ANALYZING QUANTITATIVE (NUMERIC) ASSESSMENT DATA

Office of Assessment of Teaching and Learning, Washington State University

Assessment data provides a means to look at student performance in order to offer evidence about student learning in the curriculum, provide information about program strengths and weaknesses, and guide decision-making. Analyzing the data -- in context -- gives meaning to the information collected and is essential in order to appropriately utilize and communicate the assessment results.

Quantitative data analysis relies on numerical scores or ratings and can be helpful in evaluation because it provides quantifiable results that are easy to calculate and display. Quantitative assessment data can come from a variety of assessment measures, including rubric evaluations of student work, pre-test/post-test assessments, standardized tests, embedded assessments, supervisor evaluations of interns, surveys, and course evaluations. This resource is intended to help faculty consider good practices and plan for summarizing quantitative data collected about student learning as part of program or course assessment.

ATL is available to work with undergraduate programs to analyze and create visual displays of assessment data to engage faculty in discussions of assessment results; please contact ATL for additional information.

Before You Begin: Purpose, Context, Audience

There is no "one size fits all" approach to analyzing quantitative assessment data, but there are some ways to make it more approachable. It's best to start thinking about your data analysis plan when you are first identifying your assessment questions and determining how you will collect the needed information. It is important to match the analysis strategy to the type of information that you have and the kinds of assessment questions that you are trying to answer. In other words, decisions about how to analyze assessment data are guided by what assessment questions are asked, the needs and goals of the audience/stakeholders, as well as the types of data available and how they were collected. For example:

- Targets or benchmarks can be expressed in different ways and therefore dictate how assessment data are summarized and displayed. If a benchmark is stated in the form of a percentage (i.e. 80% of students will meet the level of expectation for a specific learning outcome), it would be appropriate to provide percentages in the data summary. On the other hand, a benchmark may be related to an average (i.e. the mean score on the licensure exam for students in our program will be above national average). In that case, it would be appropriate to determine means when analyzing the data.
- Data collection processes may vary between and within different types of assessment measures. For example, assessment data may be collected from all students in a program (a census) or a subset of those students (a sample) and the number of students included can be quite large or very small, depending on the size of the program. Pieces of evidence may have been reviewed/scored by one rater or many. Assessment data may also be collected at one point in time or over several years.

Typically, assessment data are intended for discussion and use by program faculty, who are familiar with the discipline, curriculum, and other sources of related, complementary data. When carefully analyzed and interpreted in the context that they were collected, assessment data can offer useful insight into curricular coherence and effectiveness. Data can be misleading, or worse, when they are taken out of context or used for purposes other than originally intended and agreed upon. As a result, you will want to understand the purpose and scope of the project, the assessment questions that guided the project, the context, and the audience for the results before any type of analysis occurs. You should be familiar with

the basic data collection processes, including how the data were collected, who participated, and any known limitations of the data, as this can help you make an informed decision about what the data can reasonably reveal. Other factors to consider may include: How was the random sampling/sample size determined? What was the response rate? Were well-established, agreed-upon criteria (such as a rubric) used for assessing the evidence for each outcome? How were raters normed/calibrated? Did multiple raters review each piece of evidence? Has this measure been pilot tested and refined? As a good practice, a short written description of the data collection processes, number of participants, and a copy of any instrument used (i.e. rubric, survey, exam) should accompany the data analysis file, data summary, and/or final report.

Levels of Quantitative Data

There are three main levels of quantitative data in assessment: nominal, ordinal, and interval/ratio.

- **Nominal or categorical data** are items which are differentiated by a classification system, but have no logical order. Each category may be assigned an arbitrary value, but there is no associated numerical value or relationship.

  Example 1: Male = 0, Female = 1
  Example 2: No = 0, Yes = 1

- **Ordinal data** have a logical order, but the differences between values are not constant. Again, each category may be assigned a value, but there is no associated numerical value or relationship beyond order. For example, numbers assigned to the categories convey "greater than" or “less than” relationships; however, how much greater or less is not implied.

