Evaluating Score Consistency through Score Change Probabilities

Kimberly A. Barchard
University of Nevada, Las Vegas

Reference: Barchard, K.A. (May, 2009). *Evaluating Score Consistency through Score Change Probabilities.* Paper presented at the Association for Psychological Science convention in San Francisco, CA.

Contact Information: Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, 4505 Maryland Parkway, P.O. Box 455030, Las Vegas, NV, 89154-5030, USA, barchard@unlv.nevada.edu

Abstract

This paper demonstrates a new method for calculating the probability that observed test scores would change by a certain amount under slightly different measurement conditions. Score Change Probabilities provide information that complements and goes beyond traditional reliability coefficients and can be used with virtually any type of measurement data.

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 other TA 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 a different TA, 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, Latent Trait Theory, Generalizability Theory) do not answer this question directly, and require a number of assumptions in order to estimate this answer indirectly[1]. 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.

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 Changes. Third, create a frequency table which shows how often each Score Change value appeared. Finally, convert the frequencies into probabilities, by dividing by the total number of participants.

In this paper, I will use an extended example in which we compare the scores that were assigned by two raters. However, the logic and calculations of Score Change Probabilities can be easily applied to times or items. Of course, the interpretation of Score Change values will depend upon the ways in which the two measurements differ. If the two measurements consist of the same items, which were scored by the same person, but which were administered at two times, then there are three possible causes of changes in the score: change in the person over time (e.g., the person is more conscientious now that they are older), the interaction of the person with time (e.g., the person is more conscientious this year, because of worries about job loss in the poor economy), and random error (e.g., the person feels happy today and therefore is agreeing with most items). If the two measurements consist of the same items, which were administered once, but which were scored by two different people, changes in the score would be attributed to different factors. The purpose of calculating Score Change values is not to create a mathematical model of *why* scores change. Instead, the purpose is to calculate precisely and directly *how* the scores change. These calculations are equally applicable if the scores differ in terms of the time, the rater, or even the items that test takers completed.

Score Change Probabilities

Imagine a class of students completes an essay. 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, we will have both graders 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.

To show that Score Change Probabilities complement the information given by traditional reliability coefficients, we will consider three examples. In each, the correlation between the scores assigned by the TA and the Instructor is .92. This would usually be considered a high level of inter-rater reliability. As we shall see, however, high inter-rater reliability does not guarantee high levels of score consistency.

In the first example, the TA is a slightly easier grader. This can be seen from the Score Change Probabilities given in Table 1. The second column of the table shows that the grade will likely go up, if the Instructor was the first grader and now the TA is going to mark an essay. For example, there is a 26% chance that the grade will increase by 1 point, and an 8% chance that it will increase by 2 points. Conversely, the third column shows that the grade is likely to go 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 essays are divided between the two graders and which person marks each essay is random: the grade has an equal chance of increasing or decreasing. 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 first example, the TA and Instructor are relatively consistent but the TA is a slightly easier grader. 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-. The fourth column of Table 1 shows that there is a pretty good chance (68%) that the grade would remain the same or differ by 1 point (e.g., 71 C- increases to 72 C- or decreases 70 C-) if the other person did the grading. There is a moderate chance (8%) that the grade will change by 3 or 4 points (e.g., 71 C- increases to 75 C or decreases 67 D+). There is virtually no chance (in this data set, we estimated it as a 0% chance) that the other person would grade 7 points higher or lower (e.g., 71 C- increases to 79 C+ or decreases to 64 D). This level of consistency might be considered relatively good.

In the second example, the TA and Instructor are less consistent. As shown in Table 2, there is only a small chance (16%) that the grade would remain the same or differ by 1 point. There is also some chance (8%) that the other person would grade 7 or more points higher or lower (e.g., 71 C- changes to 78 C+ or 64 D). This level of consistency might be considered acceptable but not ideal. This second example has the same level of inter-rater reliability as the first example: the correlation between the first and second raters is .92. This inter-rater reliability coefficient does not reflect the lower agreement between the two raters.

In the third example, the TA is a much easier grader. As shown in Table 3, there is now a 22% chance that the grades will change by 10 points or more (e.g., e.g., 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 is still .92. These examples demonstrate that the inter-rater reliability coefficient by itself does not fully describe the consistency (or inconsistency) of scores.

Table 1
Example 1 Grade Change Probabilities

Grade Change	1: Instructor 2: TA	1: TA 2: Instructor	1: Random 2: The Other
-6	.04	.02	.03
-5	.04	.02	.03
-4	.04	.02	.03
-3	.00	.02	.01
-2	.04	.08	.06
-1	.12	.26	.19
0	.30	.30	.30
1	.26	.12	.19
2	.08	.04	.06
3	.02	.00	.01
4	.02	.04	.03
5	.02	.04	.03
6	.02	.04	.03

Note. Inter-rater reliability = .92. Standard Change = 2.38. The calculation of Standard Change is explained below.

