Evaluating Score Consistency through Score Change Probabilities

Kimberly A. Barchard
University of Nevada, Las Vegas

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Abstract

Test users often want to know how much scores would change if different measurement conditions were used: if the test was taken at a different time, contained different items, or was scored by someone else. If the student took the test on another day, what is the probability they would pass the course? If a different professor read the applications, what is the probability they would admit the same students to the graduate program? If a different clinician interviewed the client, what is the probability the client would receive the same diagnosis? This paper demonstrates a new method for calculating the probability of changes in observed test scores under slightly different measurement conditions. It also demonstrates the relationship of these new calculations to the Reliability Coefficient used in Classical Test Score theory and to the Concordance Correlation Coefficient developed by Lin (1989) to measure absolute consistency.

For interval or ratio level data, Score Change Probabilities are calculated in three steps. First, subtract the two sets of scores to obtain the Score Change values. Second create a frequency table to show how often each Score Change value occurs. Third, convert the frequencies into probabilities, by dividing by the total number of cases. Score Change Probabilities can be summarized using the Standard Change, which is the square root of the average squared Standard Change value. For ordinal level data, two analyses are possible: Rank Change Probabilities (and the Standard Rank Change), or the proportion agreement among the top ranked participants. Finally, for nominal level data, the probability of category changes can be calculated.

These calculations are demonstrated using data from the Levels of Emotional Awareness Scale (LEAS; Lane & Swartz, 1987). 18 research assistants rated 48 participants on the LEAS. Two pairs of RAs are examined in detail. One pair had high consistency (CCC = .98, Standard Change = 1.53). The other pair had acceptable inter-rater reliability (.79) but detailed examination revealed that these two RAs did not have adequate levels of agreement, even for research purposes (CCC = .72, Standard Change = 6.25). Thus, Score Change Probabilities provide information that complements the information given by existing measurement theories.

Introduction

Mary failed an essay exam. She says she knows the material and that she failed for three reasons. First, the questions weren't fair. If the exam had included different questions, she would have been fine. Second, the TA was too hard. If the instructor had done the grading, she would have passed. Finally, she couldn't sleep the night before because the apartment above her flooded at midnight. If the exam had been on another day, she would have been better rested and able to demonstrate what she knew. Is Mary right? How much would her exam grade have changed if the exam had different items, was marked by someone else, or was given on a different day? What is the chance that she would have passed the exam? This paper will demonstrate how to answer these questions.

Test takers (such as Mary) and test users (such as Mary's teacher) often want to know how much scores would change if slightly different measurement conditions were used: if the test contained different items, was scored by someone else, or was taken at a different time. Existing measurement theories (e.g., Classical Test Score Theory, Lord & Novick, 1968; Latent Trait Theory, Lord, 1980; and Generalizability Theory, Cronbach, Gleser, Nanda, & Rajaratnam, 1972) do not answer this question directly, and require a number of assumptions in order to estimate this answer indirectly. This paper develops a new method that answers this question directly, by calculating the probability that scores will increase or decrease by any given amount.

Score Change Probabilities

Imagine a class of students completes an essay exam. Grading the essay will be time consuming and so two raters (the instructor and the TA) plan to divide the grading. What effect will this have? What if one rater is "easier" than the other? To assess the comparability of scores that are assigned by two graders, both graders should provide scores for a relatively large number of students (for example, 50 students). We can then assess the comparability of these scores using traditional inter-rater reliability coefficients and Score Change Probabilities. If these calculations show that the two scorers are comparable, then we can justifiably divide the grading between the two people; otherwise one of them will have to grade every essay.

Table 1
Interval Data: Grade Change Probabilities

Grade Change	1: Instructor 2: TA	1: TA 2: Instructor	1: Random 2: The Other
-14	.00	.04	.02
-13	.00	.02	.01
-12	.00	.06	.03
-11	.00	.08	.04
-10	.00	.02	.01
-9	.00	.02	.01
-8	.00	.02	.01
-7	.00	.06	.03
-6	.00	.02	.01
-5	.00	.06	.03
-4	.00	.04	.02
-3	.00	.18	.09
-2	.00	.24	.12
-1	.00	.04	.02
0	.10	.10	.10
1	.04	.00	.02
2	.24	.00	.12
3	.18	.00	.09
4	.04	.00	.02
5	.06	.00	.03
6	.02	.00	.01
7	.06	.00	.03
8	.02	.00	.01
9	.02	.00	.01
10	.02	.00	.01
11	.08	.00	.04
12	.06	.00	.03
13	.02	.00	.01
14	.04	.00	.02

Note. Inter-rater reliability = .92. Standard Change = 6.55.

