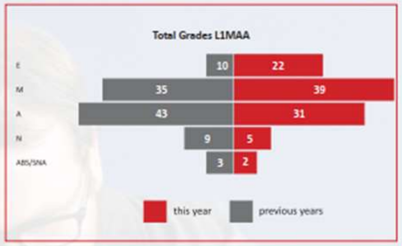
Methodology

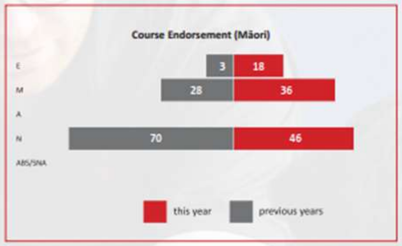
After hearing positive results from the Art and Māori departments in instituting Wānanga days for 2019 the Mathematics department decided for all students sitting the external MCAT to have a day of organised preparation.

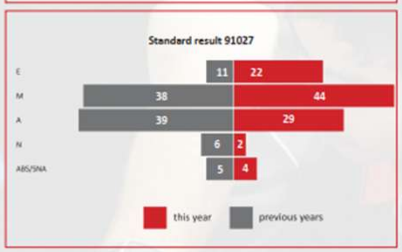
- [redacted] used the library and ran various presentations and activities for students the day prior to the external.
- Students where supplied lunch during a break.

Results Summary

- L1MAA students gained better overall grades in all topics. This difference was highly significant ($p = < 0.0001$).
- Course endorsement improved overall and this improvement was due to Māori students. This was large and significant ($p = 0.002$, $G = 0.475$).
- MCAT showed significant improvement of grades earned including double the percentage of students gaining excellence ($p = 0.016$).







Conclusion

- Using a Wānanga day has enhanced student outcomes.

It is an opportunity to establish a learning focused relationship between students, and teachers. It is a time to fully focus on a standard without external distractions.
Although a significant advantage was established it was only for one standard, the results indicate that this practice should be promoted within the department and school. It would require a major change to the timetable.

GOAL: Was to significantly lift the rate of higher grades and endorsements at Level 1.

HERES WHAT HAPPENED IN 2019:

Highlights

• 90914 ***(IN)*** - Historically 19% Non achieved.
2019 - Large, highly significant improvement.

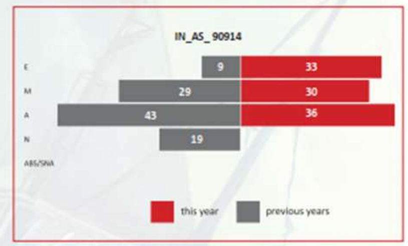
• 90916 ***(EX)*** - Historically 23% - SNA, 9% NA.
2019 - Similar to previous years.

What did we do?

• The Wānanga day was a huge success. 100% of the students achieved their internal allowing us to move as a majority onto the external submissions (closure/peace of mind/admin done).
• Proposition had a clear direction. ***Unpack Arrival***
• Collaboration - Two classes - SM/WO

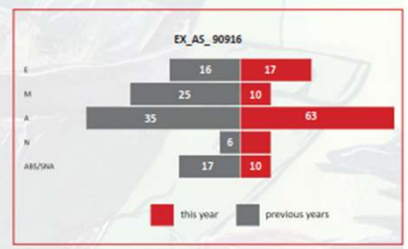
Work on ...

• We achieved our goal from 2019 for the internals by significantly improving both earned and quality of grade. However there was a shift in the externals both up and down.



Externals

SNA - DOWN 7%, N/A - DOWN 6%,
Achieved - UP 28%, Merit - DOWN 15%
Excellence - UP 1%



• As with the internals from 2019, the externals show a ***large portion*** of students gaining achievement. We have not yet fixed this issue!

• Externals are a ***school concern*** and our subject is no different.