Table 2
Example 2 Grade Change Probabilities

Grade Change	1: Instructor 2: TA	1: TA 2: Instructor	1: Random 2: The Other
-9	.00	.04	.02
-8	.00	.02	.01
-7	.00	.02	.01
-6	.00	.04	.02
-5	.00	.12	.06
-4	.02	.04	.03
-3	.02	.12	.07
-2	.06	.34	.20
-1	.02	.12	.07
0	.02	.02	.02
1	.12	.02	.07
2	.34	.06	.20
3	.12	.02	.07
4	.04	.02	.03
5	.12	.00	.06
6	.04	.00	.02
7	.02	.00	.01
8	.02	.00	.01
9	.04	.00	.02

Note. Inter-rater reliability = .92. Standard Change = 3.73.

Table 3

Example 3 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.

Score Change Probabilities for Different Types of Data

Score Change Probabilities can be calculated for virtually any type of measurement data. They can be calculated based upon raw scores, as described above, or transformed scores (z-scores, T-scores, Stanines, etc). They can be calculated based upon interval or ratio level data, as described above, but they can also be calculated for ordinal and nominal level data. In this section, I will describe the calculation of Score Change Probabilities for each of these data types.

Measurement data is usually divided into four levels: nominal, ordinal, interval and ratio. In the nominal level of measurement, numbers are used to name people (e.g., driver's license numbers) or categories (e.g., diagnoses). In the ordinal level of measurement, numbers are used to represent the rank order of the individuals. In the interval level of measurement, the differences between the numbers represent the differences between the people. Finally, in the ratio level of measurement, the ratios of the numbers reflect the relationships between the people and a score of 0 indicates that the person has none of the characteristic being measured.

When data are interval level, the differences between scores are meaningful. This allows us to justify adding and subtracting scores, so that we can calculate the mean and use linear statistical models such as ANOVA, correlation, and factor analysis. This also allows us to calculate reliability coefficients using the Pearson product-moment correlation (Lord & Novick, 1968) as we did above. When data are instead at the ordinal or nominal level, we use different statistics to assess reliability. At the ordinal level, we might examine the reliability of single scores using Spearman's rank order correlation or Kendall's coefficient of concordance (Kendall, 1948), and we might examine the reliability of composite scores using ordinal coefficients alpha or theta (Zumbo, Gadermann, & Zeisser, 2007). At the nominal level, Kappa (Cohen, 1960) is often used to estimate reliability. Thus, the level of measurement influences how we calculate reliability coefficients. The level of measurement will also influence how we calculate Score Change Probabilities.

Most psychological data are technically ordinal level. We are confident that higher scores represent more of the construct, but we cannot mathematically justify the claim that differences between numbers mean the same thing at different points of the scale. In these situations, researchers can treat the data as ordinal or they can argue that the data is very close to being interval. For example, consider an agreement scale where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. Often, researchers will claim that such scores are very close to being evening spaced and thus are very close to being interval level. They will then proceed to calculate reliability coefficients that are based upon the assumption of interval level data. Similarly, researchers who are interested in Score Change Probabilities might treat such data as ordinal or as interval, depending upon the research questions they are interested in. I will begin by describing Score Change Probabilities for interval and ratio level data, and then discuss Score Change Probabilities for ordinal and nominal data.

Interval or Ratio Level of Measurement

For interval or ratio level data, we can either work directly with the raw scores or we can transform the raw scores.

Raw Scores

When we are working with raw scores for interval or ratio level data, calculating basic Score Change Probabilities requires four steps, as described above. First, obtain two sets of scores that you want to compare. Second, calculate Score Change values for each participant. Third, create a frequency table for the Score Change values. Finally, calculate Score Change 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. For detailed instructions on how to do these calculations using SPSS, see the Appendix.

Transformed Scores

Sometimes we transform scores before we use them. For example, we sometimes know that scores from different raters or different forms of a test will have different means and standard deviations. We might want to eliminate those differences by transforming the data. One common transformation is to mean-deviate the scores. To do this, the researcher subtracts the mean from each set of scores, so that each set of scores now has a mean of zero. Another common transformation is to standardize the scores. To do this, the researcher converts each score to a z-score, by subtracting the mean and dividing by the standard deviation. Each set of scores then has a mean of zero and a standard deviation of one. Sometimes, we take the transformation one step further to give the data a new mean and standard deviation. For example, with IQ scores, we multiply the z-score by 15 and add 100, so that the new scores have a mean of 100 and a standard deviation of 15. This type of transformation makes the scores easier to interpret than z-scores, because all scores are positive whole numbers. Other common transformations include T-scores, CEEB scores, Stanines, and percentiles (see Murphy and Davidshofer (2005) for a discussion of these transformations).