  Example 1: Education Level (High School – 1, College Graduate – 2, Advanced Degree – 3)
  Example 2: Agreement Level (Strongly Agree – 1, Agree – 2, Neutral – 3, Disagree – 4, Strongly Disagree – 5)

- **Interval/ratio data** are continuous, with a logical order standardized differences between values.

  Example 1: Years (2010, 2011, 2012)
  Example 2: # of Credit Hours

How does the level of measurement impact data analysis?

The following sections contain multiple strategies for analyzing quantitative data and it is up to you to decide which analysis methods make sense for your specific data and context. Keep in mind that statistical computations and analyses assume that variables have specific levels of measurement. While nominal/categorical and ordinal data may be assigned numerical values, it may not make sense to apply certain analysis techniques to these data.

For example, a question may ask respondents to select their favorite color (1 – red, 2 – yellow, 3 – blue, 4 – green). While it is possible to calculate the mean or median response based on the assigned arbitrary values, it does not make sense to calculate a mean or median favorite color. Moreover, if you tried to compute the mean education level as in example 1 in the previous ordinal data section (High School – 1, College Graduate – 2, Advanced Degree – 3), you would also obtain a nonsensical result as the spacing between the three levels of educational experience is uneven. Sometimes data can appear to be "in between" ordinal and interval; for example, a five-point Likert scale with values "1 – strongly agree", "2 – agree", "3 – neutral", "4 – disagree" and "5 – strongly disagree". If you cannot be sure that the intervals between each of these values are the same, then you would not be able to say that it is interval data (it would be ordinal).

Descriptive Statistics

While statistical analysis of quantitative information can be quite complex, relatively simple techniques can provide useful information. Descriptive statistics can be used to describe the basic features of your data and reduce it down to an understandable level. Descriptive statistics form the basis of virtually every quantitative analysis of data. Common methods include:

- **Frequency/Percentage Distributions.** A frequency distribution is tallies/counts of the number of individuals or scores located in each category. A percentage distribution displays the proportion of participants who are represented within each category (i.e. the number of participants in a category divided by the total number of participants). Tabulating your results for the different variables in your data set will give you a comprehensive picture of what your data looks like and assist you in identifying patterns. Frequency/percentage distributions are generally appropriate for all types of quantitative data. In some cases, it may be useful to group categories when examining frequency distributions. For example, examining tallies/counts of the number of students with GPAs between 0.0-0.99, 1.0-1.99, 2.0-2.99, 3.0-4.0) as opposed to creating a frequency distribution containing counts of every possible GPA.
- **Measures of Central Tendency.** Measures of central tendency are used to describe the number that best represents the "typical" score or value of a distribution. The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.
  - **Mean** – the average score for a particular variable. *Note: Meaningful averages can only be calculated from interval/ratio data that are roughly normally distributed (i.e. bell-shaped); the median (see following) is a better measure of central tendency for skewed data. Means may be of limited or no value for nominal/categorical and ordinal data, even where numbers are assigned.*
  - **Median** – the numerical middle point of a set of data that had been arranged in order of magnitude (i.e. the median splits the distribution in half). *Note: Meaningful medians can only be calculated from ordinal and interval/ratio data. Medians may be of limited or no value for nominal/categorical data, even where numbers are assigned.*
  - **Mode** – the most common number score or value for a particular variable. *Note: Mode is appropriate for nominal/categorical, ordinal, and interval/ratio data. A set of data can have more than one mode.*
- **Measures of Spread.** Measures of spread describe the variability in a set of values. Measures of spread are typically used in conjunction with a measure of central tendency, such as the mean or median, to provide a more complete description of a set of data. In other words, a measure of spread gives you an idea of how well the mean, for example, represents the data. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small.
  - **Standard deviation** – a measure used to quantify the amount of variation or dispersion of a set of values. It is important to distinguish between the standard deviation of a population and the standard deviation of a sample, as these two standard deviations (sample and population standard deviations) are calculated differently. A smaller standard deviation indicates that the data points tend to be close to the mean, while a larger standard deviation indicates that the data points are spread out over a wider range of values. The standard deviation is often reported along with the mean to summarize interval/ratio data. *Note: Meaningful standard deviations can only be calculated from interval/ratio data that are roughly normally distributed (i.e. bell-shaped); quartiles (see following) are a better measure of spread for skewed distributions. Standard*

deviations may be of limited or no value for nominal/categorical and ordinal data, even where numbers are assigned.

