DOCUMENT RESUME

ED 334 207 TM 016 610

AUTHOR Thayer, Jerome D.
TITLE Use of Observed, True, and Scale Variability in Combining Students' Scores in Grading.
PUB DATE Apr 91
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Elementary Secondary Education; *Grading; Mathematical Models; Scaling; *Scoring; Test Results; *Weighted Scores
IDENTIFIERS *Composite Scores; *Variability

ABSTRACT

Combining student scores to form subtotals and finally a total score to determine a grade is discussed. The composite score reached by combining measures or subtotals is only valid when the scores are combined so that the actual weight of each measure or subtotal in the total score is the same as the intended weight. Three types of variability should be considered in combining scores: (1) variability of true achievement (true variability); (2) variability that would exist if all possible levels of achievement had been measured (scale variability); and (3) variability of the observed scores (observed variability). Weighting procedures are summarized for teachers who want to use standard or converted scores and for those who want a simple weighting scheme or no weighting scheme. It is suggested that true variability and scale variability should be considered in addition to observed variability in weighting scores. If differences in true achievement cannot be ignored, one of the following options should be used: using a common scale for each measure; modifying possible points; and modifying scoring procedures. Sixteen tables provide illustrative examples. (SLD)

Use of Observed, True and Scale Variability in Combining Students' Scores in Grading

Jerome D. Thayer

Andrews University

A Paper Presented at the Annual Meeting of the National Council on Measurement in Education

Chicago, Illinois

April 5, 1991

Importance of Weighting

Most teachers base student grades on more than one measure. A "measure" as used in this paper is anything that is evaluated in such a way as to assign a number to student achievement. A measure could be a quiz, an assignment, an examination, a project or a term paper. The scores for each of these measures are typically entered into a grade book and then combined at the end of the grading period to form subtotals (perhaps for quizzes, assignments, tests, etc.) and finally a total score which is used to determine the grade for the student. In order to allow the measures to have appropriate weights, adjustments are frequently made to the measures or subtotals before combining.

The intended weight for each measure is a function of the importance of the objectives that the measure covers. If an exam and a project each cover objectives of equal importance they should be combined in such a way as to have equal weight. The intended weight for the exam and project would then be equal. The intended weight is the weight assigned by the teacher. A teacher who makes a course outline that states that quizzes and the final exam each count for 50% of the grade is giving equal intended weights to these two types of measures.

The actual weight is the weight that truly exists which may or may not be the same as the intended weight. In the example in Table 1, Joe and Fred take a 50 point true-false exam and turn in a 50 point term paper. Fred knew nothing on the exam and simply marked true for every answer, scoring 50% and receiving a grade of F. Joe was too busy practicing for the basketball team and so did not turn in a term paper, receiving a score of 0% and a grade of F. Joe knew all of the objectives covered on the exam and Fred turned in a "perfect" term paper so they each received 100% (A) on these measures. The teacher wanted the exam and term paper to have the same weight (intended weights were equal) so therefore gave each 50 points. Even though Joe and Fred each achieved equally (perfect on one measure and nothing on the other), when the scores were added together, Joe received a much lower score than Fred, because the term paper had an actual weight twice as high as the exam. The actual weights were much different than the intended weights.

Table 1

Grade Book Data -- Example 1

	True/False Exam		Term Paper		Total
Student	Score	Grade	Score	Grade	Score
Joe	50 (100%)	A	0 (0%)	F	50 (50%)
Fred	25 (50%)	F	50 (100%)	A	75 (75%)

This example supports the idea that variability should be equated before combining scores. The example in Table 2 supports the idea that there are times when variability should not be equated before combining scores.

Table 2

Grade Book Data -- Example 2

Student	Test 1	Test 2	Total
Joe	90	10	100
Fred	100	100	200

Fred knew all of the objectives for both tests and received a perfect score on each. Joe was almost perfect on the first test but failed the second one. If Fred and Joe were the only students in the class and if both measures were converted to scores that would make the variability of the two measures equal, it would appear as if Joe scored equally well on both tests. If both tests were composed of 100 spelling words randomly chosen from a list of 500 words and if upon observation they appear to be equally difficult tests, Joe's scores of 90 and 10 should not normally be treated as equally lower than Fred. There would be some situations, however, when they would want to be equated. This might be desired if the instructional period for Test 2 was short, the instruction was inferior, etc.

