МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2003
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2003
Proceedings of the Thirty Second Spring Conference of
the Union of Bulgarian Mathematicians
Sunny Beach, April 5–8, 2003

THE EFFECT OF TEST COMPONENT WEIGHTING ON RELIABILITY

Eugenia Stoimenova, Radost Vassileva

The purpose of this study is to illustrate how test component combinions and differential weights affect the reliability of the Composite scale. Reliability provides information on how consistent the test scores are. An important consideration in combining the components is Composite reliability because this index is affected when multiple components are differentially weighted.

1. Introduction. In many educational achievement test situations, composite scores are formed from test scores on different tests. The composite scores are then considered as a single index of achievement, and are often used for decision making. When composite scores are used for such decisions, it is necessary to consider the precision of the composite scores as well as of their components. If the component test scores are positively correlated, and the components are positively weighted, the error variance in the composite will be smaller than that of any of the component scores. The reliability of the composite would be correspondingly higher.

The proposed didactic system consists of five quizzes on First degree Ordinary Differential equations (ODE). The 18 test problems (items) are allocated into five Quizzes as follows. Quiz 1 is an entry test and includes 7 items: problems 1 to 7. Students should have a basic knowledge in Calculus to solve the items. Quiz 2 includes 2 items: problems 8 and 9. These items train cover separate variables equations and equations reducible to homogeneous type. Quiz 3 includes 3 items: problems 10, 11 and 12. They cover linear equations, Bernoulli equations and Riccati equations. Quiz 4 includes 2 items: problems 13 and 14. They cover equations allowing integrating factor and exact differential equations. Quiz 5 is a final exam and includes 4 items: problems 15, 16, 17 and 18. They cover Clairaut's equations, equations allowing integrating factor, Bernoulli equations and homogeneous equations.

The situations described in this paper are ones in which tests having similar purposes are combined to form a single composite score. Today, many large-scale testing programs use composite scores to make decision about examinees. The general procedures and results of this study can be applied to other situations, and will provide information about selecting a combining method and weight based on reliability.

2. Scale distributions. The scale scores are calculated just as a sum of the item scores included in the corresponding Quiz; the composite score is a sum of the five quiz scores. The maximum possible score on Quiz 1 is 36 points, on Quiz 2 – 25 points, on

Quiz 3 – 28 points, on Quiz 4 – 16 points, on Quiz 5 – 36 points, and on the Composite scale of all 18 items – 141 points. The tests are administered to 66 students, (Third Year Degree Course). The first four columns in Table 1 refers to the item's number, its item points(maximum possible item score), its obtained mean and standard deviation, respectively. The second part of the table refers to the scale's number, its mean and standard deviation.

Item	max points	item mean	standard deviation	Scale	scale mean	scale st. dev.
Item 1	4	2,47	1,62	Quiz 1	16.74	8.63
Item 2	5	3,79	1,51			
Item 3	7	2,38	2,67			
Item 4	4	2,27	1,71			
Item 5	6	2,76	2,39			
Item 6	5	1,74	2,14			
Item 7	5	1,33	1,94			
Item 8	6	3,48	2,40	Quiz 2	14.15	7.21
Item 9	19	10,67	5,85			
Item 10	6	4,32	1,96	Quiz 3	14.56	7.21
Item 11	11	5,50	4,01			
Item 12	11	4,74	3,60			
Item 13	6	4,00	2,25	Quiz 4	9.26	5.55
Item 14	10	5,26	3,87			
Item 15	7	5,20	2,02	Quiz 5	21.88	10.76
Item 16	7	5,39	2,25			
Item 17	11	5,44	4,51			
Item 18	11	5,85	4,43			

Table 1. Item points, means and standard deviations.

The scale difficulty is calculated as a ratio of the scale mean compared to the scale points. Since the statistic is a percentage, its range is from 0 to 100.Higher difficulties indicate easier quizzes (see [2] for more information).

The difficulties of all quizzes are middle; the easiest Quiz is Quiz 5 collected 60.1% of the possible score; the most difficult quiz is Quiz 1 collected 46.5% of the possible score

Scales intercorrelations. The matrix in Table 2 shows how each scale relates to the other scales, and to the composite scale. All correlations are very high which indicates that students with high score on one scale also have high score on the others. The correlations of the scales with the composite one are higher because the scales are parts of the last.