- **Quartiles and Interquartile Range** – quartiles split an ordered data set into four equal parts, just like the median splits the data set in half. For this reason, quartiles are often reported along with the median. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2 (the median), and Q3, respectively. The interquartile range (IQR) is the difference between the third and first quartiles. *Note: Meaningful quartiles can only be calculated from ordinal and interval/ratio data. Quartiles may be of limited or no value for nominal/categorical data, even where numbers are assigned.*
- **Range** – the difference between the highest and lowest value for a particular variable. *Note: Meaningful ranges can only be calculated from ordinal and interval/ratio data. Ranges may be of limited or no value for nominal/categorical data, even where numbers are assigned.*

- **Correlation.** Correlation is a commonly used technique for describing the relationship between two quantitative variables. Correlation quantifies the strength and direction of the linear relationship between a pair of variables. An important thing to remember when using correlations is that a correlation does not explain causation. A correlation merely indicates that a relationship or pattern exists, but it does not mean that one variable is the cause of the other. As with other descriptive statistics, there are different types of correlations that correspond to different levels of measurement. For example, Pearson's product-moment correlation can be used to determine if there is a relationship or association between two interval/ratio variables, while Spearman's rank-order correlation can be used if one or both sets of data are ordinal.

While descriptive statistics can provide a summary that may enable comparisons across groups or units, every time you try to describe a set of observations with a single indicator (such as the mean or median) you run the risk of distorting the original data or losing important detail. Frequency distributions, means, and medians can tell very different stories, especially in the presence of extreme scores or skewed distributions. Consider the following example where a random sample of students completed a survey designed to assess student engagement.

Frequency/Percentage Distributions:

How much has your experience contributed to your knowledge and skills in the following areas?

	% (#) of students				
	Very much (5)	Quite a bit (4)	Some (3)	Very little (2)	Not at all (1)
Thinking critically	**5%** (3)	**18%** (12)	**62%** (40)	**11%** (7)	**5%** (3)
Writing clearly	**63%** (41)	**9%** (6)	**5%** (3)	**5%** (3)	**18%** (12)

Measures of Central Tendency & Spread:

How much has your experience contributed to your knowledge and skills in the following areas?

	Mean	Median	Mode	St Dev	Q1	Q3	IQR	Min	Max	Range
Thinking critically	3.1	3	3	0.8	3	3	0	5	1	4
Writing clearly	3.9	5	5	1.6	3	5	2	5	1	4

Looking at the previous frequency distributions, the majority of students (62%) said that their experience contributed *some* to their knowledge and skills related to "thinking critically". When you have roughly normally distributed data (i.e. bell-shaped), as in the previous responses for "thinking critically", you'll notice that the mean, median and mode are roughly equal. When the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data.

However if you look at the distribution of responses for "writing clearly", the majority of students (72%) said that their experience contributed to their knowledge and skills related to "writing clearly" *very much* or *quite a bit*. Additionally, you can see that 23% of students said that their experience contributed to their knowledge and skills related to "writing clearly" *very little* or *not at all*. While more than half of students responded *very much*, you might conclude that students typically answered *quite a bit* or *some* by looking at the mean alone. The 12 students who responded *not at all* severely affected the mean in this example. Looking at the median tells you that students typically answered *very much*; however, the median does not give you any indication of the variation in scores. Therefore, it can be helpful to consider frequency distributions in addition to measures of central tendency and spread. You'll want to know that most students felt their experience contributed to their ability to write clearly. But, you also want to know that 12 students said that their experience did not contribute to their ability to write clearly, so you can think of ways to address that.

