Statistical Analysis of Test Results

ANTONY J. KUNNAN AND NATHAN T. CARR

Introduction

Statistical analysis of test results provides information to test developers and users that can be used as evidence to build an overall validity argument which supports links between the observations of performance on language tests to score interpretations, and score-based decisions (Kane, 1992, 2006; Xi, 2008). The choice of which statistical analyses to perform, and their interpretation, depend in part on whether the test results to be analyzed are from a norm-referenced test (NRT) or a criterion-referenced test (CRT). In the former, a test taker's performance is compared to other test takers' performances who took the same test; while in the latter approach, a test taker's performance is interpreted with reference to a set of performance criteria (such as standards or mastery levels with respect to test content) previously set by the test developer or a stakeholder. Sometimes, a test is designed to be used as both an NRT and a CRT, but this fact is also relevant to statistical analysis.

Statistical Analyses

Statistical analyses provide a means of systematically summarizing examinees' performance on a test to provide information about both the examinees and the test as a measure for a particular group of examinees. The latter use of statistical analysis is the main focus here—the use of statistics for evaluating how good a test is for its intended purpose. The statistical procedures are grouped according to four of the perspectives on test quality, which can be used in demonstrating the validity of test interpretation and use: evaluation, generalization, explanation, and utilization (following Bachman, 2005, and Chapelle, Enright, & Jamieson, 2008).

Evaluation

Evaluation issues pertain to the defensibility of the manner in which examinees' performance is summarized into test scores. Examinees' scores are used to reflect their performance and therefore the link between performance and score needs to be supported in several ways.

First, it is important to understand the distributions of the test takers. Statistical analysis can reveal the distributions through calculation of the following: mean, median (used for small samples), mode, and standard deviation. These results will provide information on the examinees' scores in terms of the average score (mean), its mid-point (median), most frequently occurring score (mode), and the overall variation from the average (standard deviation). For large groups and in contexts where further statistical analysis is planned, skewness (scores bunching toward to the right, center, or left of a score distribution shape) and kurtosis (whether the scores are peaked or flat in terms of the score distribution shape) would also be examined. Overall, these descriptive statistics inform the test researcher or classroom teacher about how a test-taking sample performed on a test. (See Carr, 2008, for a more complete discussion.)

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal1115

Second, we need to evaluate the statistical characteristics of items, tasks, and tests. One approach in assessing the quality of test items and tasks or improving their overall quality is to pilot (or pretest) the test items or tasks. The data obtained need to be scored and coded to provide a database for statistical analyses. Statistical analyses can then be conducted with a view to evaluating whether the items and tasks behave as intended. The identification of which items are appropriate or problematic is accomplished through *item analysis*, consisting of analysis of *item difficulty*, *item discrimination*, and *distractor analysis*.

Item difficulty is most commonly estimated by using *item facility* (*IF*; also commonly referred to as *p*), the proportion of test takers who responded to a test item correctly. In the case of NRT, ideally, items will have an IF near .50; that is, half of the test takers will answer each item correctly. The item will then provide the maximum information for the largest group of test takers (i.e., those in the middle ability range, if scores are in fact distributed normally, as they should be on an NRT). Allen and Yen (1978) point out, however, that in practice, items ranging between .30 and .70 are generally acceptable. Items outside this range are deemed to be too difficult or too easy, respectively. On the other hand, in CRT we hope that 50% of the test takers whose ability level is at the cut score will answer a given item correctly. In practice, unfortunately, this is a very difficult proposition to manage; as a result, in CRT item analysis, IF is usually only reported for general information, and to assist with the interpretation of discrimination and distracter analysis results.

Item discrimination can be analyzed following two approaches for both NRT and CRT: correlational and subtractive. In NRT, the correlational approach uses the point-biserial (pb(*r*) or $r_{\text{p-bis}}$) correlation coefficient between each item and the total score on that section of the test. If the test is one homogeneous whole (e.g., composed entirely of discrete-point grammar items, rather than separate grammar and vocabulary sections), of course, the items should be correlated with total test score. The correlational index is interpreted as the relationship between performance on the item and on the section or test as a whole; it is therefore simply an item-total correlation. The subtractive NRT index is the upper–lower discrimination (*item discrimination*, abbreviated ID or ID_{UL}). This is the difference between the IF for high-scoring and low-scoring test takers; these two groups are usually defined as the top and bottom 25%, 27%, or thirds of the test takers taking the test. One drawback to using ID is that it ignores the performance of all the test takers in the middle. Thus, when sample sizes are large enough ($n > 35$, perhaps), the point-biserial may be preferable.