Table 2
Comparing Ranks Assigned by Two Evaluators

Applicant's Initials	Rank based on Evaluator A	Rank based on Evaluator B	Rank Change
HD	1	2	1
JT	2	3	1
AB	3	1	-2
CF	4	7	3
KA	5	4	-1
OL	6	5	-1
PS	7	6	-1
RF	8	9	1
WB	9	10	1
PB	10	8	-2

To show that Score Change Probabilities complement the information given by traditional reliability coefficients, we will consider an example in which the correlation between the scores assigned by the TA and the Instructor is .92. See the Score Change Probabilities given in Table 1. The second column of the table shows that the grade may change if the Instructor was the first grader and now the TA is going to mark an essay. For example, there is a 10% chance that the grade will stay the same, and a 4% chance that it will increase by 1 point. The third column shows the opposite probabilities: the probabilities that the grade will go up or down if the first grader was the TA and now the Instructor will mark the essay. Finally, the fourth column shows what happens if the Instructor and TA each grade half the essays first, where the division of essays between the Instructor and TA is random: The grade has an exactly equal chance of increasing or decreasing when the second person marks the essay. The probabilities in the last column are simply the average of the probabilities in the second column and the probabilities in the third column. In this example, there is a 22% chance that the grade will change by 10 or more points. To make this example more concrete, consider the grading scheme that I use in my undergraduate classes: A grade of 60 is D-, 70 is a C-, 80 is B-, and 90 is A-. A grade change of 10 points means that a 71 C- changes to 81 B- or 61 D-. This level of consistency would be considered unacceptable by most instructors. However, the inter-rater reliability coefficient was .92. This example demonstrates that the inter-rater reliability coefficient by itself does not fully describe the consistency (or inconsistency) of scores.

Interval or Ratio Data

For interval or ratio level data, calculating basic Score Change Probabilities requires four steps. First, obtain two sets of scores that you want to compare. For example, these might be scores from the same research participants at two testing times, or from two forms of a test that contain different items, or from two raters. Second, for each participant, calculate the difference between the two scores. These differences are called Score Change values. Third, create a frequency table that shows how often each Score Change value appears. Finally, convert the frequencies into probabilities, by dividing by the total number of participants. If there are a large number of possible Score Change values, it may be helpful to group the Score Change values into intervals for presentation purposes, rather than presenting the probabilities for each possible Score Change value, or to show the probabilities on a graph.

Sometimes we transform scores before we use them. For example, we might convert them into T-scores or CEEB scores (see Murphy & Davidshofer, 2005). To calculate Score Change Probabilities for transformed scores, start by calculating the transformed scores. Then calculate the Score Change Probabilities using the same steps as you would use for raw scores.

Ordinal Data

Table 3
Ordinal Data: Proportion Agreement regarding which Students to Admit

Number admitted	Proportion agreement (fraction)		(decimal)
1	0 / 1	=	0.00
2	1 / 2	=	0.50
3	3 / 3	=	1.00
4	3 / 4	=	0.75
5	4 / 5	=	0.80
6	5 / 6	=	0.83
7	7 / 7	=	1.00
8	7 / 8	=	0.875
9	9 / 10	=	0.90
10	10 / 10	=	1.00

Data can be measured at the ordinal level or can be transformed to the ordinal level for decision making. One common ordinal level measure is an agreement scale where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. Even if data were not measured at the ordinal level, psychologists often convert data to the ordinal level for decision making purposes. For example, if 10 students applied to a graduate program, faculty might be rank order the applicants to determine who to admit. When we use ranked data such as this, we might wonder if the rank order of the applicants would be the same if someone else evaluated the files. We can determine this directly: we can have another person evaluate the materials and rank order the applicants. Then we can calculate the change in the rank orders for each applicant (see Table 2).