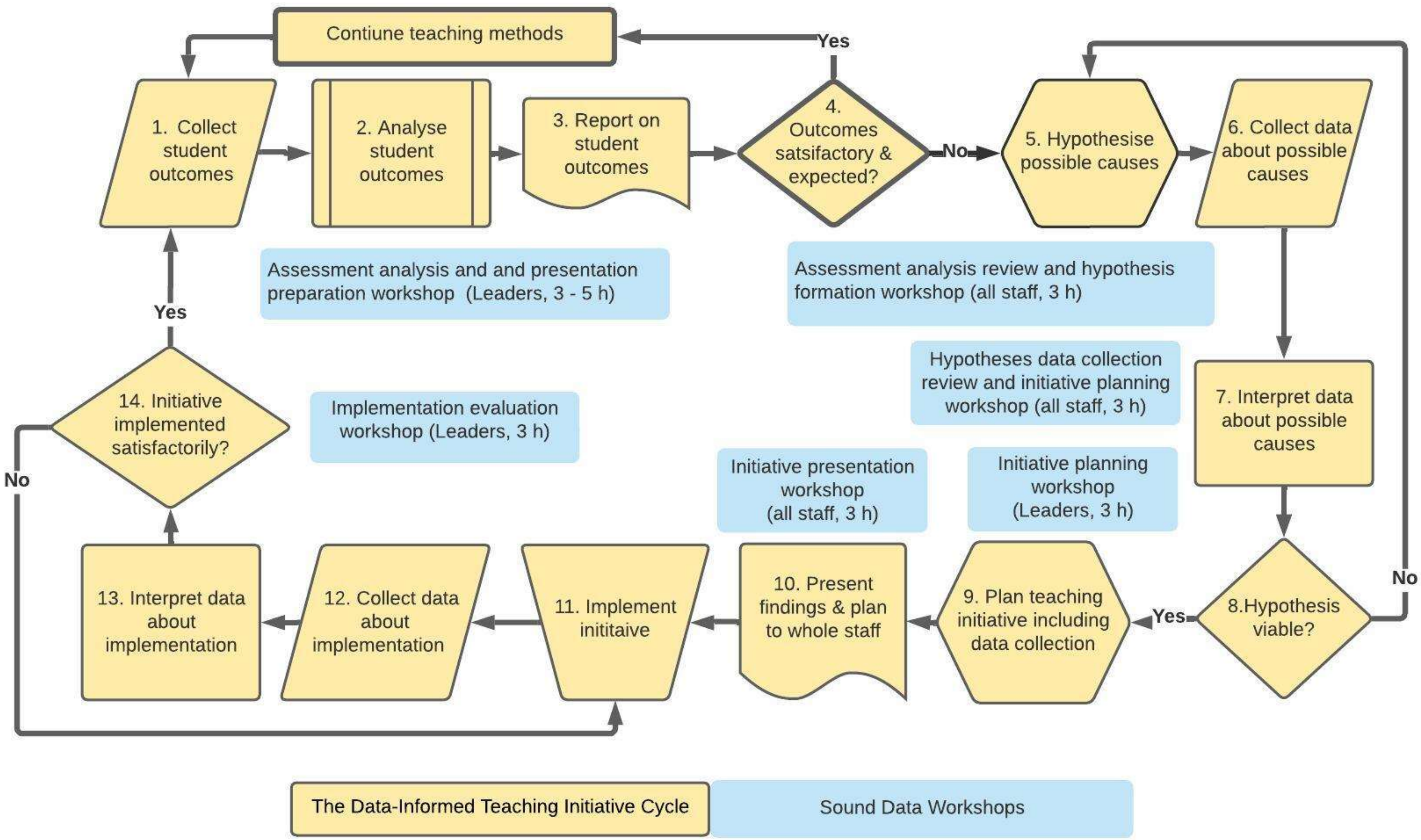
Data presentations are useful and have an impact!

