Fact Sheet 16

Raw scores, centiles and standardisation

The following information is split into two main sections, namely A and B. Section A addresses three questions that some users may ask and Section B explains some basics about the standardisation process, raw scores and standard scores (such as centiles). We recommend that you read all parts of this document.

Section A:

- **Why can the same raw score give a different centile score?**
- **How can child C get fewer items correct than child D but still obtain a higher centile score?**
- **Why can small changes in raw scores result in large changes in centile scores?**

Section B contains:

- **Standardised or normative tests**
- **Why use Standard Scores such as centiles anyway?**

Section A

1) Why can the same raw score give a different centile score?

The straightforward answer is that the two raw scores use different look-up tables to map to the centile score (see Section B). This is because the children belong to different age bands and therefore different norm groups based on those age bands are used when calculating centile scores.

For example, both child A and child B obtain a raw score of 20 yet child A's centile score is 44 and child B's centile score is 32. Child A is 6 years old and child B is 7 years old.

A scoring outcome of this nature is normal, proper and to be expected for any good normative or standardised test (i.e. statistically standardised) that cover various ages. Each age band will have different "norms" (i.e. an appropriate population group or reference population) and child A's score is mapped (from the raw score to the centile score) using a different look-up table.

In accordance with expectations, population statistics would show that on average Child B's peers obtain higher raw scores than child A's peers. Therefore child B's raw score of 20 is a relatively worse performance as compared to his/her peers than that of child A. A centile of 44 for child A means that 44 percent of child A's age related peers obtain a raw score of less than 20. Whereas only 32 percent of child B's age related peers obtain raw scores less than 20.

2) How can child C get fewer items correct than child D but still obtain a higher centile score?

Following on from the information provided in 1) above, take another example where child C gets 10 items correct (a raw score of 10) and child D gets 15 items correct (raw score of 15). Yet child

C obtains a centile score of 65 and child D obtains a centile score of 40. This can be a normal and proper outcome for any normative (standardised or statistically standardised) test. It will occur when different look up tables (norms tables) are used by the software for different population groups.

From the scores given in the example you would expect that child C is younger than child D. Therefore it is expected that child C's age related peers would, on average, obtain lower scores that child D's age peers. In fact a centile of 65 for child C means that 65 percent of child C's age related peers (the population reference group) obtain a raw score lower than 10. In contrast 40 percent of child D's age related peers obtain a raw score lower than 15.

3) Why can small changes in raw scores result in large changes in centile scores?

Firstly, we recommend that you read the answers to the two previous questions then read the following answer.

Small changes in raw scores can give rise to 'large' changes in centile scores. The norms look up table determines these centile scores and their incremental values. The norms tables were compiled as a direct result from real data collected in the standardisation process (statistical standardisation that is – see note on standardisation below). Where a small change in a raw score results in a large change in a centile score it indicates that the population scores cluster around those particular raw scores and are not spread out relatively thinly. This would be a typical outcome where an assessment has a limited range of actual raw scores (imagine the centile scores for a test that had only 4 test items – you have to spread the whole population of scores over 5 possible outcomes [scores of 0, 1, 2, 3 and 4] assuming that the actual range of scores matches the possible range of scores) or where there is a statistical "skew" in the distribution of population raw scores. Even if there is a wide range of possible raw scores which could be obtained, the real results from a given population of individuals may not always closely follow the "normal distribution" (i.e. bell-shaped distribution curve). Despite a "statistical skew", the test may still be worthwhile and advantageous to use because the results obtained from individuals may still have of value in diagnostic and assessment terms.

These effects may be observed to some extent in the "Rhymes" test in the Lucid CoPS program. Very few people would question the usefulness and appropriateness of using a rhyming (and alliteration) test to assess the phonological awareness of young children between the ages of 4 to 8 years. (Indeed, out of the 27 computerised tests originally used in the prospective longitudinal validation study for CoPS, the Rhymes test proved to be the best single predictor test for this skill area.) Yet the distribution of performances from the population on such a test, especially for the older children in the age range, has a tendency to bunch at the top end of scores (the distribution is said to have a "negative skew"). In an attempt to reduce the 'skew' of frequency of scores one thing you could do is to increase the number of test items. However, this will not necessarily improve (reduce) the 'skewedness' for a given individual – making 35 correct responses for example may not be much harder (in terms of cognitive difficulty) than say, making 12 correct responses. Furthermore, adding more test items would extend the time that the assessment takes and many teachers are highly resistant to increasing the assessment duration, especially if it adds little or no extra diagnostic value. What in fact tends to happen with this age group on a task like rhyming is that the children have a tendency to be able to respond correctly (with little or no error) or alternatively, they have considerable difficulty (where several errors are made). There isn't that fine grading in between. This is of course is a simplification and there is in fact some spread of data, but nevertheless there is a bunching of scores separating those that generally have realised the

skill from those that generally haven't fully realised the skill. In a normative test every raw score must be mapped to a centile score (or some other "standard score" such as a standard deviation). Therefore the centile scores that correspond to adjacent raw scores may have rather large jumps. This is merely a function of the distribution of the performance of the population on this particular test.

To some extent this type of effect can also be seen in the Wock test of Lucid CoPS. Again increasing the number of test items in an attempt to reduce the 'skew' of the distribution of scores would result in the test becoming overlong (increasing testing time and without significant gain in the diagnostic value of the test) and also may have little impact on improving the distribution of performances.

Section B

1) Standardised or normative tests

Many people misunderstand the true meaning of what is generally referred to as a 'standardised test'. It does not simply mean that the test is delivered in a consistent way, nor does it simply mean that the test uses the exactly the same measures. Most importantly, it means that the test is 'statistically standardised' in that it uses norms to map raw scores to population statistics or "standard scores". In order to reduce misunderstandings it is sometimes better to refer to a 'normative' test rather than to refer to a 'standardised' test.]