In CRT, the correlational approach uses the ϕ (phi; also referred to as item ϕ) correlation coefficient between each item and mastery/nonmastery classification (mastery = 1, nonmastery = 0). Items should, of course, be correlated with section mastery/nonmastery when there are distinct sections. The most commonly used CRT subtractive approach is the B-index, which is calculated by subtracting the IF for all test takers who failed the test (nonmasters) from that of test takers who passed the test (masters). Another less commonly used subtractive CRT index is the difference index (DI). This requires administering a test twice, either to the same group before and after instruction, or to one group of test takers who are presumed to have mastered something and another group who presumably have not (e.g., test takers in two levels of a language program). Like the B-index, it is calculated by subtracting the IF values for the two groups. Both item phi and the B-index are cut-score dependent; that is, as with phi lambda ($\Phi(\lambda)$), if there are multiple cut scores, they must be reported for each cut score. Similarly, when more than two groups are used for the DI (as in a placement test), it must be reported for each pair of adjacent groups (levels one and two, levels two and three, and so on).

When interpreting discrimination indices for an NRT or CRT, the same rules of thumb apply. For the point-biserial or item ϕ, ideally, items should have discrimination values of .30 or above. For ID, the B-index, and the DI, values should be .40 and above. These

values are what is typically expected in the context of professional item development for high-stakes testing, though; in the case of locally developed tests, particularly those to be used for classroom testing, it is often necessary to accept lower values. These rules of thumb—both for discrimination and for difficulty—are *not* absolutes, and it is important to keep in mind that these analyses matter most when it is necessary to improve a test with poor internal consistency. When a test already has high reliability or dependability, it is probably only necessary to worry about the worst items, unless one is hoping to shorten the test.

This discussion of difficulty and discrimination assumes that scoring is dichotomous; that is, that each question is scored as either correct or incorrect, with no partial credit possible. With polytomous scoring (i.e., when questions are worth more than one point, with partial credit possible), however, some adjustments must be made. First, rather than IF, we use IF*, which is equal to the average score for a given item divided by its number of points possible. We then use IF* in calculating the $ID_{U\text{-}L}$, B-index, or DI. Technically, this also changes the correlation coefficient being used, so that the point-biserial and phi are not being calculated. Practically speaking, however, there is no change, since these are both Pearsonian coefficients, and a computer will use the same formula. The only difference is that the correlation should technically be called a Pearson *r*, rather than a point-biserial. These slight complications notwithstanding, though, the interpretations for IF*, $ID_{U\text{-}L}$*, B*, DI*, and the two correlational estimates of discrimination are the same as for their NRT equivalents.

Distractor analysis: Item analysis, or consideration of difficulty and discrimination, tells us which questions on a test are problematic. In the case of multiple-choice tests, we can then examine the distractors (incorrect choices) on those questions to help improve them. One way to do this is to see what percentage of test takers selected each option, and then revise or replace each one that was not selected by at least 10% of the test takers (Bachman, 2004). Two other approaches involve examining the relationship between test takers' total scores and their distractor choices. The first of these methods is to calculate the point-biserial correlation of each option with total score. This is done by, in effect, "scoring" an item as many times as it has options; thus, a three-option question would undergo the process three times, once for each option. The correlation for the correct answer will, of course, be the same as the item's point-biserial estimate. The correlations for the distractors, on the other hand, should all be negative. A *positive* correlation for a distractor is problematic, particularly if it is not close to zero—that is, if choosing a particular incorrect option is associated with having a higher score on the test, then something is clearly wrong. A final method is to compare the proportions of test takers in the high and low groups (the same ones as are used in the ID, B-index, or DI) who selected each distractor.

Third, we turn to analyzing rating scales and raters. When analyzing scores based on ratings, we are concerned with both inter-rater consistency (how similar the ratings are that are produced by different raters), as well as intra-rater consistency (how consistent each rater is; that is, whether the rater would give the same score at different times or on different occasions).

Perhaps the simplest approach to estimating the consistency of ratings is to correlate sets of ratings. This correlation is often reported by itself, but to be used as a reliability estimate, it should be adjusted using the Spearman-Brown prophecy formula. When there are more than two ratings, the Fisher *Z*-transformation should be applied to each correlation, the correlations then averaged together, and the result retransformed to a correlation coefficient. Another approach that can be used when there are multiple scores (for multiple raters, multiple tasks, or for individual subscales of an analytic scoring rubric) is to calculate Cronbach's alpha as an estimate of the internal consistency of the ratings.

