Van Blerkom Chapter 17: Alternate Ways to Report Test Scores

1. Percentile Ranks (PR)

 A PR score indicates the percentage of scores that are less than that score. For example, a GRE raw score of 500 may correspond with a PR of 50, which means that 50% of the raw GRE scores are less than 500.

PRs range from a low of 1 to a high of 99. Usually, PR are whole numbers, although one may occasionally see PR with decimals.

PRs do, and do not, represent equal intervals. For PR the differences between each rank represents an equal change – 1 percentage point – so that is equal. So the difference between PRs of 47 and 53 is 6 percentage points, which represents the same difference as between PRs of 93 and 99, again 6 percentage points. The problem measurement specialists warn against is when one converts from PR to raw scores. The difference between PRs of 47 and 53 in raw scores is not nearly the difference observed in raw scores for PRs of 3 and 9 or 93 and 99.

(a) Computation of Percentile Ranks

 PRs are relatively easy to calculate. Simply count the number of raw scores below the raw for which you wish to establish the PR, divide this count by the total number of raw scores, multiple by 100, drop the fraction, and the result is the PR. In formula format:

PR = [and remove fractions]

where nbelow is the number of raw scores below the raw score for which you wish to find the percentile rank and ntotal is the total number of raw scores.

Another method is similar to the above, exact half the raw scores at the raw score for which one wishes to calculate the PR are include in nbelow. In formula notation:

PR = [and remove fractions]

where nbelow+half is the number of raw scores below the raw score for which you wish to find the PR and half represents half the number of raw scores at the raw score for which you wish to calculate the PR.

Once may also use a cumulative relative frequency (cumulative percent in SPSS) in a frequency distribution to obtain the percentile rank. Below is an example SPSS frequency table taken from presentation notes 01b Basic Statistical Concepts. The Cumulative Percent column shows the PR for each score value. For example, score 2 has a PR = 25 while score 6 has a PR = 75.



(b) Unequal Intervals of PRs

 Note that PRs tend to change quickly near the middle raw scores of the distribution, and less so near the extreme scores. For example, using information in table 22.1 (p. 403), note that the difference between scores 40 and 41 represent PRs of 44 and 56, respectively. Yet, the same difference in raw scores of 47 and 48 correspond with PRs of 97 and 99. The reason for this dramatic difference is due to cluster of scores around the middle of the distribution. As more scores cluster on one raw value, PR reflect the larger number of scores more dramatically.

(c) Inappropriateness of Averaging PR in Some Cases

 Often one will wish to perform statistical analysis of test scores. While PR provides an informative score for interpretation purposes, they can be misleading if used for statistical analysis if one attempts to convert PR to raw scores. As noted above, PR are equal intervals and can be used for statistical analysis (i.e., t-tests, correlation, regression, ANOVA), but should not be converted to raw score interpretations when statistically analyzed. If PR are analyzed, then PR should be used for interpretation, not raw scores. Performing analyses on PR then converting to raw scores provides misleading results. Analyzing PR then converting those results to raw scores is problematic because most inferential statistics are based upon the mean, *M*. The mean assumes that scores are of an interval level, which requires that scores represent a scale for which equal difference between scores represent equal intervals. For example, the difference between 43 and 44 is one point; the difference between 98 and 99 is one point. However, if the 50 and 51, and the 95 and 96 are PRs, then the one-point difference between them does not represent the same difference in terms of raw scores (see point (b) above). Because of this, PR are not appropriate for statistical analysis when *M*, or any statistical procedure based upon *M*, is used.

(d) Normal Curve Equivalent (NCE) vs. PR

 NCE scores are standard scores with a mean of 50 and SD of 21.06. The reason NCE are based on a SD of 21.06 is to allow one to convert PR to NCE. Use the 21.06 SD will enable one to match PR on three critical PR scores: 1, 50, and 99. Note that NCE scores, like PR, range from 1 to 99. However, unlike PR, NCE theoretically have scores that represent equal differences for equal differences in raw scores.

2. Distinction Between PR and Percentage Scores

 Both PR and percentage scores are based upon percentages; however, both represent very different phenomena. Recall that PR indicates the percentage of scores below a given score. Percentage scores represent, usually, the percentage of items on a test that were correctly answered. Note that PR enable one to make norm-referenced interpretations, but do not allow for criterion-referenced interpretations. Percentage scores enable one to make criterion-referenced interpretations provided one’s performance objective provides an appropriate criterion performance.

3. Determination of Grade Equivalents

 Grade equivalent scores represent inferred levels of performance in many cases. Usually content taught at a particular grade level is administered in a test to several grade levels (e.g., 2nd, 3rd, and 4th) students at some pre-specified time of the school year (often at the beginning). Median levels of performance are determined for scores at each grade level. If the test was administered at the beginning of the school year, then the median scores will represent the beginning performance anticipated for the respective grade level. For example, if the median raw score for 2nd graders is 31, then 31 will correspond to the 2.0 grade level or grade equivalent.