Indices of Reliability. A useful way to conceptualize the consistence of a set of scores on an exam is to think about what would happen to those scores if we could give the exam a second time to the same group of students under identical conditions. The square of the correlation between the scores on the two occasions is called the "reliability coefficient" and it tells us how well we can predict scores on the second occasion from the

Scales	Composite	1	2	3	4	5
Composite	–					
Quiz 1	.9393	–				
Quiz 2	.9176	.8242	–			
Quiz 3	.9521	.8550	.8495	–		
Quiz 4	.9030	.7940	.7768	.8384	–	
Quiz 5	.9803	.9086	.8683	.9230	.8838	–
Scale reliabilities (alpha)	.9566	.7194	.4853	.7297	.7151	.7761
Composite reliability	.9306					

Table 2. Scales intercorrelations.

scores on the first occasion using a linear equation. The more homogeneous the items are in terms of the content they cover, the higher the inter-item correlations, the higher the reliability.

We estimate two indices of test reliability: Cronbach's coefficient alpha and Composite reliability. Estimates of reliability show how consistent scores are likely to be from one administration of a test to another. If a person answers to a similar set of items on another occasion, the resulting score should be similar, too. The near the reliability is to its upper limit 1, the more likely it is that a person's score remains near the score achieved on the current scale. In practice, values of the estimates greater than 0.95 are rare.

Although calculated differently, both indexes are indicators of the internal consistency of the test or the extent to which parts of the test could work together to measure the same underlying construct. Reliability indices range from 0 to 1. Cronbach's coefficients for the six scales (Quiz 1, 2, 3, 4 and 5, and composite scale) are presented in Table 2. Their values are between 0.71 and 0,78 except Quiz 2 which value is 0.4853. Cronbach's alpha for the composite scale is 0.9566.

When the composite score is a sum of scores on two or more tests, the Composite reliability should be used instead of Cronbach's alpha. The components of such a composite score may well measure different things; nevertheless, the idea of its reliability remains the same. The reliability estimates for different sections are combined to obtain an estimate of the composite score reliability. When the tests that are combined into the composite score have unequal precision at different score levels, it is not a simple matter to combine the unequal score variances.

The formula for the composite reliability (see [1]) uses scales intercorrelations and takes into account that items are grouped into consistent scales. The composite reliability of the test is 0.9306.

3. Single test component weighting. Scores for many tests are computed as a weighted sum of scores on two or more items. Define the score as

$$X = \sum w_i Y_i, \tag{1}$$

where Y_i is the the score on item i and w_i is the weight for the item i.

There are many ways to state what weights are for such linear combination. Instructors often weight items or sections of classroom tests based on their feel for the task difficulty. This approach appears to be attractive because it provides additional reward for mastering particularly difficult concepts.

Cronbach's alpha for weighted score takes the form

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum w_i^2 \sigma_{Y_i}^2}{\sigma_X^2}\right], \tag{2}$$

where $\sigma_{Y_i}^2$ is the variance of the score on item i, σ_X^2 is the variance of the scale score, and k is the number of items.

Consider the Quiz 2 on ODE that comprises Item 8 and Item 9 and has obtained the lowest reliability, 0.48. (see Table 2). Score on Item 8 is allocated half of the testing time and the maximum score on this item is 6 points while the maximum score on Item 9 is 19 points (see Table 1). The observed correlation between scores on these two items is approximately 0.45. We want to include weights in the model in order to increase the reliability of this scale.

We calculate the score as weighted combination of the score on Item 8, weighted by 3/4, and the score on Item 9, weighted by 1/4. The 3:1 ratio of the weights was selected to represent the 3:1 ratio in maximum scores of the two items. The Cronbach's alpha of the weighted scale increases to 0.617.

The formula (2) for the reliability of weighted scores assumes that the weights are known. Suppose that we do not know how we would like to weight the items in the Quiz 2. Our goal is to choose weights such that to maximize reliability.

If there are no constrains on weights values, then all of them could be chosen as small as we like and therefore the reliability would tend to 1. In order to exclude this case we assume that weights are nonnegative and satisfy $\sum w_i = 1$.

The score on Quiz 2 is presented by

$$X = wY_8 + (1-w)Y_9. \tag{3}$$

The exact weights for maximizing scale reliability can be determined by setting the first derivative of (3) with respect to w to zero. The solution is

$$w = \frac{\sigma_{Y_9}}{\sigma_{Y_8} + \sigma_{Y_9}},$$

where σ_{Y_8} and σ_{Y_9} are standard deviations of the scores on Item 8 and Item 9, respectively.

Thus optimal weights in Quiz 2 are 0.7 and 0.3 for Item 8 and Item 9, respectively. The maximum possible reliability of Quiz 2 is 0.624. The weighted score on Quiz 2 reflects to Cronbach's alpha and Composite reliability of the Composite scale. Cronbach's alpha decreases to 0.9367 while Composite reliability increase to 0.9345. The last is much more important since Composite reliability takes into account the internal consistence of the five components.

4. Composite test component weighting. Scores for many tests are computed as a weighted sum of scores on two or more components. Define the composite score as

$$Z = \sum w_i X_i, \tag{4}$$

where X_i is the the score on component i and w_i is the weight for the score on component i.