To calculate Score Change Probabilities for transformed scores, start by calculating the transformed scores. Then calculate the Score Change Probabilities using the same four steps as you would use for raw scores.

Ordinal Level of Measurement

Data can be measured at the ordinal level of measurement or can be transformed to the ordinal level of measurement for decision making. A variety of data is measured at the ordinal level using ranks. For example, psychologists might measure rank order in a competition such as a science fair or a poster session. In addition, psychologists often convert data to the ordinal level for decision making purposes.

Consider the example below in which 10 students have applied to a graduate program (see Table 4). For each applicant, imagine that an evaluator rated the letters of recommendation and the statement of purpose (see columns 2 and 3, below) and calculated the combined GPA across all institutions attended (see column 4). The evaluator then calculated a linear combination of the scores by transforming each raw score to a z-score (see columns 5, 6, and 7) and then averaging them (see column 8). The evaluator then used this linear combination to calculate rank orders for the 10 applicants (see column 9). These rank orders could be used to determine which applicants to admit and which to fund.

Table 4
Transforming Data to the Ordinal Level

Applicant's initials	Letters of Recommendation	Statement of Purpose	GPA	z-score Recommendation	z-score Purpose	z-score GPA	Average z-score	Rank
HD	3.90	3.90	3.93	1.15	0.84	0.86	0.95	1
JT	3.75	4.00	3.94	0.86	1.03	0.89	0.93	2
AB	3.50	3.60	4.00	0.37	0.27	1.07	0.57	3
CF	3.20	4.00	3.84	-0.21	1.03	0.59	0.47	4
KA	3.90	3.25	3.55	1.15	-0.39	-0.29	0.16	5
OL	2.90	3.80	3.67	-0.80	0.65	0.07	-0.03	6
PS	3.50	3.20	3.42	0.37	-0.48	-0.68	-0.26	7
RF	3.20	2.70	3.84	-0.21	-1.42	0.59	-0.35	8
WB	2.25	3.60	3.26	-2.06	0.27	-1.17	-0.99	9
PB	3.00	2.50	3.01	-0.60	-1.80	-1.93	-1.44	10
Mean	3.31	3.46	3.65					
Standard Deviation	0.51	0.53	0.33					

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 5).

Table 5

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

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. This tells us the probabilities that ranks will increase or decrease if we use Evaluator B rather than Evaluator A (or vice versa). In this data set, rank orders were likely to change quite a bit (see Table 6).

Table 6

Rank Change Probabilities

Rank Change	1: Evaluator A 2:Evaluator B	1: Evaluator B 2: Evaluator A	1: Random 2: The Other
-3	.00	.10	.05
-2	.20	.00	.10
-1	.30	.40	.35
0	.00	.00	.00
1	.40	.30	.35
2	.00	.20	.10
3	.10	.00	.05

Note. Inter-rater reliability = .85. Standard Change = 1.52.

When data is measured at the ordinal level of measurement, the differences in the numbers do not represent the same differences for low, medium, and high ranking people. Therefore, it may or may not be useful to calculate how much the ranked data changes. There is another way this data could be analyzed. We can focus specifically on the information that is most desired. In this case, it might be that the information that is most desired is the proportion of agreement regarding the top ranking individuals – the people who will be admitted to the program.

In the above dataset, if the program were to admit just one applicant, there is 0 agreement between the two evaluators in terms of who that top applicant is. Evaluator A gave HD the highest ratings, whereas Evaluator B gave AB the highest ratings. However, if the program were to admit the top 2 applicants, then the two evaluators agree on 1 of those 2 individuals: both would recommend admitting JT, but Evaluator A would also recommend admitting HD and Evaluator B would recommend admitting AB. If the program were to admit the top 3 applicants, then the two evaluators agree completely on who those students are.

The proportion agreement can be calculated for each possible number of admitted students (see Table 7). 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, they may want to change the evaluation procedures to create greater score consistency for top ranked applicants.

Table 7

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

Nominal Level of Measurement

When we measure data at the nominal level of measurement, the numbers are used like names for the individuals or groups. Therefore, it would be inappropriate to subtract the scores to examine Score Change values and it would be inappropriate to order numbers to determine how the rank orders change. However, we can still examine how scores change from one measurement to the next. The first rater may have assigned a participant to one category and the next rater may have assigned a 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 8 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. Table 9 converts these frequencies 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.