The raw score distance between grade level medians is used to infer the performance for students at various junctures of the school year. For instance, the raw score difference between 2nd and 3rd grade medians may be 5 points, i.e., 31 (for 2nd graders) and 36 (for 3rd graders). Since 31 is the 2.0 grade level and 36 is the 3.0 grade level, the intervening grade levels are calculated by simply dividing 10 by the number of raw score points between the two medians. Thus, 32 would be 2.2, 33 = 2.4, 34 = 2.6, 35 = 3.8, and then 36 = 3.0.

Typically grade equivalent scores are interpreted to mean performance anticipated at a given grade level with respect to month during the year. Thus, 2.4 indicate the expected performance for 2nd graders in their fourth month.

Suppose a child in the 2nd grade obtains a score of 4.3. This score does not mean that this child can adequately learn content taught in 4th grade. Rather, it simply means that the child is performing on this content in a manner that would be anticipated of 4th graders in their third month.

4. Use of Standard Deviation to Facilitate Interpretation of Scores

(a) Interpretation and Calculation of Standard Deviation

 The standard deviation, SD or s, is a measure of variability, or spread, of scores. The SD indicates how much scores differ or vary. The SD does not indicate how far a score is from the mean, but it does indicate how much scores vary from the mean.

In general, smaller values of SD indicate less variability and larger values indicate more variability. SD may be calculated using the following formula:

SD =

where X is the raw score obtained for each student (such as Bill scored a 650 on the verbal section of the GRE), *M* is the average or mean of the scores in the distribution (e.g., 500 is the average score obtained), and n is the sample size—the number of scores one has in the distribution (i.e., 20 students took the GRE).

Below is an example calculation for the standard deviation. Suppose the following IQ scores were obtain for a class of students:

|  |  |
| --- | --- |
| Bill | 95 |
| David | 103 |
| Will | 111 |
| Willy | 106 |
| John | 87 |
| Paul | 98 |

The mean, *M*, is

M = = (95+103+111+106+87+98)/6 = 100.00.

Calculation of the SD is illustrated below using raw scores (X), mean (M), deviation scores (X-M), and squared deviation scores (X-M2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Student | X | M | X-*M* | (X-*M*)2 |  |
| Bill | 95 | 100 | -5 | 25 |  |
| David | 103 | 100 | 3 | 9 |  |
| Will | 111 | 100 | 11 | 121 |  |
| Willy | 106 | 100 | 6 | 36 |  |
| John | 87 | 100 | -13 | 169 |  |
| Paul | 98 | 100 | -2 | 4 |  |
|  |  |  |  | S(X-*M*)2= | 364 |

Using the information obtained from the table above, the SD may be calculated as:

SD = = = = 8.532.

(b) Using Standard Deviation to Interpret Test Scores

 b1. Characteristics of Standard Deviation

 The SD indicates the amount of variability in a set of scores. As scores become more diverse, the SD will correspondingly become larger. Thus, larger SD’s typically mean more spread and dispersion of scores.

When the distribution of scores from a test form a normal (or near normal) shape, then approximately 68% of all scores will be within one SD of the mean, i.e. 1 SD about the mean. Similarly, about 95-96% of scores will be within two SD of the mean, i.e., 2 SD.

 (2) Expressing Scores in Standard Deviation Units

 Note that knowledge of a test score, such as 53, does not enable one to make an interpretation which more information. If one knew both M and SD, then one could make a norm-referenced interpretation. For example, if the 53 represents a score derived from a test with M = 50 and SD = 3, then one could state that the 53 is a score that is one SD above the mean. This student, then, probably performed better than most on the test (assuming an approximately normal distribution).

One may convert any score into a z-score, which is a standard score, with the following formula:

z = = = 1

Note that z score indicates how far above or below then mean a given score is in SD units. Thus, if a z score were -1, then the corresponding raw score is one SD below the mean. If the z score is 1.7, then the raw score is 1.7 SD above the mean.

 (3) Comparing a Student’s Performance on Two Exams

 If one wishes to compare a student’s performance on multiple exams to other students, then one must convert scores into standard scores (e.g., z scores) to make the norm-referenced interpretation. The example a student scores 12 of 20 on one exam and 15 of 20 on a second exam. Relative to other students, how well did this student perform? After converting the scores to z scores, on exam one the student was 1 SD above the mean (z = 1), and on the second exam, the student was half a SD above the mean (z = 0.5). This information indicates that the student probably scored better than more than half the students who took the tests.

(c) Limitation of Standard Deviation Units When Interpreting Scores

 Standard scores (scores based upon standard deviations) do not indicate absolute performance, only relative performance. Thus, standard scores do not enable one to make criterion-referenced interpretations.