The composite score arrived at by combining measures or subtotals is only valid to the extent that the scores are combined properly so that the actual weight of each measure or subtotal in the total score is the same as the intended weight. The way variability should be treated is complex: in some cases the variability should be used in the combining, in other cases it should not be.

Typical Teacher Procedures

The usual way teachers deal with weighting is to select the possible points for each measure to be proportional to the weight intended and then just add up the scores at the end of the grading period with no adjustment at that time. If scores, measures or subtotals are adjusted before combining, it is done so that the possible points are proportional to the intended weight.

This method is appropriate if two conditions both exist: the course grades are based on percentage of total points (a type of criterion-referenced grading) (Oosterhof, 1987), and the scales used for every measure have the same scale variability (defined later in this paper). Where course grades are not based on percentage of points earned, usually a non-mathematical system is used where subjective judgment is used to combine the measures and precise weighting is not an issue. This system will not be discussed in this paper. Where percentage of points is used, measures frequently do not have the same scale variability. In these situations, teachers usually do not deal with variability in the proper manner.

To equate the variability of measures, measurement textbooks such as Gronlund and Linn (1990) usually recommend converting all grade book entries to standard scores such as T scores or stanines and then multiplying each score by its intended weight before being combined. Teachers who are aware of this recommendation typically ignore it either because they do not understand it, conversion is too difficult or takes too much time, or they feel that conversion is not worth the effort. Teachers who use computer grading or spreadsheet programs, however, can implement this procedure quite easily. However, even if a way was found to make the use of standard scores easy for teachers, converting scores to standard scores in some respects just replaces one kind of invalid weighting with another. The reason for using standard scores is the fact that the weight of a measure is related to the variability of the measure. But the recommendation does not account for different types of variability: observed, true, and scale variability. It assumes that there is one type of variability which must be equated for all measures before combining.

This paper attempts to clarify how different types of variability should be taken into consideration when combining scores.

Types of Variability To Consider When Combining Scores

Variability is a characteristic of scores that must be considered when combining scores. If variability is not considered, the combined total scores are likely to be invalid. Whereas most teachers think that the possible points of a measure is the major determining factor in the weight of the measure, in fact only as modifying the possible points changes the variability of the measure does it influence the weight. The possible points for different measures may or may not be related to their variabilities.

There are three types of variability that should be considered in combining scores: true variability, scale variability, and observed variability.

True variability is the variability of true achievement on the variable that is being measured. It is the variability of true achievement that actually exists for the students in a given class on a scale from 0% for no achievement to 100% for perfect achievement. If the scores on a test equalled the true achievement for each student, a student with no achievement would get a score of 0%; it would be impossible to gain any points by guessing or bluffing. A student with perfect achievement would get a score of 100%; it would be impossible to lose points by poor test-taking behavior or imprecise subjective scoring by the teacher. If all students in a class had perfect achievement the true variability would be zero, no matter what type of measure was used. In a perfectly valid measure the observed scores are perfectly correlated with true achievement but the observed variability is not necessarily equal to the true variability. Observed variability would only equal true variability if the scale measured achievement on a scale from 0%-100%. On a true-false test in which some students knew nothing and some knew everything, using the range as the measure of variability, the observed variability would be from 50%-100% while the true variability would be from 0%-100%. The test could be considered to be a valid measure of true achievement because the observed and true scores would be perfectly correlated: the true scores could be computed using a correction for guessing formula (Observed Correct - Observed Incorrect = True Score).

Scale variability is the variability that would exist if persons of all possible levels of achievement (from perfect achievement to no achievement) had been measured. Scale variability would be based on scores that might range from 0%-100% for an essay test to 50%-100% for a true-false test. Unless a class is extremely large, it is unlikely that either the observed or true variabilities would equal the scale variability b cause a class seldom contains both a student that knows everything and a student that knows nothing. Scale variability is usually larger than true variability.

Observed variability is the variability of the observed scores (the actual student scores as recorded in a grade book). Observed variability is related to both true variability and scale variability but is usually smaller than either because poor students usually gain some points by guessing or bluffing and good students usually lose some points through poor test-taking or poor teacher scoring.

