Generalizability theory, or G theory, is a statistical framework for conceptualizing, investigating, and designing reliable observations. It is used to determine the reliability (i.e., reproducibility) of measurements under specific conditions. It is particularly useful for assessing the reliability of performance assessments. It was originally introduced in Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963).

Overview [edit]

In G theory, sources of variation are referred to as facets. Facets are similar to the "factors" used in analysis of variance, and may include persons, raters, items/forms, time, and settings among other possibilities. These facets are potential sources of error and the purpose of generalizability theory is to quantify the amount of error caused by each facet and interaction of facets. The usefulness of data gained from a G study is crucially dependent on the design of the study. Therefore, the researcher must carefully consider the ways in which he/she hopes to generalize any specific results. Is it important to generalize from one setting to a larger number of settings? From one rater to a larger number of raters? From one set of items to a larger set of items? The answers to these questions will vary from one researcher to the next, and will drive the design of a G study in different ways.

In addition to deciding which facets the researcher generally wishes to examine, it is necessary to determine which facet will serve as the object of measurement (e.g. the systematic source of variance) for the purpose of analysis. The remaining facets of interest are then considered to be sources of measurement error. In most cases, the object of measurement will be the person to whom a number/score is assigned. In other cases it may be a group or performers such as a team or classroom. Ideally, nearly all of the measured variance will be attributed to the object of measurement (e.g. individual differences), with only a negligible amount of variance attributed to the remaining facets (e.g., rater, time, setting).

The results from a G study can also be used to inform a decision, or D, study. In a D study, we can ask the hypothetical question of "what would happen if different aspects of this study were altered?" For example, a soft drink company might be interested in assessing the quality of a new product through use of a consumer rating scale. By employing a D study, it would be possible to estimate how the consistency of quality ratings would change if consumers were asked 10 questions instead of 2, or if 1,000 consumers rated the soft drink instead of 100. By employing simulated D studies, it is therefore possible to examine how the generalizability coefficients (similar to reliability coefficients in Classical test theory) would change under different circumstances, and consequently determine the ideal conditions under which our measurements would be the most reliable.

Comparison with classical test theory [edit]

The focus of classical test theory (CTT) is on determining error of the

measurement. Perhaps the most famous model of CTT is the equation X = T + E {displaystyle X=T+E} , where X is the observed score, T is the true score, and e is the error involved in measurement. Although e could represent many different types of error, such as rater or instrument error, CTT only allows us to estimate one type of error at a time. Essentially it throws all sources of error into one error term. This may be suitable in the context of highly controlled laboratory conditions, but variance is a part of everyday life. In field research, for example, it is unrealistic to expect that the conditions of measurement will remain constant. Generalizability theory acknowledges and allows for variability in assessment conditions that may affect measurements. The advantage of G theory lies in the fact that researchers can estimate what proportion of the total variance in the results is due to the individual factors that often vary in assessment, such as setting, time, items, and raters.

Another important difference between CTT and G theory is that the latter approach takes into account how the consistency of outcomes may change if a measure is used to make absolute versus relative decisions. An example of an absolute, or criterion-referenced, decision would be when an individual's test score is compared to a cut-off score to determine eligibility or diagnosis (i.e. a child's score on an achievement test is used to determine eligibility for a gifted program). In contrast, an example of a relative, or norm-referenced, decision would be when the individual's test score is used to either (a) determine relative standing as compared to his/her peers (i.e. a child's score on a reading subtest is used to determine which reading group he/she is placed in), or (b) make intra-individual comparisons (i.e. comparing previous versus current performance within the same individual). The type of decision that the researcher is interested in will determine which formula should be used to calculate the generalizability coefficient (similar to a reliability coefficient in CTT).

Notes [edit]

References [edit]

Reference

Anatomy & Physiology Super Review

Ethnocultural Diversity and the Home-to-School Link (Research on Family-School Partnerships)