A more informative approach is to use generalizability theory to estimate the consistency of scoring. This offers the added advantage of identifying which facets of the measurement process are contributing most to unreliability. The main disadvantage of using generalizability theory is that it requires a greater degree of technical expertise. The same is also true of many-faceted Rasch theory, which can be used to complement generalizability theory. Its greatest advantages are that it provides ability estimates for each test taker, and difficulty or severity estimates for each rater, task, or rubric subscale. It can also provide diagnostic information on whether individual raters are showing more or less variability in their ratings than would be expected.

In "messy" rating situations, those in which there is inconsistent overlap in terms of which rater scored which test takers, the rater agreement proportion (RAP) can be used. This is the proportion of ratings for a given task (essay, spoken response, etc.) that are the same; for example, if three out of four ratings are the same, the RAP is .75.

Fourth it is important to understand differences in group performance. When different test-taker groups take a test, some groups may perform better than others on some test items. These groups may self-report memberships through a questionnaire (examples of self-reporting are gender, race and ethnicity, age, native language, second language learning, etc.), assigned by a test designer (examples, test accommodations, planning conditions, computer use, etc.) or assigned by a researcher in an experimental research setting (examples, experimental group versus control group; treatment 1 group versus treatment 2 group, etc.).

The analysis of variance and *t* test statistics can provide information as to whether group differences on certain test items or tasks are statistically significant. Test-score differences among test-taker groups will normally also require an examination of test bias.

Generalization

To evaluate the assumptions associated with generalization, test scores need to be analyzed in terms of their reliability, dependability, and generalizability. In NRT, the concern is with the consistency of scores (technically, consistency of ranking, but in practice it is generally treated as consistency of scoring), which is referred to as reliability. In CRT, we are concerned with two areas of consistency, both referred to as dependability. The first is the dependability of scores, and the second is the dependability of classification. In general, however, it is important to provide an index of score consistency (and classification consistency, for CRT), as well as an estimate of the margin of error associated with the estimate (for NRT, the standard error of measurement, or SEM; for CRT, the criterion-referenced confidence interval, abbreviated the CRT CI or CI_{CRT}).

Two important alternative paradigms for looking at measurement consistency have been developed in the last few decades: generalizability theory (G theory) and item response theory (IRT). Generalizability theory is particularly suited to analyzing rated performances, as in tests of speaking or writing, but is not limited to these contexts, however. G theory can be used to analyze both NRT and CRT results, and also provides estimates of the margin of error for test scores, as well as classification consistency for CRTs. IRT is particularly suited for working with large (i.e., numbering hundreds or thousands) test-taker groups. Detailed explanation is beyond the scope of this entry, but it is important to mention an IRT variant known as the many-facet Rasch model, which is particularly useful with rated performances.

Explanation

To support explanations of score meaning the key applications of statistics are explaining the constructs being assessed, explaining test performance, and conducting concurrent validation studies. One way of looking at score meaning is through an analysis of the

composition or structure of the scores; whether the test structure (or language abilities as operationalized in the test) is unitary or divisible (or multicomponential). This can be examined through exploratory factor analysis (EFA), which seeks to identify hypothetical factors that account for the patterns of correlations that are observed in test scores (from individual items or tasks). In confirmatory factor analysis (CFA), a proposed test structure is specified in advance and the data available in the form of test scores are used to evaluate it. Structural equation modeling can also be used for this purpose (see Kunnan, 1995, 1998). Concurrent validation studies are conducted by examining how well a new test's scores correlate with some other test's scores that have already been demonstrated to be useful in decision making, or have been demonstrated to be a good measure of the intended construct. Such studies use correlations to analyze the data.

Utilization

The utilization portion of a validation argument needs to provide support for the intended test uses. Statistics are helpful in this area in standard setting, estimating classification errors, and understanding test consequences and washback.

When test takers receive scores for their test performance, they are typically accompanied by other classifications such as "pass" or "fail," or categories or levels of performance (such as "needs improvement," "basic," "proficient," "advanced," or "unqualified," "qualified" and so on). These classifications are based on standards that have been set either in terms of pre-specified percentages of "pass" or "fail" (such as 5% or 10% pass rate) as in NRTs, or pre-specified performance or content standards (operationalized in terms of cut-points or cutoff score) as in CRTs. While both types of standard-setting procedures need to be defended, performance and content standards are more complex because the procedure involves many steps and a clear research design. Understanding consequences and washback are other ways of appreciating the effect of utilization or decision making of the test.