Differences in variability between measures may result from three things: differences in true achievement levels of students, differences in the difficulty of the measure, and difference in the type of scale being used. The following chart illustrates the relationship between the three types of variability and the influences on them.

Table 3

Influences on Various Types of Variability

	Influences on Variability		
Types of Variability	True Achievement	Difficulty	Type of Scale
True Variability	x		
Scale Variability			x
Observed Variability	x	x	x

Differences in True Achievement

On measures of equal difficulty and scale, usually the top scores of students are consistently in the 90-100% range but frequently the lower end of the distribution varies markedly due student-related differences such variations in effort for some students or external factors affecting the amount of learning of some students such as social events the night before an exam. Grades should take into consideration these differences in achievement; they should not be removed by the weighting process. Where students learn more, they should normally get higher grades.

Differences in Difficulty

Teachers frequently construct measures of varying difficulty. Measures used primarily for motivation or practice such as daily assignments or daily quizzes are usually easier while measures used solely for grading purposes such as exams and term papers are usually harder. Whereas the level of difficulty should only affect the mean score of the class, extremely easy measures frequently result in negatively skewed distributions because of an artificial upper limit to the distribution of scores and thus the observed variability is smaller than it would be if the measure was more difficult.

Teachers frequently adjust their grading scheme for measures that are unusually difficult (either too easy or too hard). Particularly for measures that are more difficult than expected extra points will be given to all students or they will be graded from the highest student rather than the total possible. These procedures only adjust the central tendency of the scale and not the variability. Usually only measures that are much easier than normal will affect the scale variability. In these cases the artificial upper limit in a sense changes the scale variability. While not always reducing the range of scores (the poorest student may approach the bottom no matter what the difficulty is), the standard deviation is usually reduced due to the truncated upper portion of the distribution. These differences in variability due to differences in difficulty should be removed in the weighting process. This frequently cannot be done after the measure has been scored. If there is an upper limit that truncates the distribution, the distribution of observed scores cannot be converted to a distribution that is perfectly correlated with the distribution of true achievement. The correct way to deal with differences in difficulty is to construct each measure to have sufficient difficulty so there is no artificial upper limit to the distribution (no 100% scores unless there is "perfect" achievement). Any difference in difficulty between the measures would then only affect the central tendency of the distributions and not the variabilities.

Differences in Type of Scale

Teachers frequently construct both objective (multiple choice, true-false) and subjective (essay, compositions, term papers) measures for the same course. The variabilities of these scales range from 0%-100% for essay tests to 50%-100% for true-false tests. Differences in observed variability between measures due to differences in scale should be removed in the weighting process.

Variability and Weighting

Textbooks do not make a distinction between variability due to these three causes when recommending use of standard scores to equate variability in weighting. If the weighting procedure uses observed variability (standard deviation) as textbooks recommend, any differences in variability due to differences in true achievement will be eliminated in addition to those due to difficulty and type of scale. If there are no differences in observed variability between measures due to differences in true achievement level and if none of the measures have an artificial upper limit, then differences in observed variability will be the same as the differences in scale variability and textbook recommendations using standard scores would result in valid total scores. In most cases, however, this would not be true. The procedures described in this paper show how to remove variability due to difficulty and type of scale while not removing differences due to true achievement.

The following three examples illustrate how differences in observed variability can be caused by differences in the type of scale used, differences in true achievement, and differences in difficulty. In each example there are two measures, both with 50 points possible, that are to have equal weight when combined to form a total score.

1. Difference in the type of scale used.

For the example in Table 4 Joe was "perfect" on the exam and did nothing on the term paper. Fred knew nothing on the exam and did a "perfect" term paper. Since each student was perfect on one measure and knew nothing (or did nothing) on the other measure they should get the same total score. Differences in the type of scale make the total scores invalid if the scores are combined without weighting as are done here.

