Dear Mr Marwood,

Thank you for your query regarding the statistical significance used in RAISEonline.

It may be of interest that Ofsted and the Department for Education are currently engaged in a review of the statistical significance used in RAISEonline and are working with stakeholders from schools and local authorities. We would welcome feedback on the methodology to inform ongoing discussions.

Use of Z-score (pages 14 and 15 of the guidance)

The context section of RAISEonline contains a table displaying the prior attainment of each year group, as measured by the Average Points Score (APS) at the end of the previous key stage, and compares this to the national picture.

A standardised normal score (or Z-test) is used to assess whether the difference between the cohort’s prior APS is different from the national average. As you can see from the notation, this calculation uses the national population standard deviation.

It is appropriate to apply a Z-test to a sample mean when the standard deviation is known, regardless of sample size (see figure 1). As we are comparing a school’s prior APS to the national APS for that year group, and therefore the population standard deviation is known, we should apply a standardised normal score.

(Figure 1: a flow chart describing the situation in which to apply a Z or t-test.)

Use of t-test (pages 25 to 26 of the guidance)

APS reports in RAISEonline contain significance tests to compare school level and pupil group means to the national or school average respectively.

To test for statistical significance in APS reports, a confidence interval is generated from the mean and sample standard deviation for the target group and critical values of a student’s t-distribution. It is appropriate to apply a Z-test to samples based on more than 30 observations and use a t-test for smaller samples.

RAISEonline performs significance tests on both the school level figures and pupil group level figures too. It is a fairly safe assumption that almost all school level APS figures will be based on at least 30 observations, especially at Key Stage 4, but this assumption does not hold when we consider pupil groups.

Based on these assumptions, it is appropriate to apply a Z-test to school level APS values (as they are likely to be based on more than 30 observations) but use a t-test for the pupil group values (often based on fewer than 30 observations). This is not a practical solution and would add a layer of complexity that can be avoided by applying a single methodology.

Statisticians within Ofsted and the Department for Education decided to accept simplifying assumptions and apply a t-test to APS significance tests. This has three major benefits:

1. Applying the same formula throughout APS significance tests makes the methodology more accessible to users and avoids concerns about applying different statistical tests to the same metric within the same school.

2. Using a t-test is more robust than a Z-score for small cohorts (as it is less sensitive to outliers), but converges on the results of a Z-test as the sample size increases; applying a t-test to all cohorts provides robust outcomes whilst not being overly-sensitive to larger sample sizes.

3. The t-test is a more conservative test than using a Z-test and as such, it reduces the probability of making a type I error. A type I error occurs when the difference is said to be significant when it is not significant. This in turn increases the probability of making a type II error, i.e. stating that the difference is not significant, when perhaps it is significant. The impact of a type I error is judged to be greater than the impact of a type II error in this situation.

I hope that this helps to explain why different tests have been applied.

Best wishes,

RAISEonline Team

From: Jack Marwood

Sent: 14 July 2014 17:44

To: Enquiries

Subject: Re: FAO Raise team

Dear Ms \*\*\*\*\*,

Thank you for your response to my enquiry.

I am concerned about the use of ‘significance’ within RAISEonline. it appears that statistical tests of significance have been used inconsistently within RAISEonline reports.

All page numbers below refer to the document called ‘A Guide to calculations of RAISEonline reports 2013’ in RAISEonline’s Methodology Library (https://www.raiseonline.org/documentlibrary/ViewDocumentLibrary.aspx).

On page 14 of the guide, a formula is given for calculating significance. This suggests that a Z-test is being used to test for significance, as a Z-score is tested against a figure of 1.96.

On page 25-26, a t-test is used to test for significance.

Could you explain why the two different tests have been used?

Yours sincerely,

Jack Marwood