5. Transforming Scores to a Standard Scale

 When raw scores are converted to a format that has a predetermined mean and SD, such as z scores, then the transformed scores are call standard scores. Thus, z scores represent standard scores because all z scores for a distribution of scores have a mean of 0.00 and SD of 1.00 (assuming the original distribution had a SD greater than 0.00).

6. Interpretation of Commonly Used Standard Score Scales

 There are many standard score scales available. Below are several and their corresponding interpretations.

6a. T-Scores

 T-scores have a mean of 50 and SD of 10. Thus, students with scores greater than 50 have scored above the mean, and those with scores less than 50 scored below the mean. Since the SD of T-scores is 10, one can say that a student with a t-score of 60 is one SD above the mean, and one with a score of 35 is -1.5 SD below the mean.

6b. Deviation IQ Scores

 Many intelligence tests standardize scores to have a mean of 100 and a SD of 15 or 16. These standardized scores are referred to as deviation IQ scores.

6c. Stanine Scores

 The name stanine is derived from standard nine. Stanine scores have a mean of 5, SD of 2, and range of 1 to 9. Individuals with a score that is 1 SD above the mean would have a 7, 1 SD below the mean a 3, and 0.5 SD below the mean a score of 4. Raw scores that are 2 or more SD above the mean are assigned 9, and 2 or more SD below the mean are assigned 1.

6d. Normal Curve Equivalent (NCE) Scores

 NCE scores have a mean of 50 and SD of 21.06. More information on NCE is presented in supplemental reading linked on the course web page – see Hill’s Handy Hints.

6e. Considerations When Interpreting Standard Scores

 When interpreting standard scores, it is important to know for which test and group of students the scores were derived. One cannot compare standard scores derived from different tests or groups of subjects. For example, if one has a t-score of 73 for the GRE, one cannot say that this individual would be more than 2 SD above the mean for a different test, such as the MAT. In short, carefully interpret standard scores with respect to test and population.

The figure below provides a pictorial display of the various standard scores discussed in terms of a normal distribution.



6f. Advantages of Standard Scores

* Standard scores produce equal interval scores unlike percentiles and grade-equivalents. This does not mean, however, that standard scores represent an interval level of measurement.
* Standard scores may be averaged across students unlike percentile ranks and grade-equivalents.
* Standard scores can be used to compare a student’s performance across tests or testing situations. (Percentile ranks allow this comparisons).

6g. Limitations of Standard Scores

* Standard scores often suggest a degree of precision that cannot be justified by the test. So while standard scores may produce equal intervals, this does not mean the variable represents an interval scale. Further, when making comparisons across students on the same test, one should interpret scores, such as IQ, to be different unless more than about 1/3 of a SD exist. For example, IQ tests usually report scores with a SD of 16, so one should not conclude that people with scores of 133 vs. 137 are really different. However, one may claim that scores as discrepant as 133 vs. 123 do represent a real difference.
* Standard scores, as previously mentioned, represent relative position, so other interpretations (criterion, growth) are not possible with standard scores.

7. Using Standard Deviation to Interpret Errors in Measurement

 When one takes a test, one responds to a sample of test items. Because of this, it is not possible to obtain a student’s true score on the skill measured—a true score can only be obtained if all items are included, or if a test is perfectly reliable. Rather than measure true scores, test provide an estimate of a student’s true score. This estimate is called typically call one’s test score, observed score, or just score. Every test measures, with some degree of error, each student’s true score. The error of measurement is defined as the difference between the observed score and true score (if it were every observed). Since error exists for every score obtained, one could theoretically calculate the standard deviation of these errors.

7a. Nature of Measurement Error

 As noted above, measurement error occurs when one obtains a score that differs from one’s true score. One problem faced by researchers is that one does not know whether an obtained score over or under estimates the true score because measurement error is random. So for some the obtained score will underestimate their true score and for others it will overestimate their true scores.

While one cannot determine the degree of error for a given individual, it is possible to estimate the degree of error for a group of examinees. More specifically, the SD of measurement errors can be estimated. The SD of measurement errors is referred to as the **standard error of measurement (SEM).**

7b. Interpreting Standard Error of Estimate

 The estimate of SEM is simply a standard deviation for errors in scores. Because SEM is a SD, and because measurement error is random and therefore likely to be normally distributed, one may conclude that about 2/3 of all obtained scores will be within ±SEM of their true scores. For example, suppose the SEM of an IQ test is 3.1. Then one could expect that about 67% of all scores reported for the test will be within 3.1 points of the true score. Further, about 96% of the scores will be within ±2SEM of the true score. For example, if one scores 116 on the test, then one could be about 96% confident that the individual’s true score is within the range of 116-6.2 = 109.8 to 116+6.2 = 122.2.