Table 4

Grade Book Data -- Example 1

	True/False Exam		Term Paper		Total
Student	Score	Grade	Score	Grade	Score
Joe	50 (100%)	A	0 (0%)	F	50 (50%)
Fred	25 (50%)	F	50 (100%)	A	75 (75%)

2. Difference in true achievement.

The two 50-point essay exams in the example in Table 5 are scored in a similar manner -- 0% for no achievement and 100% for perfect achievement (same scale variability) and are equivalent in difficulty. Joe did not study at all for the second essay exam and got the score/grade he deserved (0%/F). Fred studied a bit for the first exam and learned half of the material and got the score/grade he deserved (50%/C). Both students learned everything on one of the exams. Since Fred learned more of the material covered on the two exams he should get a larger total score which in fact he did. The difference in variability between the two measures which was only due to true achievement should not be removed.

Table 5

Grade Book Data -- Example 3

	Essay Exam 1		Essay Exam 2		Total
Student	Score	Grade	Score	Grade	Score
Joe	50 (100%)	A	0 (0%)	F	50 (50%)
Fred	25 (50%)	C	50 (100%)	A	75 (75%)

3. Difference in difficulty.

The two 50-point essay exams in the example in Table 6 are scored in a similar manner -- 0% for no achievement and 100% for perfect achievement (same scale variability). Exam two had a truncated upper limit -- its mean score was 90% while the mean of Exam one was 80%. Joe was "perfect" on the first exam and Fred was "perfect" on the second exam. Both Joe and Fred forgot to study the same amount of material for one of the exams. Since both students learned the same amount of material for the two exams, their total scores should be the same. Differences in scores due to the differences in difficulty should be removed before combining to form the total score.

Table 6

Grade Book Data -- Example 4

	Essay Exam 1		Essay Exam 2		Total
Student	Score	Grade	Score	Grade	Score
Joe	50 (100%)	A	45 (90%)	B	95 (95%)
Fred	40 (80%)	B	50 (100%)	A	90 (90%)

This paper will deal with the procedures for properly dealing with true, observed, and scale variability for correct weighting with two types of teachers:

1. teachers who want to use standard scores (precise control of variability using textbook recommendations) in weighting
2. teachers who want to use a simple weighting scheme or no weighting scheme.

Weighting Procedures for Teachers Who Want to Use Standard or Converted Scores

The procedures for this section deal with conversion of individual scores into either standard scores or converted scores. In the conversion process, standard scores use the observed mean and standard deviation. For this paper the term converted score will be used when a variability measure other than the observed standard deviation is used during the conversion.

Teachers who want to use standard or converted scores to have precise weighting should follow these steps:

1. At the end of the grading period, for each measure determine whether the variability of the measure is different from the other measures in the class due to variations in true achievement or due to differences in difficulty or type of scale used.

2. Decide whether to use an observed or estimated measure of variability as the variability score to use in computing the standard or converted scores.

   a. If the observed variability is equal to the scale variability then it can be assumed that there are no differences due to difficulty or true achievement and the observed variability is the proper one to use. The observed standard deviation can also be used if grades are to be based only on relative achievement (norm-referenced) and not absolute achievement (criterion-referenced). If, for example, the lowest scores on two 10 point true-false quizzes (such as 6 on quiz one and 3 on quiz two) are to be treated as equally poor achievement (ignoring or assuming no differences in difficulty or true achievement), the observed standard deviation is appropriate for equating the variability of the two quizzes.

   b. If there are differences in variability due to difficulty or true achievement then an estimated measure of variability should be used. Using the two quizzes of the previous example where the highest score on each quiz was 10 points, it could be assumed that the differences in range of 6-10 on quiz one and 3-10 on quiz two were differences in difficulty and/or true achievement and therefore the two quizzes should be treated as measures with equal variability. In this case the same estimated variability number (range or standard deviation) should be used for the two quizzes.

   The range (observed or estimated) can be used if the class is large and only approximate adjustment of scores for variability is needed. If the range is used to convert the scores to a constant scale variability, the resulting scores are not standard scores but will be called converted scores in this paper.

3. Convert each score to a standard score (z or T) or converted score.

4. Multiply each score by its intended weight.

5. Add up the scores to form the total score.

The following three examples illustrate how this method can be implemented.

1. Using the actual range to make converted scores.

   Each score in the example in Table 7 is converted to the values found in Table 8. Since each quiz is to have half the weight of the exam (based on their intended weights), the scores for the quizzes should be converted to have a range of 16. This can be done by multiplying each Quiz 1 score by 4 and each Quiz 2 score by 2 to give ranges of 16, 16, and 32 which are proportional to the intended weight. The numbers are then summed to form the total score. The Possible, Mean, and Standard Deviation numbers are not used in this procedure.