These two tables show that 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, 5 people who were classified as passive-aggressive by Rater 1 were classified as antisocial by Rater 2. There are three possible causes of these disagreements. First, the criteria for each diagnosis may be unclear. Second, the information that was available to the two raters might have been different (for example, perhaps they completed separate interviews with the clients and obtained different information). Third, the raters may have had inadequate training in how to complete the diagnoses. Category Change Frequencies do not tell the researcher why there are inconsistencies between the two raters, but discovering where there are inconsistencies is the first step in correcting them.

Table 8
Category Change Frequencies

	Second Rater					
First Rater	Antisocial	Bipolar	Borderline	Dependent	Passive-aggressive	Total
Antisocial	16	1	6	1	3	27
Bipolar	3	23	1	2	0	29
Borderline	5	1	18	0	3	27
Dependent	1	0	1	28	3	33
Passive-aggressive	5	1	2	0	26	34
Total	30	26	28	31	35	150

Table 9
Category Change Probabilities

	Second Rater				
First Rater	Antisocial	Bipolar	Borderline	Dependent	Passive-aggressive
Antisocial	.11	.01	.04	.01	.02
Bipolar	.02	.15	.01	.01	.00
Borderline	.03	.01	.12	.00	.02
Dependent	.01	.00	.01	.19	.02
Passive-aggressive	.03	.01	.01	.00	.17

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

Category Change Probabilities provide more information than traditional reliability statistics such as proportion agreement and Kappa. Proportion agreement is calculated as the proportion of participants for whom the two raters gave identical scores. Kappa (Cohen, 1960) is a similar statistic that corrects for agreement that would occur by chance when there are a limited number of options to choose between. Although these statistics give an overall sense of the consistency of the scores, they do not provide information about the specific location of the disagreements, which are clearly indicated by the Category Change Probabilities.

Standard Change

In some cases, test users may want to summarize Score Change Probabilities using a single number. One summary score that could be used with interval or ratio level data is the Standard Change. The Standard Change tells us how much the scores typically change from one measurement to the next. The three examples given at the beginning of this paper demonstrate the usefulness of the Standard Change. In these examples, two raters had high levels of inter-rater reliability (r = .92). However, in the first example they had relatively high consistency; in the second example, they had moderate consistency; and in the last example, they had low consistency. See Tables 1, 2, and 3 for the Score Change Probabilities. The decreasing level of agreement across these three examples can also be seen in the Standard Change values. These are given at the bottom of each table. In Table 1, the Standard Change was 2.38. In Table 2, it was 3.73. In Table 3, it was 6.55. Standard Change can therefore summarize the degree of inconsistency between two sets of scores, and it is easier to read than the entire Score Change Probability Table. For completeness, the researcher should always calculate the full set of Score Change Probabilities. However, they may not need to report the full set of probabilities. In some cases the researcher may be able to convey their conclusion about score consistency with just the Standard Change value and one or two other summary comments.

The Standard Change summarizes how much scores tend to change, on average. It is calculated as the square root of the average squared Score Change. The formula is

$$\hat{\Delta} = \sqrt{\frac{\sum_{i=1}^{n}(X_{ia} - X_{ib})^2}{n}}$$

where X_{ia} is the observed score of person i from measurement a,

X_{ib} is the observed score of person i from measurement b, and

n is the number of people for whom both measurements are available.

To calculate the Standard Change, there are four steps, which are demonstrated in Table 10 below. First, calculate how much the scores change from the first measurement to the second for each participant in the study. In the example, these numbers are given in the second-to-last column. You might already have calculated these Score Change values in order to calculate the Score Change Probabilities. Second, square these numbers. In the example, these numbers are given in the last column. Third, calculate the mean of the squared Score Change values. For the data given in Table 10, this is (1 + 4 + 16)/3 = 21 / 3 = 7. Finally, calculate the square root. This number is the Standard Change. For the data in Table 10, the Standard Change is the square root of 7, which is 2.65. It could be reported at the bottom of a Score Change Probability Table or included in the text. The population value Δ is calculated identically to the sample value, but is summed over all people in the population. For detailed instructions on how to do these calculations using SPSS, see the Appendix.

Table 10
Example Calculation of Standard Change

Test Taker	Measurement A	Measurement B	Score Change	Squared Score Change
A	10	11	1	1
B	8	10	2	4
C	11	7	-4	16

The minimum value of the Standard Change is 0. This value occurs when every number in the first set of data is identically equal to the corresponding number in the second set of data. The maximum value of the Standard Change is the difference between the highest and lowest possible scores. This value occurs when the two sets of numbers are extreme opposites: when one set of data has the smallest possible value, the other set of numbers has the largest possible value.

