CAMBRIDGE ASSESSMENT

# *Calculating the number of marks needed in a subtest to make reporting subscores worthwhile*

Tom Benton

Cambridge Assessment Research Report

27th November 2012

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

**Author contact details:**
Tom Benton
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
Benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

Cambridge Assessment is the brand name used by the University of Cambridge Local Examinations Syndicate (UCLES). UCLES is a non-teaching department of the University of Cambridge and has various subsidiaries. The University of Cambridge is an exempt charity.

**How to cite this publication:**

Benton, T. (2012). *Calculating the number of marks needed in a subtest to make reporting subscores worthwhile*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

**Key Finding**

Assuming that a test comprises of 180 marks there is little point in presenting subscores based on less than 30 marks within this test. This should be seen as a minimum value and in practice 50 or even 60 marks may be required to be confident that subscore results will provide additional useful information.

**Background**

Any individual GCSE subject covers a wide variety of skills. These skills are described within the subject specification and may be in terms of different topics within the subject or different generic skills as described within the assessment objectives. At face value, providing more detailed information to candidates about their scores on examination questions relating to each possible skill could help them identify areas where they have secure knowledge and areas where they need to improve. However there are two dangers in providing such information:

- Subscores based on a subsection of items in a test will be inherently less reliable than the overall test score. As with any examination, scores only provide an estimate of candidates' skills and these estimates will be tend to be less accurate if they are based on smaller numbers of questions.
- There is often a very strong correlation between candidates' skills within one topic or assessment objective and their skills within another. For example, in a combined science qualification, those candidates who are good at Biology also tend to be good at the Physics. Obviously there will be individual candidates who are exceptions, but in general a candidate's overall score will be highly informative about their ability across a number of different areas.

Together the above points raise the possibility that, under certain circumstances, subtest scores could in fact be less informative about candidates' abilities within particular topics or assessment objectives than the overall test score. For example, suppose a candidate performs well on an exam overall but badly in a particular section, it is possible that their poor performance was caused by something very particular about the individual questions they answered within that section (or, for example, a temporary loss of concentration) rather than by a real weakness in the relevant area. At worst this could mean that providing additional feedback on subtest scores could provide misleading information to candidates and lead to them wasting time trying to improve their skills within a topic or assessment objective where they are already strong.

The aim of this paper is to identify the number of marks needed to asses a sub-skill to ensure that achievement within the relevant subsection of a test will be more informative of students skills in this area than overall test score. In undertaking this research it is understood that is possible to combine information from overall test performance and performance within subsections to estimate each student's level of ability in any particular sub-skill. However, even if this were used in practice, we would hope that such estimates would be more heavily determined by achievement within the relevant subsections of the test than by overall test scores; otherwise we run the risk of largely repeating students' overall test scores to them under a variety of different titles. For this reason, the rationale behind this paper provides a valid approach to answering the question regarding how many marks should be within a subtest in order for reporting of scores on the relevant sub-skill to be worthwhile.

Previous studies of this issue (for example, Sinharay, 2010) have demonstrated that in practice it can be quite difficult for subscores to provide worthwhile additional information over and above the total test score. Indeed empirical analysis of 94 subtests across 25 different tests[1] found that only 16 of these provided added value over and above the total test score. All of the subtest scores that were

[1] Subtests ranged between 11 and 69 marks in length with an average of 29 marks.

found to add value occurred either for subtests with large numbers of items or for tests where the separate subscores measured relatively distinct dimensions of ability[2].

### Mathematical model

The approach developed in this paper builds upon classical test theory. A similar mathematical model to the one developed below has been separately described by Sinharay (2010) and by Sinharay, Haberman and Puhan (2007). Indeed the equations developed within these papers are almost identical to those derived below with just minor differences in the starting assumptions about the composition of the overall test[3] and in the notation.

In order to answer our research question we employ the following mathematical model. We first assume that each student has a general level of ability in the subject that is being assessed as a whole which we will label θ. We will also assume that the overall test comprises *k* subtests measuring sub-skills and that a student's true ability in sub-skill *i* is labelled $\varphi_i$. We assume that for all *i* the correlation between θ and $\varphi_i$ is equal to the square root of ρ. Finally we assume that each sub-skill is measured by a subsection of the test, that score on this subtest is labelled $X_i$ and that the reliability of each subtest is equal to α. With little loss of generality we can assume that each of θ, $\varphi_i$ and $X_i$ are standardised to have a variance equal to 1[4].