Software and Reporting

Software

Many software programs, some general and others written specifically for assessment, can be used for analyzing test results. For entering, organizing, and cleaning data, Microsoft Excel or a similar spreadsheet is a simple, easily available, and easy-to-use choice. It also provides adequate descriptive statistics, has very powerful features for creating graphs, and can calculate the Pearson *r*, although non-Pearsonian correlations must be calculated manually. For anything beyond the simplest statistics, however, a true statistics package should be used. Some of the most common ones are IBM SPSS Statistics (formerly SPSS), STATA, BMDP, R, and SAS.

For NRT and CRT item and distractor analyses, Microsoft Excel can be used fairly easily, and can also be employed for calculating reliability (Carr, 2011). As for software specifically developed for handling reliability and item analysis, Lertap and ITEMAN are two commonly used packages. Lertap runs within Microsoft Excel, taking advantage of its file formats, output options, and familiar interface. ITEMAN, on the other hand, works with ASCII-format data files.

For generalizability theory, while the basic estimates can be calculated using results from general statistics packages, there are three packages developed by Brennan that are available for free and can calculate results automatically—aside from calculating proportions of variance, which requires manual addition and division by the user. These are GENOVA, the original program; urGENOVA, which accommodates unbalanced designs;

and mGENOVA, which accommodates both unbalanced and multivariate designs (e.g., when there are multiple subscales within a rubric).

For IRT, BILOG and MULTILOG are perhaps two of the best-known packages, with BILOG used for handling dichotomous data, and MULTILOG for polytomous items. The program PARSCALE, on the other hand, can apply both dichotomous and polytomous models. Two well-known programs for applying the many-facet Rasch model, as opposed to IRT models in general, are FACETS and WINSTEPS.

For factor analytic studies, IBM SPSS Statistics can be used to perform exploratory factor analyses. Other programs are needed, however, to perform confirmatory factor analyses and structural equation modeling (SEM), the best-known are EQS, LISREL, AMOS, and M+.

Reporting

Reports of statistical analysis results should contain descriptive statistics and correlations among all the variables. The number of decimal places to be reported generally depends upon the scale of the variables, but for correlations, two or three places is the norm, although five places is appropriate if the correlations might be used as data for factor analytic studies replicating or extending the project. Test- or survey-based studies should include reliability or dependability estimates for the overall test or survey, as well as for each section or scale for which separate scores are reported. Each reliability or dependability estimate should be accompanied by its corresponding CRT confidence interval or standard error of measurement.

In reports or papers discussing CFA or IRT results, model fit must be discussed. In these studies, as well as in those using EFA, it is also important to explain how the final model was developed, how it compared to other plausible models, which other models were considered, and why they were rejected. Factor loadings should be reported in all factor analytic studies, as well as any correlations among factors. In the case of CFA and SEM studies, diagrams or figures of models are important for ease of understanding by readers. All journal articles or research reports based on factor analytic studies, particularly those employing CFA and SEM, should include an appendix containing the correlation or covariance matrix and standard deviations of all variables included in the study. This information is essential in allowing others to replicate and verify results.

Conclusion

These statistical analyses of test results provide an overview of the types of procedures that are used to investigate how tests perform with particular groups of examinees. These types of analyses are needed to build a validity argument which supports the links that test users hope to be able to make between test performance, test-score interpretations, uses, and decisions. Only fundamental statistical procedures used in standard test analysis contexts were included in this entry. Variations in test constructs, research design, data collection, and research questions might need approaches not discussed here.

SEE ALSO: Bias in Language Assessment; Comparing Two+ Independent Groups; Correlational Research in Language Assessment; Cut Scores on Language Tests; Factor Analysis; Generalizability Theory in Language Testing; Inference; Structural Equation Modeling; Validation of Language Assessments; Washback in Language Assessment

References

Allen, M. J., & Yen, W. M. (1978). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1–34.

Carr, N. T. (2008). Using Microsoft Excel® to calculate descriptive statistics and create graphs. *Language Assessment Quarterly, 5,* 43–62.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.

Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. London, England: Routledge.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–35.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–65). Westport, CT: Praeger.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling study*. Cambridge, England: Cambridge University Press.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment. *Language Testing, 15,* 295–332.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 177–96). New York, NY: Springer.

Suggested Readings

Brennan, R. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Embretson, S., & Reize, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

McDonald, R. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.