The Standard Change should not be confused with the Standard Deviation of the Score Change values. The sample Standard Deviation of the Score Change values would be calculated as

$$s_{X_{ia} - X_{ib}} = \sqrt{\frac{\sum_{i=1}^{n}((X_{ia} - X_{ib}) - (\overline{X}_a - \overline{X}_b))^2}{n-1}}$$

where $\overline{X}_a$ is the mean of the observed scores from measurement a, across all participants, and

$\overline{X}_b$ is the mean of the observed scores from measurement b, across all participants,

so that $(\overline{X}_a - \overline{X}_b)$ is the difference in mean scores and is therefore the mean of the Score Change values.

The Standard Deviation subtracts the mean Score Change in the numerator, but the Standard Change does not. If the two measures have different means this will increase the Standard Change, but will have no effect on the Standard Deviation of the Score Change values.

Other Measures of Score Consistency

Statisticians have developed several measures of score consistency beyond the traditional reliability coefficient. In this section, I will compare four of these to each other and to the Standard Change. First, the Pearson Product Moment Correlation measures the degrees of linear association between the two measures. 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. Second, ICC(C,1) measures the degree of additivity between the two measures. It measures the extent to which Y = X + b. It assumes that the variances of the two measures are identical. The symbol ICC(C,1) denotes that this is an Intraclass Correlation Coefficient that measures the consistency of 1 measure to another. Third, ICC(A,1) measures the degree of absolute agreement between the two measures, the extent to which Y = X. The symbol ICC(A,1) denotes that this is an Intraclass Correlation Coefficient of absolute agreement between 1 measure and another. This formula also assumes that the variances are equal to each other. Fourth, the Concordance Correlation Coefficient (CCC; Lin, 1989) is a second measure of the extent to which

Y = X. However, CCC does not assume equal variances. Finally, Standard Change is a third measure of the extent to which Y = X, and it does not assume equal variances. It has the advantage of being expressed in the metric of the original scores, and thus is immediately useful to test users.[1]

The CCC is a well-accepted measure of score consistency in biology and medicine. Some papers state that it is the standard measure of agreement in these disciplines. However, the CCC is virtually unheard of in psychology. Neither Lin's (1989) paper not the Concordance Correlation Coefficient is mentioned in *Psychometrika*, *Psychological Methods*, or *Educational and Psychological Measurement.*

Each of these indices of score consistency has been calculated for the data in Examples 1, 2, and 3. Table 11 shows that ICC(C,1), ICC(A,1), and Standard Change are influenced by differences in means and variances, but the Correlation is impervious to these differences in the scores. 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.

The Standard Change behaves similarly to ICC(A,1) and CCC. This is because all three use an identity definition of agreement: If X and Y differ in any way, this is considered inconsistent.

Table 11

Coefficients of Consistency for the Data in Examples 1, 2, and 3

Coefficient	Example 1	Example 2	Example 3
Means	60.00; 59.98	60.00; 62.54	60.00; 65.02
Variances	35.20; 35.66	35.20; 47.73	35.20; 80.70
Correlation	.92	.92	.92
Covariance	32.60	37.75	49.14
ICC(C,1)	.92	.91	.85
ICC(A,1)	.92	.85	.70
CCC	.92	.84	.70
Standard Change	2.38	3.73	6.55

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. In this study, 48 undergraduate students completed 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.

Table 12

LEAS Score Change Probabilities for Scorers 8 and 12

Score Change	First: Scorer 8 Second: Scorer 12	First: Scorer 12 Second: Scorer 8	First: Random Second: 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.

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, and by comparing the means and variances for the two scorers. The second column of Table 12 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

[1] Two additional Intraclass Correlation Coefficients exist, which measure the degree of consistency of scores that are based upon the average or sum of k measures. These are denoted ICC(C,k) and ICC(A,k), for relative consistency and absolute agreement, respectively. These indices parallel coefficient alpha, which also assesses the consistency of total scores based upon k items. I will not discuss these indices further in this paper, but future research should explore the calculation of Standard Change values for the sum or average of k measures, and should compare the Standard Change with ICC(C,k), ICC(A,k), and coefficient alpha.

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.

I compared the means for Scorers 8 and 12 using a dependent samples t-test. This test shows that the these two scorers do have different means (Scorer 8 mean = 68.33, Scorer 12 mean = 67.60, $t(46) = 3.70$, $p = .001$). In absolute terms, the mean difference is small (.73); however, the effect size for the difference is moderate (Cohen's d = .53). This indicates that the two scorers are quite consistent with each other, but when there are differences between them, Scorer 12 usually gave a lower score.