Table 7

Grade Book Data -- Example 5

	Quiz 1	Quiz 2	Exam
Student 1	10	0	80
Student 2	6	8	100
.			
.			
.			
Possible	10	10	100
Mean	8	6	80
Standard Deviation	2	4	15
Range	4	8	32
Intended Weight	1	1	2

Table 8

Converted Scores -- Example 5

	Quiz 1	Quiz 2	Exam	Total
Student 1	40	0	80	120
Student 2	24	16	100	140
.				
.				
.				
Possible	40	20	100	160
Mean	32	12	80	124
Standard Deviation	8	8	15	
Range	16	16	32	

2. Using the observed standard deviation to make T scores.

The scores in the example in Table 9 are converted to the T scores in Table 10. These scores are multiplied by their intended weight and summed to form the scores in Table 11. The formula used to compute the T scores is:

$$T = 10\left[\frac{score\text{-}mean}{student}\right] + 50$$

Table 9

Grade Book Data -- Example 5

	Quiz 1	Quiz 2	Exam
Student 1	10	0	80
Student 2	6	8	100
.			
.			
.			
Possible	10	10	100
*Mean	8	6	80
*Standard Deviation	2	4	15
Intended Weight	1	1	2

*Used for conversion to T scores

Table 10

T Scores -- Example 5

	Quiz 1	Quiz 2	Exam
Student 1	60	35	50
Student 2	40	55	63
.			
.			
.			
Mean	50	50	50
Standard Deviation	10	10	10
Intended Weight	1	1	2

Table 11

Combined Scores -- Example 5

	Quiz 1	Quiz 2	Exam	Total
Student 1	60	35	100	195
Student 2	40	55	126	221
.				
.				
.				
Mean	50	50	100	200
Standard Deviation	10	10	20	

3. Using an estimated scale standard deviation to make T scores

For the example in Table 12 let us assume that the two quizzes contained the same type of questions and were of similar difficulty. However, students did much better on quiz one than quiz two, which resulted in a smaller observed standard deviation. In a similar manner, students did better on the exam, resulting in a small observed standard deviation.

Since both quizzes and the exam were of the same type of questions (equivalent scale variability), the difference in observed variabilities could be attributed to a difference in true achievement rather than in difficulty or scale. Therefore, estimated standard deviations were used that were proportional to the possible points for the quizzes and exam. The observed standard deviation of Quiz two (4) which was 40% of possible (10) was used as a standard and the estimated standard deviations for Quiz one and Exam were also selected to be equal to 40% of possible. The estimated standard deviations were used to form the T scores in Table 13, then multiplied by the intended weights and combined to form the total scores in Table 14.

Table 12

Grade Book Data -- Example 5

	Quiz 1	Quiz 2	Exam
Student 1	10	0	80
Student 2	6	8	100
.			
.			
Possible	10	10	100
*Mean	8	6	80
Observed Standard Deviation	2	4	15
*Est. Scale Standard Deviation	4	4	40
Intended Weight	1	1	2

*Used for conversion to T scores

Table 13

T Scores -- Example 5

	Quiz 1	Quiz 2	Exam
Student 1	55	35	50
Student 2	45	55	55
.			
.			
.			
Mean	50	50	50
Standard Deviation	10	10	10
Intended Weight	1	1	2

Table 14

Combiued Scores -- Example 5

	Quiz 1	Quiz 2	Exam	Total
Student 1	55	35	100	190
Student 2	45	55	110	210
.				
.				
.				
Mean	50	50	100	200
Standard Deviation	10	10	20	

Since the observed standard deviation is frequently influenced by differences in true achievement, if the observed standard deviation is used in computing the standard scores, these important differences are removed. If an estimated standard deviation is used as recommended in this procedure, much of the advantage of the textbook recommended method is lost: it is no longer objective and precise--subjectivity is introduced into the method, and a computerized grading program can not be used to automatically combine the scores. It would generally be more convenient and probably just as accurate to use one of the methods that will now be described.