Inferential Statistics

Descriptive statistics are typically distinguished from inferential statistics (such as from t-tests, ANOVA, chi square, etc.). Inferential statistics are produced by more complex mathematical calculations to reach conclusions that extend beyond the immediate data alone. In other words, inferential statistics can be used to generalize findings from sample data to make assumptions about the population at large given a representative sample and minimal sampling bias. For instance, inferential statistics may be used to examine the relationships between variables or differences between groups within a sample, and make generalizations or predictions about a larger population. Thus, we use inferential statistics to make inferences from our data to more general conditions.

While inferential statistics can be valuable when it is not convenient or possible to examine each member of the entire population, keep in mind that assessment is not controlled experimental research.

- Like research, assessment involves asking specific questions, using good practices, collecting and analyzing evidence, and evaluating results. Like research, assessment may use quantitative or qualitative methods, and often benefits from mixed methods.
- Unlike research, assessment lacks control of many outside variables that affect students and instruction, doesn't include a control group, and isn't intended to develop theories or test concepts. Many factors limit assessment, including limitations on time, resources, design and implementation.

Rather, assessment uses available time and resources to produce reasonably accurate information about student learning in the context of a particular program or institution, which can guide local practice or decisions. As a result, there may be no need for significance testing if there is no interest in making generalizations.

There are two common forms of significance testing:

- Using probability theory, **statistical significance** indicates whether a result is stronger than what would have occurred due to random error. To be considered significant, there must be high probability that the results were not due to chance.

- **Clinical significance** compares results to a pre-established standard that has been determined to be meaningful (such as national standardized test scores or program benchmarks). Clinical significance is sometimes seen as having more practical value, but only when there is a clear rationale for establishing the standards.

It is important to keep in mind that any significant differences – or lack thereof – may be due to factors beyond your control. If you are looking at historical trends and you see an improvement in skills, it may be that the improvement is simply because this year's class was better prepared than last year's class. Additionally, several factors influence the likelihood of significance, including the strength of the relationship, the amount of variability and bias/error in the data, and the size of the sample. For example:

- When examining large samples, significant testing can be misleading because even small or trivial effects are likely to produce statistically significant results.
- Statistical significance can be difficult to obtain when examining small sample sizes. Contrary to popular opinion, statistical significance is not a direct indicator of size of effect.

What is Effect Size?

An inferential test may be statistically significant (i.e., unlikely to have occurred by chance), but this doesn't necessarily indicate how large the effect is. The simple definition of effect size (sometimes referred to as *practical significance*) is the magnitude, or size, of an effect. There are many different ways to calculate effect size, examples include the correlation between two variables, the regression coefficient in a regression, the mean difference, or even the risk with which something happens. Unlike statistical significance, effect sizes are not influenced by sample size.

Recall that statistical computations and analyses assume that variables have specific levels of measurement and how, for example, it is nonsensical to calculate a mean for nominal level data. The assumptions for inferential statistics, though less obvious, are equally critical to the appropriate use of the statistical technique. Every inferential statistical test has assumptions that explain when it is and isn't reasonable to perform that specific test.

Additional Resources and References

- Field, A. (2009). Discovering Statistics Using SPSS, 3rd Ed. UK: Sage.
- Suskie, L. (2009). Assessing Student Learning: A Common Sense Guide, 2nd Ed. San Francisco, CA: Jossey-Bass.
- University of Hawaii Manoa Assessment Office. (2013). Making Sense of Assessment Data.
- Upcraft, M. and Schuh, J. (2002). Assessment vs. Research: Why We Should Care About the Difference. About Campus. 7(1): 16-20.

Assessment Data Stewardship: *It is important to remember that assessment data/results are valuable resources and must be carefully managed. Each individual with access to assessment data/results has the responsibility to use those data and any information derived from them appropriately. Non-public (i.e. internal or confidential) data/results should be labeled and only used to support assigned roles and duties. For more information on data stewardship, see ATL's Assessment Data Stewardship: Tips for Programs.*

Educational Research and Publication: *While assessment utilizes many of the same qualitative and quantitative methodologies applied by traditional social science research, this resources is NOT designed for those who are interested in generating/publishing scholarly research from their assessment activities or findings. Please contact ATL if you have questions about distinguishing the data needs of program-level assessment and educational research, or about sharing results from assessment.*