From the above and from the definition of reliability for each subtest we know that

$$Cor(\varphi_i, X_i) = \sqrt{\alpha}$$

We can also easily calculate that if i≠ j then

$$Cor(\varphi_i, X_j) = \rho\sqrt{\alpha}$$

And[5]

$$Cor(X_i, X_j) = \rho\alpha$$

We now wish to compare $Cor(\varphi_i, X_i)$ to the correlation of any given true ability level in a sub-skill and total test score. This can be calculated by

$$Cor\left(\varphi_i, \sum_{j=1}^{k} X_j\right) = \frac{Cov(\varphi_i, \sum_{j=1}^{k} X_j)}{\sqrt{V(\sum_{j=1}^{k} X_j)}} = \frac{\sqrt{\alpha} + (k-1)\rho\sqrt{\alpha}}{\sqrt{k + k(k-1)\rho\alpha}} = \frac{\sqrt{\alpha}(1 + (k-1)\rho)}{\sqrt{k(1 + (k-1)\rho\alpha)}} \tag{1}$$

We can now infer that the score on any given subtest will be a more reliable indicator of skills in that subtest than total test score if

$$\sqrt{\alpha} > \frac{\sqrt{\alpha}(1 + (k-1)\rho)}{\sqrt{k(1 + (k-1)\rho\alpha)}} \tag{2}$$

[2] Even in these instances no instances were found of subtests with less than 20 marks adding value.

[3] In our paper we assume that the total test score comprises a number of subtests each of which have equal reliability. We also assume equal correlations between the true scores on different subtests. To avoid this the earlier papers start from a position where the correlation between overall true score and true subscores is known (or can be estimated) and derive rules about the length of subtests of this basis.

[4] In essence we are assuming that the standard deviations of scores on different subtests will be equal.

[5] Assuming linear relationships between sub-skills and overall skill, and sub-skills and subscores.

We can immediately see that this equation will always hold if ρ=0. Assuming that both α and ρ are greater than 0 the equation will hold if

$$\alpha > \frac{((k-1)\rho^2 + 2\rho - 1)}{k\rho} \tag{3}$$

We can see that if ρ=1 (that is, the test is unidimensional meaning that all sub-skills are in fact the same) then total test score will always be at least as good an indicator of ability in the sub-skill as the associated subscore with equality if and only if α=1. Also note that since ρ≤1, the right hand side of the equation is always less than or equal to ρ so the subscore will always be at least as good an indicator of ability in the sub-skill as the total score if α≥ρ.

Finally by setting the numerator of the right hand side of the equation to zero and solving for ρ, we can deduce that scores on any given subtest will always be a more reliable indicator of subtest abilities than total test score if

$$\rho < \frac{\sqrt{k} - 1}{k - 1} \tag{4}$$

This means that if we are interested in measuring just two sub-skills, the relevant subscores will always be a more accurate indicator than total test score if $\rho < \sqrt{2} - 1$=0.414. As the number of sub-skills increases the reliability of the total test score will increase relative to the reliability of individual subscores (provided ρ>0) and so smaller values of ρ will be required to allow the possibility of total test score being a more accurate indicator of sub-skills than the relevant subscore.

The formula given in equation 3 (determining the minimum required level for α in order for subscores to be worthwhile indicators of sub-skills[6]) can be easily adapted to work out the number of marks required using the Spearman-Brown formula. In order to do this we simply use a known estimate of reliability for a given number of marks (for example, 60 marks yielding a reliability of 0.9[7]) and use the Spearman-Brown formula to determine how many marks we need to achieve the required level of reliability.

In fact we can go further than this. If we assume that the overall number of marks available in a test is fixed (for example at 180) then the number of sub-skills we can measure (*k*) is determined by the total number of marks overall divided by the number of marks associated with each sub-skill. Now the Spearman-Brown formula allows us to plot the association between the number of marks for each subscore and their individual reliability. Using equation 1 we can now also calculate the association between the number of marks in each subscore and the correlation between total test score and each sub-skill. This allows us to identify the number of marks required in order that subscores will be a more reliable estimate of sub-skills than the overall score.