The following formulas used for combining reliabilities take the weighting of components into account

$$(5)\qquad \alpha^* = \frac{\sum_{i=1}^{n} w_i^2 \sigma_{X_i}^2 \alpha_{X_i} + 2\sum_{i<j} w_i w_j \rho(X_i, X_j)\sigma_{X_i}\sigma_{X_j}}{\sum_{i=1}^{n} w_i^2 \sigma_{X_i}^2 + 2\sum_{i<j} w_i w_j \rho(X_i, X_j)\sigma_{X_i}\sigma_{X_j}},$$

where k is the number of scales, $\sigma_{X_i}^2$ the variance of the score on i-th scale, α_i is Cronbach's alpha of the i-th scale, $\rho(X_i, X_j)$ is the correlation between score on i-th scale and score on j-th scale.

If the components of a linearly combined composite are positively correlated, there is always some weighting scheme that yields maximum reliability for the composite.

Consider the didactic system of 5 quizzes – but now imagine that we do not know how we would like to weight the score on Quiz 2 in the composite. Suppose that all components except Quiz 2 are equally weighted and denote the weight of their sum by w. The weight of Quiz 2 is then $(1 - w)$ and the composite score is formed by

$$Z = w \sum_{i \neq 2} X_i + (1 - w)X_2.$$

Now we can reconsider the composite reliability for the compose test as a function of the proportional weight w (the proportional weight for Quiz 2). Figure 1 shows reliability of the test as a function of the w.

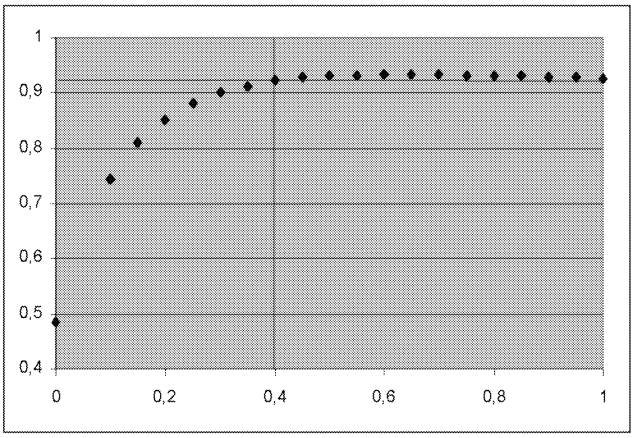

Fig. 1. Reliability as a function of w.

We note first the obvious: When $w = 1$ (that is, Quiz 2 is excluded from the composite), α^* is equal to the reliability of the rest 4 test as a composite, 0.926. When $w = 0$, α^* is equal to the reliability of the Quiz 2, 0.485. A little less obvious, for values of $w > 0.4$ reliability exceeds the reliability of 0.926, whereas for values less than 0.4 the curve drop fairly precipitously toward 0.485, the lower of the reliability of the two components.

Further, note that for the Quiz 2, the graph of $\alpha^*(w)$ is relatively flat in the general vicinity of its maximum. However, below $w = 0.4$, composite reliability drops sharply. This means that it would be wise to choose weights somewhere in the general vicinity of those that yield maximum composite reliability – or at least, one should be aware of how much less reliability than the maximum any particular set of weights might yield.

The exact maximum for the Composite reliability can be determined by setting the first derivative of (5) with respect to w to zero. If we use the original scale of Quiz 2 (unweighted Items 8 and 9), the solution is 0.36 for Quiz 2 and 0.64 for the other scales. The composite reliability increases to 0.9337. However, we can use the optimal weighted score on Quiz 2, obtained in section 3, and find the optimal solution for the component scales. The solution is 0.41 for Quiz 2 and 0.59 for the other scores. The corresponding composite reliability increases to 0.9351.

REFERENCES

[1] P. Mateev, E. Stoimenova. Reliability and correct estimating of exams and tests. *Mathematics and Education in Mathematics*, **21** (2001), 95-102 (in Bulgarian).

[2] E. Stoimenova. Measurement quality of tests. New Bulgarian University Press, 2000 (in Bulgarian).

Eugenia Stoimenova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: jeni@math.bas.bg

Radost Vassileva
South-West University
Blagoevgrad, Bulgaria
e-mail: ov80@abv.bg

ВЛИЯНИЕ НА КОМПОНЕНТИТЕ НА ТЕСТА ВЪРХУ НАДЕЖДНОСТТА

Евгения А. Стоименова, Радост Василева

В статията се разглежда влиянието на различните компоненти на един тест върху надеждността на Композиционната скала. Изследва се надеждността като функция на тегловите коефициенти. При комбинирането на тестовите компоненти се отчита повишаването на композизионната надеждност, която е мярка за съгласуваността на тестовите компоненти.