Weighting Procedures for Teachers Who Want to Use a Simple Weighting Scheme or No Weighting Scheme

There are three simple procedures that will result in accurate combining of scores. In all methods scores can be simply added up at the end of the grading period with each measure having its proper weight in the total score. Method one involves using the same type of scale for all measures while methods two and three use different scales but adjust procedures for selecting possible points and scoring subjective measures.

Method One: Use the Same Type of Scale for All Measures

A teacher who wants properly weighted scores without using a "weighting" process should do the following:

1. Use measures of sufficient difficulty to avoid truncating the variability due to an artificial upper limit.

   Have measures in which the top student can be expected to score less than 100%. On objective measures this can be done by selecting items with an appropriate mixture of difficulty. On subjective measures a more rigorous scoring system should be used. Any differences in difficulty would then only affect the central tendency of the distribution and not the variability.

2. Use the same type of scale for all measures.

   Examples of measures with different types of scales would be multiple choice, true-false, or subjective (essay, term papers, projects, etc.). The important thing that must remain constant is the range of scores between perfect (highest possible "A" grade) and absolute failure (lowest possible "F" grade). Multiple choice items or tests with 4 options per question have a scale range from 25% to 100%. True-false items or tests have a scale range from 50% to 100%. Subjective items or tests have a scale range from 0% to 100%. If the same type of scale is used for each measure and the measures are of appropriate difficulty, a simple adjustment of possible points is all that needs to be done for proper weighting.

3. Select possible points for each measure that are proportional to the intended weights.

   If the same scale is used for each measure and the measures are of appropriate difficulty, the possible points will be proportional to the actual weights.

4. Score each measure based on the chosen possible points and record the score in the grade book according to normal procedures.

5. Add up the scores in the grade book according to normal procedures.

   The total scores will be properly weighted.

To implement this procedure, at the beginning of the grading period the teacher must consider the level of difficulty for the measures and decide on the type of scale to be used for all the measures during the grading period.

1. Select the difficulty of the meast⁻⁻es.

   If items are chosen or subjective scoring is done so that the mean of the scores is in the range of 60%-80%, it is unlikely that there will be an artificial upper limit to the scores. The standard deviations of scores for typical classroom measures usually are between 10%-15%. For measures with a mean of 80%, the standard deviation is typically about 10%, with very few students getting 100%.

2. Select the type of scale for all measures.

   There are three types of scales frequently used in classroom testing.

   a. 0% (chance) -100% (perfect) scale

      This scale is found with short answer and essay questions, projects, term papers, speeches, etc. This scale could also be used for a test of objective items in which a correction for guessing formula is used.

   b. 10-25% (chance) -100% (perfect) scale

      This scale is found with most multiple choice or matching items.

   c. 50% (chance) - 100% (perfect) scale

      This scale is found with true-false items.

There would be two ways to arrive at the possible points for each of the measures to use during the grading period: deciding the possible points for all measures at the beginning of the grading period, or deciding the possible points on each measure as it is being constructed.

The following examples illustrate how each would be used. Let us assume that the teacher has decided to use all multiple choice items for all of the quizzes and the final exam.

1. Deciding possible points at the beginning of the grading period.

Table 15

Class Situation -- Example 6

Type of Measure	Weight
1 Final Exam	33%
6 quizzes of equal weight	66% (11% each)

Based on the weights given in Table 15, the possible points for each quiz should be 1/3 that of the Final Exam. If a 100 point final was anticipated, each quiz should be 33 points. If 5 point quizzes were planned, the final exam should be 15 points. The number of questions on each quiz and the Final Exam would not necessarily need to be the same as the number of points. A 100 point final could have 50 questions (2 points/question) or 200 questions (½ point/question).

2. Deciding possible points during the grading period.

Table 16

Class Situation -- Example 7

Type of Measure	Weight
1 Final Exam	33%
Many quizzes of equal weight	66%

Using the data in Table 16, if the teacher did not know how many quizzes were to be given during the quarter it would be impossible to determine the possible points for the quizzes and the Final Exam at the beginning of the grading period. The teacher should just decide on how many points to use for each quiz. Whatever possible was used for the first quiz would be used for all the others. When it was time for the Final Exam it would have half as many points as the sum of all of the quizzes combined. In order to do this it is likely that the number of points per question on the Final Exam would have to be adjusted.