The standardisation process to create a "standardised test" [1] is where a large number of individuals are selected in an unbiased manner, and performance data is collected from these unbiased and representative individuals (representative in terms of accurately reflecting the whole "population"[2]). These data are used to create the norms[3] and the ideal situation would be if the norms were in fact the whole population scores, but it is generally not possible to obtain scores from the whole population. So a sample[4] is used to create the norms which accurately represents and reflects the whole population. The importance of the quality of the sample, and specifically the importance of the quality and appropriateness of the sampling procedures, cannot be overemphasised. The manner in which the sample is assembled is crucial and it should pass strict procedural criteria for subject selection. Norms should not be created if the sample is biased is any way, otherwise the norms will be equally biased. (For example, you cannot create a standardised test to identify dyslexics from non-dyslexics if the data that forms the norms was taken from individuals whom were pre-selected because they might be dyslexic. i.e. you cannot create norms for a standardised test by obtaining data from only dyslexics, or by obtaining data from a pre-selected group of

[1] A standardised test (or a norm-referenced test) means that an individual's score is a measure of how well he or she did in comparison with a large group (generally in education this means a large group that represents the general population for given ages). [Standardised tests are in contrast to "criterion-related" tests where a specific skill is examined and achievement is measured without comparison to the scores of other individuals.]

[2] "Population" refers to its statistical meaning of a group of persons (or other subjects of study) that one wishes to describe or about which one wishes to generalise.

[3] The "norms" are standards of performance that represent the population.

[4] "Sample" refers to its statistical meaning of a group of subjects selected from a larger group in the hope that studying this smaller group (the sample) will reveal important things about the larger group (the population).

dyslexics and then comparing them with another pre-selected group of any kind.) The sample must not have biases that will make it unrepresentative of the population.

Generally, the principle of random selection ensues that there are no systematic biases (so long as the sample is large enough), and given a large enough sample size any random anomalies will be diluted to such an extent that they do not distort the group statistic (because, in theory, the positive distortions will be counteracted by the negative distortions). In addition to procedural criteria, the sample must also pass statistical criteria to say with "confidence" (statistically quantified confidence) that the sample is representative of the population.

2) Why use Standard Scores such as centiles anyway?

'Raw scores'[5] are mapped to 'standard scores'[6], such as centiles[7], using look-up tables (e.g. norms tables). The norms tables are compiled by collecting a large data set of actual test scores obtained from a large 'sample'[4] (relevant individuals – e.g. pupils of certain ages) that has been statistically verified as being unbiased and representative of the desired 'population'[2] (e.g. all UK children of a defined age). It is important that the correct sampling protocols are used in the standardisation process to ensure good quality norms (see above).

The norms table will map every single possible raw score to a standard score, for example to a centile. This mapping can be a one-to-one (every raw score has a unique standard score) or a many-to-one (more than one raw score can have the same standard score) relationship. It is usual to have different norms tables for different groups or "populations" (the different groups are usually based on age but this is not necessarily so).

Viewing test results in standard scores such as centiles means that it becomes immediately possible to make meaningful comparisons across tests and people, either within an individual's performance or across different tests and between different individuals. In contrast to this there are considerable limitations and difficulties when comparing 'raw scores' (or criterion-related test scores) between different tests and between different individuals. For example, compare a score of 8 out of 10 on a reading test (reading raw score of 8) with a score of 8 out of 12 on a maths test (maths raw score of 8). Now say that the raw score of 8 on the reading test was given to child A from a Hull school and that the raw score of 8 on the maths test was obtained by another child at a Guildford school. Now say the children are of different ages, and so on. You can see that the number and significance of the assumptions that must be made, if you were ever to compare raw scores in this manner will be highly subjective and prone to error. However, if you had all the above two scores as standard scores instead of raw scores then easy and meaningfully be comparisons can immediately be made without attempting to make adjustments for age differences, task differences, regional differences and so on. Standard scores such as a centile, show the individual's performance on a given test in relation to the relevant statistical population.

---

[5] Raw scores , are scores, data, or numbers that are in their original state and have not been statistically manipulated.

[6] A 'Standard Score' is a score of relative standing in a group arrived at by transforming raw scores in a way that allows you to meaningfully compare scores obtained from different distributions. Examples of standard scores include Centile, Z-scores, Stanine etc

[7] Centiles (also known as percentile) is a number indicating rank by stating what percentage of those being measured fall below a particular score. For example, the 20th centile means that this score cuts off the bottom 20% of the distribution. In other words this score exceeds that of 20% of the population.

A centile score means that a certain percent of the population for that reference group (usually age) achieves a raw score lower than the one obtained by that individual (see footnote 7 for a definition). **Important note: A centile (or percentile) score does not mean that a certain percent of items taken were scored as correct.**

Helpful sections in the Lucid Teacher's Manuals that accompany our software include:

Lucid CoPS Teacher's Manual 2nd Edition

- Section 4.1.2.1 (pg 61)
- Section 4.2 (pg 64)

LASS Junior Teacher's Manual 1st Edition

- Section 4.1.1 (pg 33)
- Section 4 in general (pg 33)

LASS Secondary Teacher's Manual

- Section 4.1.1 (pg 34)
- Section 4 in general (pg 34)

For more information about Lucid or the developments or research please visit the Lucid web site www.lucid-research.com. The Lucid staff can be contacted by email info@lucid-research.com, telephone +44 (0)1482 862121 or fax +44 (0)1482 882911.

Please note that the information contained in this document is correct at time of going to press.