At this point, we have two options for how to present and interpret the results. We could calculate Rank Change Probabilities, using the same procedure as we used for interval level data. Alternatively, we could calculate the proportion of agreement regarding the top ranking individuals – the people who will be admitted to the program. The proportion agreement can be calculated for each possible number of admitted students (see Table 3). This table shows that the evaluation procedures are relatively consistent if the program was going to admit three or more students. However, if the program was going to admit just one or two students, rank orders do not have adequate consistency: Differences in the rank orders assigned by the two evaluators would change who is admitted to the program. Therefore, the evaluators may want to change their evaluation procedures to create greater score consistency for top ranked applicants.

Nominal Data

Table 4
Category Change Frequencies (and Probabilities)

First Rater	Second Rater: Antisocial	Bipolar	Borderline	Dependent	Passive-aggressive	Total
Antisocial	16 (.11)	1 (.01)	6 (.04)	1 (.01)	3 (.02)	27 (.18)
Bipolar	3 (.02)	23 (.15)	1 (.01)	2 (.01)	0 (.00)	29 (.19)
Borderline	5 (.03)	1 (.01)	18 (.12)	0 (.00)	3 (.02)	27 (.18)
Dependent	1 (.01)	0 (.00)	1 (.01)	28 (.19)	3 (.02)	33 (.22)
Passive-aggressive	5 (.03)	1 (.01)	2 (.01)	0 (.00)	26 (.17)	34 (.23)
Total	30 (.20)	26 (.17)	28 (.19)	31 (.21)	35 (.23)	150 (1.00)

Note. Proportion agreement = .74. Kappa = .67.

For nominal level data, numbers are used like names for individuals or groups. Therefore, it would be inappropriate to subtract the scores to examine Score Change values and it would be inappropriate to order the scores to determine how the rank orders change. However, we can still examine how scores change from one measurement to the next – how often the first rater assigned a participant to one category and the second rater assigned that participant to a different category.

Consider an example in which 150 clients are categorized as having one of five personality disorders: antisocial, bipolar, borderline, dependent, or passive-aggressive. Table 4 shows the number of people who were placed in each category by each rater. The categories given by the first rater are listed on the left hand side, and the categories given by the second rater are given across the top. For example, the first rater categorized 27 people as antisocial. Of those 27 people, 16 were categorized as antisocial by the second rater. These frequencies can be converted into probabilities by dividing by the total number of clients. For example, this table shows that 16 / 150 = .11, so that 11% of the clients were categorized as antisocial by both of the raters.

In this example, the two raters agree most of the time for each diagnosis. However, the two raters sometimes disagree on whether someone is antisocial, borderline, or passive-aggressive. For example, the first row shows that 6 people who were categorized as antisocial by the first rater were categorized as borderline by the second rater. Frequencies do not tell us *why* there are inconsistencies between the two raters, but discovering *where* there are inconsistencies is the first step in correcting them.

Standard Change

The Standard Change tells us how much the scores typically change from one measurement to the next. It can be used to summarize the information in a Score Change Probability Table for interval or ratio level data. For example, in Table 1, the Standard Change was 6.55. The Standard Change is calculated as the square root of the average squared Score Change (see the Appendix). The Standard Change should not be confused with the Standard Deviation of the Score Change values. The Standard Deviation subtracts the mean Score Change from the numerator, but the Standard Change does not. If the two measures have different means this will influence the Standard Change, but will have no effect on the Standard Deviation of the Score Change values.

$$SC = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}} \qquad \sigma_{X-Y} = \sqrt{\frac{\sum_{i=1}^{n}((X_i - Y_i) - (\mu_X - \mu_Y))^2}{n}}$$

Other Measures of Score Consistency

The reliability coefficient is usually calculated using the Pearson Product-Moment Correlation. It measures the degree of linear association between two sets of scores. If the first measure is designated as X and the second is designated as Y, it measures the extent to which Y = aX + b. The correlation coefficient is used as a measure of reliability based upon the assumption that the means and variances of X and Y are identical. If two measures are parallel, then Lin (1989) showed that the Pearson Product-Moment Correlation equals the Concordance Correlation Coefficient, which is discussed below. If two measures are not parallel, then the correlation between them tends to underestimates reliability (Alwin, 2007).