I compared the variances for Scorers 8 and 12 using the Spear procedure developed by McCollough (1987). Wilcox (1990) evaluated several methods of comparing dependent variances and recommended this approach. This procedure showed that Scorer 8 and 12 do not have significantly different variances (Scorer 8 variance = 65.89, Scorer 12 variance = 68.80, $\rho_S(46) = .13$, $p = .40$). Considering these four analyses together, I conclude that there is a small difference in the means for the two scorers, but overall they have an excellent level of consistency.

Next I will use the same four analyses to examine score consistency for 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 13 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. Third, I found that 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). Thus, when the two scorers disagreed, Scorer 3 usually gave lower scores. Finally, I found that these two scorers have significantly different variances (Scorer 3 variance = 75.43, Scorer 9 variance = 57.28, $\rho_S(46) = -.18$, $p = .22$). Considering these four analyses together, I conclude that there is a moderately large difference in the means and a small difference in variances, which impact the overall score consistency.

Table 13
LEAS Score Change Probabilities for Scorers 3 and 9

Score Change	First: Scorer 3 Second: Scorer 9	First: Scorer 9 Second: Scorer 3	First: Random Second: 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.

Usually, an inter-rater reliability coefficient 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, even for research purposes, if the difference in the means would be relevant to the main analyses being undertaken for the study. For example, if the researcher was planning to divide the scoring between these two scorers, then the difference in the means could impact the conclusions from the study. Similarly, if the researcher wanted to compare the mean from this study to test norms for the population, the difference in means between the scorers would impact conclusions.

On the other hand, there are some statistical analyses where differences in means and standard deviations would be irrelevant. For example, differences in means and standard deviations would be irrelevant if a single rater was used to score the entire set of data and those LEAS scores were used in analyses that are not influenced by the mean or standard deviation (such as correlations and factor analyses). Alternatively, if multiple scorers were to be used in a single study, differences in means and standard deviations could be eliminated by transforming the scores. To determine if the scores from these two raters have adequate levels of consistency when differences in means and standard deviations are either eliminated or irrelevant, additional analyses are needed. Specifically, the researcher should transform each score to a z-score, and calculate the consistency between z-scores.

Table 14 shows the probability of various values of z-score change for Scorers 3 and 9. This table shows that z-scores have a 38% chance of being within .25 of each other. However, z-score change values of more than 1 are also moderately common (8%). The Standard Z-score Change is .64, indicating that scores usually change by more than half a standard deviation. I conclude that these two raters do not have an adequate level of consistency with each other even for research purposes, and we cannot achieve an adequate level of consistency by standardizing the scores within rater to eliminate differences in means and standard deviations. This example demonstrates that Score Change Probabilities provide information that is not given in the inter-rater reliability coefficient.

Table 14

LEAS Z-Score Change Probabilities for Scorers 3 and 9

Z-score Change	Probability
-2.00 – -1.75	.02
-1.75 – -1.50	.00
-1.50 – -1.25	.01
-1.25 – -1.00	.01
-1.00 – -0.75	.07
-0.75 – -0.50	.09
-0.50 – -0.25	.10
-0.25 – 0.00	.19
0.00 – 0.25	.19
0.25 – 0.50	.10
0.50 – 0.75	.09
0.75 – 1.00	.07
1.00 – 1.25	.01
1.25 – 1.50	.01
1.50 – 1.75	.00
1.75 – 2.00	.02

Note. First rater is random. Inter-rater reliability = .79. Standard Z-score Change = 0.64. CCC for Z-scores = .79.

Relationship to Existing Measurement Models

Score Change Probabilities are compatible with many different measurement models. Because of this, they can be used in conjunction with statistics that rely upon those specific measurement models. For example, Score Change Probabilities are compatible with Classical Test Theory (Lord & Novick, 1968). In Classical Test Theory, the observed score is imagined to be composed of two independent parts: the True Score and the Error Score. The ratio of true score variance to observed score variance is called the reliability coefficient. This coefficient can be estimated by correlating the data from two parallel measures, such as data from two testing times or two raters. If researchers are interested in the reliability of a composite score, such as the total score across *k* items, then they can estimate this reliability using coefficient alpha. The reliability coefficient from one study can be used to estimate the ratio of true score variance to observed score variance in a new set of data.

Classical Test Theory relies upon three assumptions. First, we assume that the measures are parallel. Specifically, we assume the measures have identical True Scores for every respondent and that the variance of Error Scores is identical for all measures (Lord & Novick, 1968). If the True Scores differ by more than a constant, then the sample reliability coefficient will underestimate the population reliability coefficient (Alwin, 2007). Second, we assume that Error Scores are independent of each other and of True Scores (Lord & Novick, 1968). When multiple factors occur on a test (for example, an Extraversion scale might include factors for Gregariousness, Assertiveness, and Activity Level), Error Scores are not independent of each other, and coefficient alpha will be lower than the reliability coefficient (Kano & Azuma, 2003). Third, Classical Test Theory assumes that the study being done to examine reliability is representative of the kinds of data to which the test user wants to generalize. For example, participants should be drawn from similar populations, and tested in similar environments with similar measurement procedures.