Two worked examples are shown below.

**Example 1 – Providing separate scores for biology, chemistry and physics**

To begin with we estimate values of inter-sub-skill correlation (ρ), and subscore reliability (α) using data from the June 2012 higher tier GCSE units in Biology, Chemistry and Physics (B632, B642 and B652). Each of these units comprised a 60 mark test. Only candidates completing all three papers during June 2012 were included in analysis. This provided 17,786 students for analysis. For each individual GSCE unit, reliability (denoted as α within our work) was calculated using Guttman's

[6] Given the level of correlation between overall true ability and true ability in a sub-skill, and the number of sub-skills beign assessed

[7] Approximate average result from reliabilities for June 2012 GCSE units in Biology, Chemistry and physics (B632, B642 and B652 respectively).

Lambda 4 (Benton, 2013). Correlations were also calculated between test scores on separate GCSE units. These correlations were then corrected for measurement error on each of the units to provide an estimate of the correlation between pupils' true abilities within each subject[8] (that is ρ). The results are shown in table 1.

Table 1: Reliabilities, correlations and corrected correlations for GCSE units in biology, chemistry and physics taken in June 2012 (B632, B642 and B652).

| Unit 1 | Unit 2 | Reliability of unit 1 | Reliability of unit 2 | Correlation between raw scores on unit 1 and unit 2 | Implied correlation between true abilities in sub-skills |
|---|---|---|---|---|---|
| Biology | Chemistry | 0.89 | 0.93 | 0.82 | 0.90 |
| Biology | Physics | 0.89 | 0.92 | 0.80 | 0.88 |
| Chemistry | Physics | 0.93 | 0.92 | 0.83 | 0.90 |

The results in table 1 indicate that for a 60 mark test the reliability is roughly equivalent to 0.9. Similarly the results above indicate that the correlations between pupils' true abilities in the different subjects are close to 0.9. Assuming that our overall test will consist of three subscores (that is, $k$=3), using these values as a basis for the Spearman-Brown formula and for equation 1 we can calculate the reliability of subscores as well as the correlation between total test score and true ability in any given sub-skill (that is, either biology, chemistry or physics). The results of these calculations are shown in figure 1.

As shown by figure 1 the reliability of any subtest increases as the length of the test increases according to the Spearman-Brown formula[9]. Likewise as the length of the test as a whole increases (remembering it is three times the length of the subscore) the correlation between total test score and true ability in the sub-skill also increases. When the number of marks in the subtest is small the total test score is more indicative of the sub-skill than the subscore, however, as the length of the subtests increase, subscores become the more reliable indicator. The crossover occurs at 60 marks, indicating that subscores based on less than this number of marks would be less indicative of skill within the individual sciences than the total score across all three sciences[10]. Thus we can recommend that if separate scores for biology, chemistry and physics are required, each should be assessed by a test of at least 60 items[11].