How useful were the poster and presentations for:	Useful or very useful
Clarifying your departmental challenges, goals and plans?	78%
Sharing your departmental challenges, goals and plans?	83%
Understanding the challenges, goals and plans of the other departments?	83%

Survey question	Impact or large impact
Do you think your POSTER PRESENTATION had an impact on your teaching colleagues?	61%

Leaders are confident and value data-informed teaching initiatives

Survey Question	x or very x
How **confident** are you that your evidence-based initiative will lead to improvement?	78%
Were the outcomes of this process **worth** the effort you had to put in?	78%
How **likely** are you to recommend this process to colleagues at other schools?	78%
How **keen** are you to engage in this process again next year?	78%



(Lai & Schildkamp 2013, Mandinach & Gummer, 2015, Schildkamp et al 2018)

Data Driven Decison Making (DDDM)

- DDDM is a cyclic process of purposeful knowledge construction and decision/action (Lai & Schildkamp, 2013).
- *Evaluation* data use: positive or negative implications for user (King & Pechman, 1984);
- How teachers make sense of, respond to, and use *evaluation data* depends upon:
  - the data:
    - high/low stakes (Lewis & Holloway, 2019);
    - the data use technology (Fjørtoft & Lai, 2020);
  - the individual user:
    - beliefs and attiudes (Prenger & Schildkamp, 2018);
    - data literacy (Mandinach et al., 2015);
    - implications of data for teacher identity (Lasater, Bengtson, et al., 2020);
      - attribute the causes of the data:
        - to instruction, student understanding, nature of the test, student charatersitics (Bertrand & Marsh, 2015);
        - to factors outside the school, school policy, students/classroom, teacher functioning (Schildkamp, Poortman, Handelzalts, 2016);
  - social interactions:
    - leaders (Datnow & Park, 2019);
    - the depth, formality and transparency of data use (Vanlommel & Schildkamp, 2019);
    - the school data culture (Lasater, Albiladi, et al., 2020);
    - the external accountability system (Hargreaves et al., 2013).

Using data to improve teaching, learning and accountability

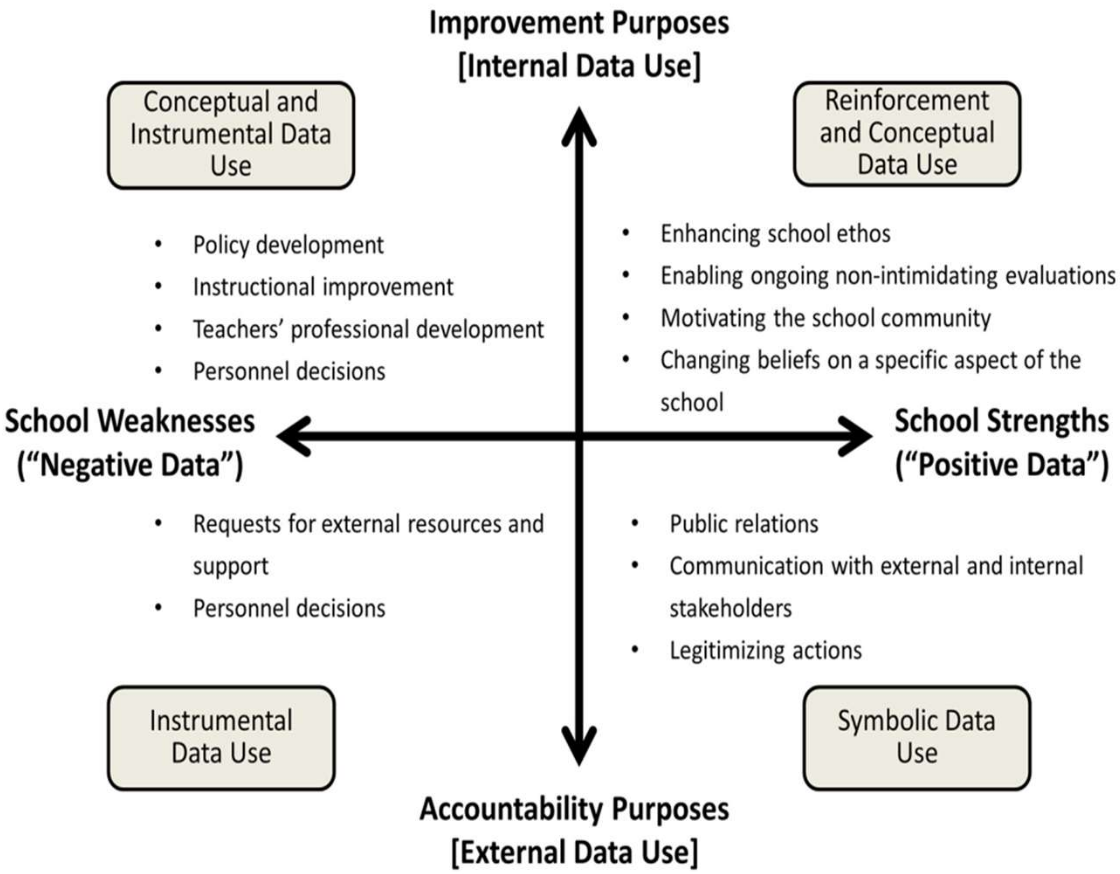
Data-informed teaching initiatives can improve teaching, learning *and* accountability when:

- They use a wide variety of *valid* data (input/process/output);
- They use postive data to celebrate *strengths* and successes as well as negative data to indicate *weakness* for improvement;
- Leaders and teachers are coached and resourced for *collaborative* data use;
- Leaders and teachers having *agency* in their data use in a *safe data-culture*.

(Marsh, Bertrand et al., 2015; Schildkamp & Poortman, 2015; Poortman & Schildkamp, 2016; Aderet-German & Ben-Peretz, 2020; Lai, McNaughton et al 2020, Lasater, Albiladi, et al., 2020)

"Accountability without improvement is empty rhetoric, and improvement without accountability is whimsical action without direction" (Earl and Katz, 2006).

The desirable uses of data in schools



- instrumental use – to directly influence actions and practice;
- conceptual use – to influence mental models and activity over time;
- persuasive or symbolic use – for a predetermined agenda
- reinforcement – to consciously highlight individual and school strengths.

(Aderet-German & Ben-Peretz, 2020)

The undesirable uses of data in schools

- non-use – ignoring or cursory data use i.e., ticking the box;
- interpreting data incorrectly – e.g., insufficient data literacy leading to inappropriate decisions or actions;
- narrowing the curriculum or teaching to the test;
- falsifying data – e.g., giving inappropriate assistance before or during assessments;
- educational triage and reshaping the data pool – focusing only on those on the threshold of passing and transferring less able students to non-examined classes;
- bullying or shaming teachers, leaders and schools

Data is misused when:

- Data use misrepresents reality to stakeholders;
- Data use has negative consequences for students;
- Data use removes the opportunity for the improvement of teaching.

(Booher-Jennings, 2005; Buly & Valencia, 2002; Ehren & Swanborn, 2012; Schildkamp & Ehren, 2013; Ford, 2018; Hardy & Lewis, 2017; Susan, 2016; Datnow & Park, 2018; Volante et al., 2020; Lockton et al., 2020; Bertrand & Marsh, 2021)

References

Aderet-German, T., & Ben-Peretz, M. (2020). Using data on school strengths and weaknesses for school improvement. *Studies in Educational Evaluation*, *64*, 100831. Retrieved 2020/03/01/, from http://www.sciencedirect.com/science/article/pii/S0191491X19300343

Bertrand, M., & Marsh, J. (2021). How data-driven reform can drive deficit thinking. *Phi Delta Kappan*, *102*(8), 35-39. https://doi.org/10.1177/00317217211013936

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231-268. http://www.jstor.org.ezproxy.waikato.ac.nz/stable/3699376

Buly, M. R., & Valencia, S. W. (2002). Below the Bar: Profiles of Students Who Fail State Reading Assessments. *Educational Evaluation and Policy Analysis*, *24*(3), 219-239. https://doi.org/10.3102/01623737024003219

Datnow, A., & Park, V. (2018). Opening or Closing Doors for Students? Equity and Data Use in Schools. *Journal of Educational Change*, *19*(2), 131-152. https://doi.org/10.1007/s10833-018-9323-6

Earl, L. M., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement.* Corwin Press.