Since the difficulty and scale differences are controlled with this method, any differences that exist in the observed variabilities of the measures would be due to differences in true variability which should not be removed. For example, if the range of scores (as an estimate of variability) on the first two quizzes (both multiple choice) were 3-10 on Quiz one and 6-10 on Quiz two, the lowest score on Quiz two (6) would indicate a higher level of achievement than the lowest score on quiz one (3).

Since this method requires all objectives to be measured using the same scale, this would be a difficult method to apply since there are some types of objectives that cannot be measured in a valid manner with objective items (i.e., multiple choice items would be invalid for measuring the ability to organize knowledge). This leaves the teacher with the alternative of using an invalid item type to measure these objectives or else measure all objectives using a 0%-100% scale which would suggest essay or short answer items which are very inefficient for many objectives (i.e., factual), especially with large classes. However, objective items could be used if a correction for guessing formula was used (this would not be viewed as desirable by many teachers). The last two methods suggested do not have this problem.

Method Two: Modify possible points

With this method different types of scales can be used to as needed to measure different types of objectives or allow for teacher preference for certain types of measures. The possible points for each measure are modified taking into consideration the type of scale used. The observed variabilities of the scores will be made proportional to the intended weights.

1. Select the difficulty level of the measures.

Use the same procedures described for method one above.

2. Select the possible points for each measure.

Choose possible points for each measure taking into consideration both the intended weight and an estimate of the scale variability of the scale used. For two measures that are to have equal weights, if the scale variability of one measure is larger, its possible points should be proportionately smaller to have the observed variability and therefore the actual weight remain equal.

Two examples illustrates this method. For the first example a test and a project that should have equal weights are to be combined. The test will be objective items (all 4-option multiple choice) with a scale ranging from 25% (no knowledge -- guessing or chance) to 100% (perfect knowledge). The project will be subjectively graded depending on how many of the specified components are present on a scale ranging from 0% (no knowledge or not handed in) to 100% (perfect project).

If the teacher assigned the same possible point to the test and the project (which students would expect since they are to be equally weighted), the scores would have to be adjusted before they are combined in order to arrive at a valid grade since they are graded on different scales.

If the actual range of scores on the multiple choice test is 50% to 100% and the actual range of scores on the project is 30% to 100%, the teacher will need to determine to what extent the differences in variability are due to differences in achievement or differences in scaling. The easiest way to do this is to assign letter grades to each measure and determine the equivalent scores on each measure. If the bottom A, B, C, and D percentages on the test are the same as the bottom A, B, C, and D percentages on the project, then the difference in variability is due to a difference in achievement and no adjustment in the possible points is necessary and setting the possible points to be equal would result in equal weighting.

Rather than convert each of the scores it would be better to assign the possible points to the project in such a way that the difference in observed variability was equal to the difference in weighting intended. If the test and project were to be equally weighted, then their observed variabilities should be the same. If the A-F range was 50% on the test and 100% on the project, the project should have a possible points half as much (50/100) as the test. If the test had a possible of 50 points, the project should have a possible of 25 points. This would give equal raw score ranges (25-50 on the test and 0-25 on the project).

For a second example a test is to be constructed in two parts, with each part having a different type of item.

1. Write the items in the objective section.

For this example we will assume 40 items have been written that will be worth 40 points.

2. Estimate the range of scores from the bottom A to the bottom D using the scoring system for the objective section.

The bottom A will be 90% and the bottom D will be 60% giving a range from the bottom A to bottom D of 30%.

3. Determine the weight for the objective section of the test.

The weight for the objective section will be 50% of the test.

4. Write the subjective section of the test.

The subjective section will have 3 essay items.

5. Determine the weight for the subjective section of the test.

The weight for the subjective section will be 50% of the test.

6. Estimate the range of scores from the bottom A to the bottom D using the scoring system used for the subjective section.

   The bottom A will be 90% and the bottom D will be 30% giving a range from the bottom A to bottom D of 60%.

7. Set the possible points for the subjective section of the test.

   Since the scale variability of the subjective section is twice as much as the objective section (60%/30% range from the bottom A to the bottom D) the number of points needed for the subjective section which would result in equal observed variabilities would be half as many points. The subjective portion of the test would have 20 points for the three questions. The raw score range from the bottom A to the bottom D would be 12 points on the objective portion -- 36 (90%) to 24 (60%) and 12 points on the subjective portion -- 18 (90%) to 6 (30%).