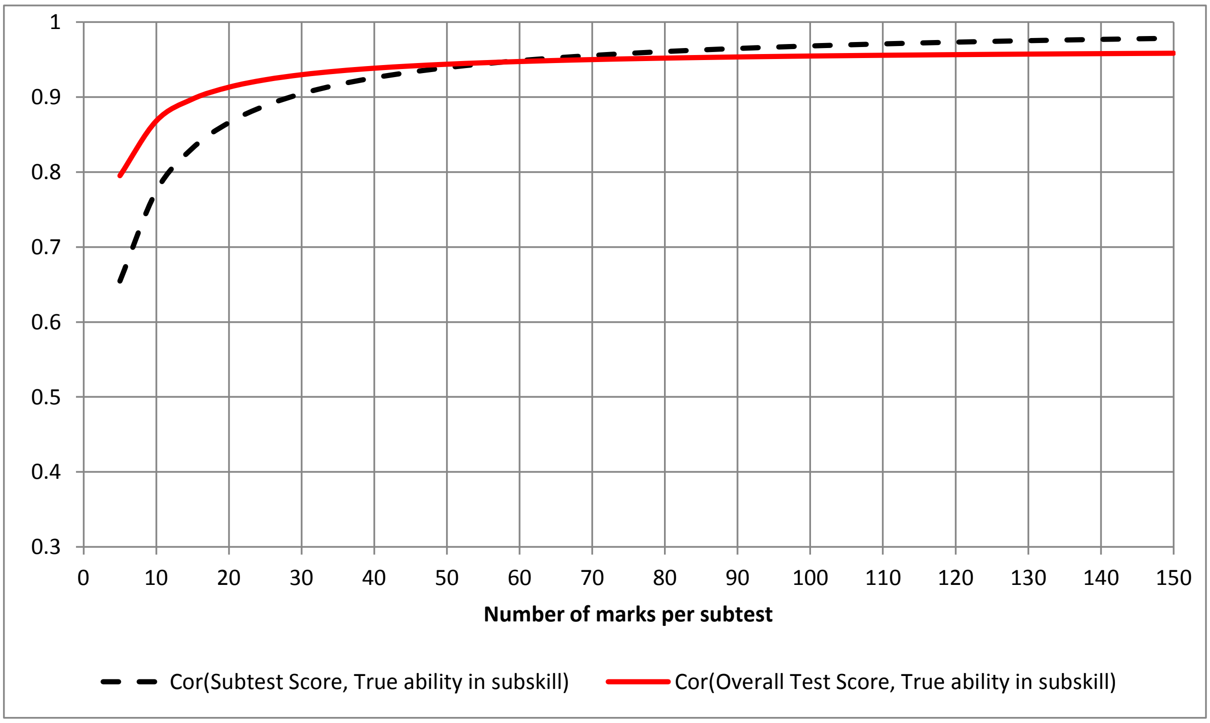
[8] This is done by dividing the correlation in raw scores by the square root of the product of the reliabilities.

[9] Or equally, and as actually displayed in the graph, the correlation between true ability in the relevant sub-skill and observed subscore increases.

[10] This fits with the earlier note that subscores will be more reliable than total test scores whenever $\alpha>\rho$.

[11] Although it is (obviously) true that if even more items were used then the subscores would become even more reliable.

Figure 1: Association of between subtest length and the strength of the correlation between true ability in associated sub-skill, subscore and total score for GCSE science



**Example 2 – Dividing a 180 mark test into subscores**

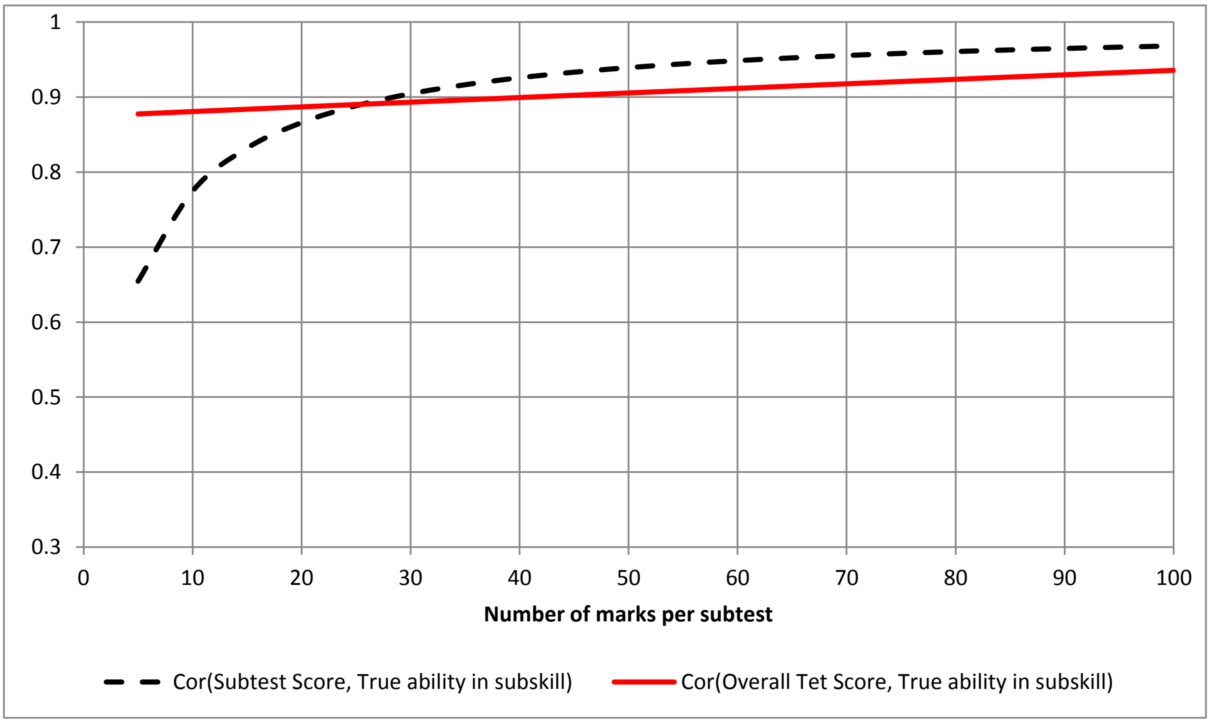
The following example assumes that our overall test length is fixed at 180 marks. We wish to ascertain how many distinct sub-skills we can both assess and reliably report on using such a test and (equivalently) how many marks are required for each sub-skill. For the purposes of this analysis we again assume that (as with each of Biology, Chemistry and Physics), were we allowed 60 marks to assess any sub-skill, we could achieve a subscore reliability of 0.9.