Table 5
Coefficients of Consistency for the Example Data in Table 1

Coefficient	Example Data
Means	60.00; 65.02
Variances	35.20; 80.70
Covariance	49.14
Correlation	.92
ICC(C,1)	.85
ICC(A,1)	.70
CCC	.70
Standard Change	6.55

Intra Class Correlations (ICC; Shrout & Fleiss, 1979) are also used to assess score consistency. ICC(C,1) measures the degree of additivity between two measures: the extent to which Y = X + b. It assumes that the variances of the two measures are identical. The symbol ICC(C,1) indicates that it measures the consistency of 1 measure to another. ICC(A,1) measures the degree of absolute agreement between the two measures: the extent to which Y = X. The symbol ICC(A,1) indicates that it measures the absolute agreement between 1 measure and another. This formula also assumes that the variances are equal to each other.

$$CCC = \frac{2\sigma_{XY}}{(SC)^2 + 2\sigma_{XY}}$$

$$(SC)^2 = 2\sigma_{XY}\left(\frac{1 - CCC}{CCC}\right)$$

The Concordance Correlation Coefficient (CCC; Lin, 1989) is another measure of the extent to which Y = X. The advantage of CCC over ICC(A,1) is that CCC does not assume equal variances. The CCC is a well-accepted measure of score consistency in biology and medicine. Some papers (e.g., Barnhart, Lokhnygina, Kosinski, & Haber, 2007) state that it is the most common measure of agreement for continuous data. However, the CCC is virtually unheard of in psychology.

Each of these indices of score consistency has been calculated for the example data in Table 1. See Table 5. The correlation is high in this example, because it is not influenced by differences in means and variances. ICC(C,1) is somewhat lower, because it is influenced by the difference in means but ignores the difference in variances. ICC(A,1) and CCC are quite a bit lower, because they are influenced by the differences in means and variances. Thus, in a particular measurement situation, if differences in means or standard deviations would be interpreted as evidence that the scores are not consistent with each other, the correlation should not be used. One of these other measures of score consistency should be used instead.

Evaluating Score Consistency for LEAS Scoring

To demonstrate the use of Score Change Probabilities with real research data, I will evaluate score consistency for the Levels of Emotional Awareness Scale (LEAS; Lane & Swartz, 1987). The LEAS is the most commonly used measure of the depth and complexity of knowledge of emotion words. It contains 20 open-ended questions, which are subjectively scored based upon the rules in the LEAS Scoring Manual (Lane, 1991). Each item is assigned a score from 0 to 5, so that total scores range from 0 to 100. Because scoring the LEAS by hand is time-consuming, I want to divide the scoring between several raters and then ignore who did the ratings when I analyze my data. I am therefore interested in an absolute definition of agreement.

To examine consistency, I asked 48 undergraduate students to complete the LEAS. Eighteen research assistants scored the data after five intensive weeks of training that included 360 practice examples. On average, consistency between the raters was high (average inter-rater reliability = .94). However, some pairs of raters had higher levels of consistency than others. I will focus on one pair of raters who had high consistency and one pair of raters who had low consistency.

First I will examine score consistency for a pair of scorers who were very consistent with each other. Scorers 8 and 12 had very high inter-rater reliability (r = .99). However, we can obtain a more complete picture of how the two sets of scores compare by examining Score Change Probabilities, the Standard Change, and the CCC. The second column of Table 6 shows that Scorer 12 often gave a score that was one point lower than Scorer 8 (this happened 33% of the time). However, it was rare for scores to be 4 points apart (this happened only 2% of the time) and scores never differed by more than 4 points. The Standard Change was 1.53, and CCC was .98.

Table 6
LEAS Score Change Probabilities for Scorers 8 and 12

Score Change	First Scorer: 8 Second Scorer: 12	First Scorer: 12 Second Scorer: 8	First Scorer: Random Second Scorer: Other
-4	.02	.00	.01
-3	.08	.00	.04
-2	.15	.06	.10
-1	.33	.10	.22
0	.25	.25	.25
1	.10	.33	.22
2	.06	.15	.10
3	.00	.08	.04
4	.00	.02	.01

Note. Inter-rater reliability = .99. Standard Change = 1.53. CCC = .98.