Score Change Probabilities are compatible with Classical Test Theory because Score Change Probabilities only require the third assumption, and do not make any additional assumptions about the data. Thus, Classical Test Theory has more restrictive assumptions than Score Change Probabilities require, and so any data that meets the requirements of Classical Test Theory will also meet the requirements for Score Change Probabilities.

Similarly, Score Change Theory is compatible with the measurement models that underlie Latent Trait Theory (Lord, 1980) and Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Each of those measurement models make some assumptions about the observed scores and the relationships between the observed scores and the unobserved characteristics of the people being measured. If these assumptions are not met, the resulting statistics (e.g., difficulty parameters, Generalizability coefficients) will be invalid (Crocker & Algina, 1986). However, Score Change Probabilities do not make any assumption that is not already made by Latent Trait Theory or Generalizability Theory, and so can be used in any situation where those measurement models can be used. The only assumption made by Score Change Probabilities is one that is already made by all other measurement models: that the study being used to examine consistency is representative of the study in which data will be used to make decisions.

All measurement models implicitly assume that the consistency study is representative of the decision study, and this is the only assumption made by Score Change Probabilities. Therefore, I will explain this assumption in more detail. This assumption has four components. First, we assume that the **participants** being used in the consistency study are similar to the participants that the test user will later make decisions about. We do not need to assume that the participants are similar in every possible way; we only need to assume similarity on the measurement of interest. For example, 5-year old Hispanic children may be similar to 8-year old Hispanic children in terms of sensitivity to sunburn, but dissimilar in terms of how they use playground equipment.

Second, we assume that the **environment** is similar between the consistency study and the decision study. Once again, we do not need to assume that the environment is similar on every possible dimension. Instead, we need to assume that the environment in the consistency study is similar to the environment in the decision study on those dimensions that would be relevant to the construct we are measuring. Let us imagine that we are measuring aggression in children. For this construct, a playground and a backyard might be considered similar, but a playground might be considered dissimilar to a classroom.

Third, we assume that the **scores** that the researcher is using in the consistency study are representative of the kinds of scores that will be used in the decision study. This is the easiest type of similarity to obtain. Often, the same types of test administrators give the same standardized sets of items to participants in the consistency study and the decision study. In both studies, the researchers score the completed surveys the same way, resulting in similar scores between the two studies. However, it would be possible for the scores to be different, especially if the scores relied upon the subjective judgments of raters. For example, consider again the example where we are examining aggression in children. If the consistency study used only female raters, those scores might or might not be similar to the scores obtained in a decision study which used both male and female raters. Empirical research might be needed to determine if the assumption of similarity is reasonable.

Finally, we assume that the **relationships between the scores** that are being used in the consistency study are representative of the relationships between the scores that are being used in the decision study. In a consistency study, the raters are usually independent of each other: they do not talk about the participants before making their ratings, nor make their ratings as a group. However, decisions studies do not always use the same procedures. It would be possible for participants to consult with each other before making their decisions, to complete a single rating form as a group, or to consult with each other and then hand in several rating forms that were identical. Score Change Probabilities can only be generalized to decision studies if the relationship between the scores is the same in the decision study as it was in the consistency study.

All measurement models implicitly assume that the consistency study (which may be called a reliability study or a generalizability study) is representative of the decision study. They all assume that the participants, environment, scores, and the relationship between the scores are similar between the consistency and decision studies. Because this is the only assumption of Score Change Probabilities, they can be calculated alongside statistics that are based upon any existing measurement model.

Summary

This paper has introduced a new method of describing observed score change: the calculation of Score Change Probabilities. These probabilities can be calculated for any type of measurement data. For interval and ratio level data, the test user can calculate differences between the scores and can summarize Score Change Probabilities using the Standard Change. For ordinal level data, it may be more useful to calculate the probability that the top-ranked test takers are the same for the two measurements. For nominal level data, users can calculate the probability that category assignments change.

This paper has also recommended that existing techniques be used to more thoroughly compare two sets of scores. For example, means can be compared using dependent sample t-tests, and variances can be compared using the Spear procedure (MuCollough, 1987). Traditional reliability coefficients assume that the two sets of scores have identical means and variances, but it is often worthwhile to test this assumption. Differences in means and variances are also interesting in themselves, especially if different measurements will be combined, such as when two raters divide the scoring between themselves.