For the purposes of this analysis we attempt to identify a best case scenario. That is, one where we are likely to underestimate the number of items required for each sub-skill in order for subscores to be worth reporting. This means that the number of marks derived from this analysis should be a bare minimum for the number of marks that will be required. In order to provide the most optimistic scenario for the number of marks that are likely to be needed per subscore we need to identify the lowest possible plausible value for the correlation between true abilities in different sub-skills. This value is chosen as 0.8. The reason for choosing this value is that the correlation between GCSE grades in any given single subject and average grade across all of the other subjects that a student is taking is approximately 0.75[12] (Benton and Sutch, 2013). However we are interested in the correlation between performance within the same GCSE subject rather than across different subjects and so would expect the correlation to be higher than this. Furthermore the correlation of 0.75 quoted here does not account for measurement error in GCSE grades. The correlation between students' *true* abilities in different subject ought to be a little higher. For these reasons 0.8 would appear to be a reasonable lower bound for the expected correlation between the true values of pupils' abilities across different sub-skills within the same GCSE subject.

Using these values and assuming that the number of subtests (*k*) within equation 1 is determined by 180 divided by the number of marks per subtest we can calculate the relative worth of subscores and total scores for evaluating any given sub-skill. The results are shown in figure 2.

[12] Based on an analysis on NPD achievement data from summer 2011.

Figure 2: Association of between subtest length and the strength of the correlation between true ability in associated sub-skill, subscore and total score for a notional 180 mark test



As with figure 1, figure 2 shows that as the number of marks in the subtest increases so does the reliability of any given subscore[13]. However, in contrast to figure 1, since the overall test length is held constant, the correlation between total test score and true ability in the sub-skill only increases slightly as the length of the subtest increases[14]. The lines cross at approximately 30 marks. That is, with less than 30 marks per subtest, the total score on the test would be a better indicator of any particular sub-skill than the relevant subscore. However, if a subtest consists of more than 30 marks, the subscore should be the better indicator.

It should be reiterated that this analysis is based upon a fairly generous scenario where different sub-skills are relatively independent of one another. They indicate that, at best, a 180 mark test could usefully provide information regarding 6 distinct sub-skills with 30 marks per subtest.

## References

Benton, T. (2013) *An empirical assessment of Guttman's Lambda 4 reliability coefficient*. Paper presented at the 78th Annual Meeting of the Psychometric Society, July 2013. Available from http://www.cambridgeassessment.org.uk/Images/141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf.

Benton, T., and Sutch, T. (2013) *Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2*. Cambridge Assessment Internal Report.

Sinharay, S. (2010) How often do subscores have added value? Results from Operational and Simulated Data. *Journal of Educational Measurement*, 47, 150-174.

Sinharay, S., Habermand, S., and Puhan, G. (2007) Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26, 21-28.

[13] Indeed the values are identical to figure 1.

[14] With the slight increase being attributable to a greater percentage of the test score being concerned with the given sub-skill.