Next I examined a pair of scorers who had a lower level of consistency: Scorers 3 and 9. First I calculated inter-rater reliability, which was adequate (r = .79). Second, I examined the Score Change Probabilities. Table 7 reveals that they rarely gave identical scores (this happened only 6% of the time). Scorer 9 tended to give higher scores than Scorer 3 (score changes of 6 or more were quite common). However, Scorer 9 sometimes gave scores that were lower (in one case, 15 points lower). The Standard Change was 6.25 and CCC was .72.

Table 7
LEAS Score Change Probabilities for Scorers 3 and 9

Score Change	First Scorer: 3 Second Scorer: 9	First Scorer: 9 Second Scorer: 3	First Scorer: Random Second Scorer: Other
-15	.02	.00	.01
-14	.00	.00	.00
-13	.00	.00	.00
-12	.02	.06	.04
-11	.00	.02	.01
-10	.00	.06	.03
-9	.00	.00	.00
-8	.00	.02	.01
-7	.00	.04	.02
-6	.00	.13	.06
-5	.02	.06	.04
-4	.00	.10	.05
-3	.02	.13	.07
-2	.06	.08	.07
-1	.02	.06	.04
0	.06	.06	.06
1	.06	.02	.04
2	.08	.06	.07
3	.13	.02	.07
4	.10	.00	.05
5	.06	.02	.04
6	.13	.00	.06
7	.04	.00	.02
8	.02	.00	.01
9	.00	.00	.00
10	.06	.00	.03
11	.02	.00	.01
12	.06	.02	.04
13	.00	.00	.00
14	.00	.00	.00
15	.00	.02	.01

Note. Inter-rater reliability = .79. Standard Change = 6.25. CCC = .72.

Examination of the means and variances is revealing. These two scorers have different means (Scorer 3 mean = 63.69; Scorer 9 mean = 67.15, $t(47) = 4.21$, $p < .001$). In absolute terms, the difference in the scores is large (3.27) and the effect size for the difference is also moderately large (Cohen's d = .61). Also, these two scorers have significantly different variances. Wilcox (1990) evaluated several methods of comparing dependent variances and recommended the Spear procedure developed by P. McCollough (1987). This test shows that Scorer 3 has a significantly larger variance (Scorer 3 variance = 75.43, Scorer 9 variance = 57.28, (46) = -.18, p = .22). I conclude that there is a moderately large difference in the means and a small difference in variances, which impact the overall score consistency.

Usually, an inter-rater reliability correlation of .79 would be considered adequate for research purposes. However, this more detailed comparison of the two sets of scores has clearly shown that these scores do not have adequate levels of consistency for us to ignore who did the scoring, even for research purposes.

These examples demonstrate that Score Change Probabilities, the Standard Change, and the Concordance Correlation Coefficient provide information that is not given in the inter-rater correlation. Calculation of these statistics can be valuable for applied users who want to know how much observed scores are likely to change from one measurement to the next, and to theoretical psychometricians who want to determine the extent to which two sets of scores are equal to each other.

Appendix: Calculating the Standard Change in SPSS

To calculate the Standard Change in SPSS 15, 16, or 17, use the following steps. First, calculate the Score Change Values. Click on the **Transform** menu and select **Compute Variable** from the drop-down menu. In the **Target Variable** box, type **ScoreChange**. In the **Numeric Expression** box, calculate the difference between the two variables. For example, if the two variables were called Score1 and Score2, then the Numeric Expression box would say **Score2-Score1**. Click **OK**. Second, to square these numbers, click on **Transform** / **Compute Variable**. Set the **Target Variable** to **SquaredChange**. In the **Numeric Expression** box, calculate the **ScoreChange** value multiplied by itself: **ScoreChange*ScoreChange**. Click **OK**. Third, to calculate the average of the SquaredChange values, click on the **Analyze** menu, and select **Descriptive Statistics** from the drop-down menu. Click on **Descriptives** from the side-menu. Move **SquaredChange** to the **Variables(s)** box. Click **OK**. Look at the output window to find the **mean** value for SquaredChange. Fourth, use a calculator to obtain the square root: This square root is the Standard Change. To calculate the Standard Change in later versions of SPSS or in other statistical packages, follow the basic idea of calculating the Score Change values, squaring them, averaging the squared values, and then taking the square root.

Presented at the April 2010 Western Psychological Association conference, Cancun, Mexico.