Ehren, M. C. M., & Swanborn, M. S. L. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement: Special Issue: Data-Based Decision Making around the World: From Policy to Practice to Results*, *23*(2), 257-280. https://doi.org/10.1080/09243453.2011.652127

Ford, T. G. (2018). Pointing teachers in the wrong direction: understanding Louisiana elementary teachers' use of Compass high-stakes teacher evaluation data. *Educational Assessment, Evaluation and Accountability*, *30*(3), 251-283. https://doi.org/10.1007/s11092-018-9280-x

Hardy, I., & Lewis, S. (2017). The 'doublethink' of data: educational performativity and the field of schooling practices. *British Journal of Sociology of Education*, *38*(5), 671-685. https://doi.org/10.1080/01425692.2016.1150155

Lai, M. K., McNaughton, S., Jesson, R., & Wilson, A. (2020). *Research-practice Partnerships for School Improvement: The Learning Schools Model*. Emerald Group Publishing.

Lai, M. K., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. *Handbook of human and social conditions in assessment*, 77-94.

Lasater, K., Albiladi, W. S., Davis, W. S., & Bengtson, E. (2020). The Data Culture Continuum: An Examination of School Data Cultures. *Educational Administration Quarterly*, *56*(4), 533-569. https://doi.org/10.1177/0013161X19873034

Lockton, M., Weddle, H., & Datnow, A. (2020). When data don't drive: teacher agency in data use efforts in low-performing schools. *School Effectiveness and School Improvement*, *31*(2), 1-23. https://doi.org/10.1080/09243453.2019.1647442

Mandinach, E. B., & Gummer, E. S. (2015). The Conceptual Framework for Data Literacy for Teachers. In M. C. Linn (Ed.), *Data Literacy for Educators : Making It Count in Teacher Preparation and Practice* (pp. 38-56). Teachers College Press. http://ebookcentral.proquest.com/lib/waikato/detail.action?docID=4513520

Marsh, J. A., Bertrand, M., & Huguet, A. (2015). Using data to alter instructional practice: The mediating role of coaches and professional learning communities. *Teachers College Record*, *117*(4), 1-40.

Poortman, C. L., & Schildkamp, K. (2016). Solving student achievement problems with a data use intervention for teachers. *Teaching and Teacher Education*, *60*, 425-433. https://doi.org/10.1016/j.tate.2016.06.010

Schildkamp, K., & Ehren, M. (2013). From "Intuition"- to "Data"-based Decision Making in Dutch Secondary Schools? In (2013 ed., Vol. 17, pp. 49-67). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_4

Schildkamp, K., Handelzalts, A., Poortman, C. L., Leusink, H., Meerdink, M., Smit, M., Ebbeler, J., & Hubers, M. D. (2018). *The Data Team™ Procedure: A Systematic Approach to School Improvement*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58853-7

Schildkamp, K., & Lai, M. K. (2013). Conclusions and a Data Use Framework. In Data-based Decision Making in Education: Challenges and Opportunities (pp. 177-191). Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_10

Schildkamp, K., & Poortman, C. L. (2015). Factors Influencing the Functioning of Data Teams. *Teachers College record (1970)*, *117*(4).

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2016). Data teams for school improvement. *School Effectiveness and School Improvement*, *27*(2), 228-254. https://doi.org/10.1080/09243453.2015.1056192

Tabachnick, B. G., & Fidell, L. S. (2014). Using multivariate statistics (Vol. 6). Pearson.

Volante, L., DeLuca, C., Adie, L., Baker, E., Harju-Luukkainen, H., Heritage, M., Schneider, C., Stobart, G., Tan, K., & Wyatt-Smith, C. (2020). Synergy and Tension between Large-Scale and Classroom Assessment: International Trends. *Educational measurement, issues and practice*.