The problem with this method is trying to convince the students that the 25 point project has the same weight as the 50 point test. The next method removes this problem.

Method Three: Modify Scoring Procedures

1. Select the difficulty level of the measures.

   Use the same procedures described for method one above.

2. Select the scoring procedure for each measure.

   This method has the teacher score each measure so that there is the same scale variability no matter which type of measure is used. The teacher would have to modify the scoring procedures for either the objective or subjective measures used in the class. All types of measures have the same percentage score for perfect performance -- 100%, so no modification of scoring is needed at the upper levels. Since students with no achievement usually get 0% on a subjective measure and a chance score on an objective measure (i.e., 25% for a 4-option multiple choice test) scoring must be modified to make the lower scores equal.

   The easiest way to change the scoring of objective items is to use a correction for guessing formula which will convert the scores to a range of 0% - 100%. This, however, takes a bit of time and mathematical manipulation. It is usually less work to change the scoring of the subjective items.

   Subjective items must be scored in such a way that the range of scores from perfect achievement to no achievement is the same as for the objective items with which the subjective items are to be combined. If the objective items are multiple choice with 4 options (a range from 25%-100%), the subjective items must be scored in such a way that a response that indicates no achievement is given a score of 25%. Since giving a score of 25% for a meaningless (or missing) response is not very satisfactory, a good way to accomplish this is to score the subjective items in the normal way and then convert the scores to a 25%-100% scale before entering them in the grade book. This could be done by converting the initial scores to letter grades (with +/- gradations) and then converting the letter grades to a score that is equivalent to the objective score of equal achievement on the 25%-100% scale.

   An alternative method suggested by Hopkins, Stanley, and Hopkins (1990) is to score projects and assignments that are evaluated subjectively by using a scale that will have the middle two-thirds of the scores within a range equal to double the standard deviation and total range equal to four to six standard deviations of the objective component with which it is to be combined.

The following example illustrates how scoring can be modified. This example is similar to the second one for the previous method.

1. Write the items in the objective section

   For this example we will assume 40 items have been written that will be worth 40 points.

2. Estimate the range of scores from the bottom A to bottom D using the scoring system for the objective section.

   The bottom A will be 90% and the bottom D will be 60% giving a range from the bottom A to bottom D of 30%.

3. Determine the weight for the objective section of the test.

   The weight for the objective section will be 33% of the test.

4. Write the subjective section of the test.

   The subjective section will have 3 essay items.

5. Determine the weight for the subjective section of the test.

   The weight for the subjective section will be 66% of the test.

6. Score the subjective items in such a way as to have the same scale variability as the objective items.

   Even though this would not be the natural way to score the subjective items, to use this method the bottom A will be 90% and the bottom D will be 60% giving a range from the bottom A to bottom D of 30%.

7. Set the possible points for the subjective section of the test

   Since the subjective part of the exam is to be twice as important as the objective portion, it should have twice as many possible points (80 points) since the scale variabilities are equal.

The major problem with this method is assigning a non-zero score to a measure that indicates no achievement. As long as the range of achievement is within a "normal" range, this problem is avoided. Since very few students have "no learning" this is not usually a problem.

Conclusions

When combining scores, consider the true variability and scale variability in addition to the observed variability. Consider the extent to which true achievement, difficulty level, and type of scale used affects the observed variability. Only use observed variability as the basis for weighting scores if differences in true achievement can be safely ignored. Otherwise, use one of the options presented in this paper: using a common scale for each measure or modifying the possible points or scoring procedures.

References

Gronlund, Norman E. and Linn, Robert L. (1990). Measurement and Evaluation in Teaching, 6th Edition. New York, NY: Macmillan.

Hopkins, Kenneth D., Stanley, Julian C., and Hopkins, B. R. (1990). Educational and Psychological Measurement and Evaluation, 7th Edition. Englewood Cliffs, NJ: Prentice Hall.

Oosterhof, Albert C. (1987). Obtaining Intended Weights When Combining Students' Scores. Educational Measurement: Issues and Practice, Winter, 1987, 29-37.