Score Change Probabilities and comparisons of means and variances can be used in conjunction with statistics that are motivated by popular measurement models, such as Classical Test Theory, Latent Trait Theory, and Generalizability Theory. This will provide a more comprehensive evaluation of score consistency. Even if the inter-rater reliability coefficient is high, these statistics might sometimes demonstrate that scores are not very consistent. In some cases, the inconsistency can be corrected by transforming the scores, but in other cases, it cannot. Such calculations demonstrate how Score Change Probabilities extend traditional reliability analyses. Score Change Probabilities can be calculated alongside other reliability analyses, because the only assumption that they require – that the consistency study represents the test takers, environments, and measurements that will be used in the decision study – is already an implicit assumption of all measurement theories.

Appendix
SPSS Calculations

Calculating Score Change Probabilities

It is easy to calculate Score Change Probabilities for interval or ratio level data. In SPSS 15 or 16 use the following steps. First, enter the two sets of scores into SPSS. To calculate the difference scores for the second step, 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**. Finally, click **OK**. Next, to calculate the frequencies and probabilities of score changes for the third and fourth steps, click on the **Analyze** menu, select **Descriptives** from the drop-down menu, and select **Frequencies** from the side-menu. Move **ScoreChange** to the **Variable(s)** box and click **OK**. The resulting table will give the values of the Score Changes in the first column, the frequencies in the second column, and the percentages in the third column. To convert percentages into probabilities, divide the percentages by 100.

Although these precise steps may not work in later versions of SPSS or in other statistical packages, the basic idea of calculating the Score Change values, obtaining the frequency table, and converting the frequencies into probabilities should be easy to implement.

Calculating the Standard Change

To calculate the Standard Change in SPSS 15 or 16, 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 on the output window to find the **mean** value for SquaredChange. Fourth, use a calculator to calculate the square root to obtain the Standard Change.

To calculate the Standard Change values 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.

Footnotes

1. For example, Classical Test Score Theory could be used to estimate the probability that Mary would pass the exam if she took it on another day. First, the test-retest reliability coefficient can be calculated as the correlation between total test scores when the test is administered on two different occasions. Second, we can estimate Mary's theoretical "true score". This calculation requires that we assume a linear relationship between observed scores and true scores, and that true scores and errors are independent and additive. Third, we can estimate the Standard Error of Measurement, which is the standard deviation of observed scores around the true score. This calculation requires the assumption of parallelism, which is unlikely to be met in practice. Finally, we can estimate the probability that Mary's observed score on a test that was taken on another day would exceed the cut-off score for passing the exam. This calculation requires we assume that error scores have a normal distribution.

References

Alwin, D.F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, NJ: John Wiley & Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement, 20,* 37-46.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory for generalizeability of scores and profiles.* New York: Wiley.

Kano, Y. & Azuma, Y. (2003). Use of SEM programs to precisely measure scale reliability. In H. Yanai, A. Akada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.), *New developments in psychometrics* (pp. 141-148). Yokyo: Springer-Verlag. Available from http://www.sigmath.es.osaka-u.ac.jp/~kano/research/paper/dvi/kano_azuma.pdf Accessed Feb 19, 2009.

Kendall, M.G. (1948). *Rank correlation methods.* London: Griffin.

Lane, R.D. (1991). *LEAS Scoring Manual and Glossary.* Unpublished manual for the Levels of Emotional Awareness Test. Available from Richard D. Lane, General Clinical Research Center, University of Arizona, PO Box 245002, Tucson, AZ 85724-5002.

Lane, R.D. & Schwartz, G.E. (1987). Levels of emotional awareness: A cognitive-developmental theory and its application to psychopathology. *American Journal of Psychiatry,* 144, 133-143.

Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, Inc.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McCulloch, C.E. (1987). Tests for equality of variances with paired data. *Communications in Statistics: Theory and Methods, 16*, 1377-1391.

Murphy, K.R. & Davidshofer, C.O. (2005). *Psychological testing: Principles and applications, 6th edition.* Upper Saddle River, New Jersey: Pearson Education, Inc.

Wilcox, R.R. (1990). Comparing the variances of two dependent groups. *Journal of Educational and Behavioral Statistics, 15,* 237-247.

Zumbo, B.D., Gaderman, A.M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.

Acknowledgements

The author wishes to thank the following people for their helpful comments on previous drafts of this paper. Content suggestions: Dr. Samuel Green, University of Arizona; Dr. Kathleen Slaney, Simon Fraser University. Writing feedback: Dr. Chad Cross, Dr. Alexis Kennedy, Danalee Brehman, Garrett Farnes, Amanda Haboush, Saira Haseebullah, Annastasia Koerner, Viridiana Linares, Jennifer Lindley, Mary Pierce Moran, Lauren Shore, and David Weintraub.