Beyond Rigor: Appropriate Analysis

Patricia B. Campbell, PhD
Eric J. Jolly, PhD

Data analysis with diverse populations involves a number of decisions about appropriate independent variables to use in the analysis as well as the levels of aggregation and disaggregation that should be done. Included here are some ways to make the analysis of quantitative and qualitative data both rigorous and responsive to the need to determine:

1 what works,
2. for whom it works, and
3. the context in which it works for different groups.[1]

Demographic Data as Independent Variables

Tip: When race/ethnicity, gender, or disability status is used as an independent variable, specify the reason for its use and include the reasons in documentation of the results.

Tip: When using a variable as a proxy for another variable, as in using educational level as a proxy for socio-economic status, indicate that the proxy is being used and include a rationale for why this is being done.

Rationale: In educational areas, evaluators may unconsciously accept a pattern of demographic differences in educational achievement or attainment as natural rather

[1] To determine what works for whom in what context, it is often necessary to include both qualitative and quantitative methods. Quantitative methods help provide the answer to what works for whom while qualitative methods are key for understanding the context.

This material is based upon work supported by the National Science Foundation under Grant No. 1146249. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Illustrations by Lee Abuabata

than looking for reasons to explain such a pattern. When group membership is accepted as an explanation for a pattern of performance, the truth may be distorted.[1]

Tip: When interpreting demographic differences, consider such conceptually relevant and possibly confounding factors as socioeconomic status, individual and family educational backgrounds, immigrant status, and place of residence. Where possible include statistical controls.

Rationale: There are large differences by race/ethnicity and disability status in a variety of areas. For example, in 2010, the Census Bureau reported the median household net worth for Whites was $110,729, versus of $7,424 for Hispanic households and $4,995 for Black households.[2] Results from the 2006 American Community Survey found significant disparities in the median incomes of those with and without disabilities. Median earnings for people with no disability were over $28,000 compared to the $17,000 median income reported for individuals with a disability.[3] Unless such differences are addressed, in any analysis, there is a great danger of generating inaccurate conclusions.

Levels of Aggregation and Disaggregation

Tip: Use crosstabs to break down the demographic characteristics of participants to help determine where levels of disaggregation can be done. If almost all of the participants are from one demographic subcategory, for example, the middle class, then it is not necessary nor perhaps appropriate to disaggregate data for analysis by socioeconomic status. However the results could not then be generalized to other socioeconomic groups.

Rationale: Cell size is an important factor in determining areas of disaggregation. If you are disaggregating by gender and by race/ethnicity, then the number of White women would be one cell, the number of Black men would be another cell, and so on. If some cell sizes are very small at a specific level of disaggregation, then statistical analysis at that level may not be appropriate. For example, a nonparametric statistical test such as a chi square test requires a minimum expected value of 5 for each cell in order to be valid. For parametric statistical tests such as t-tests, a cell size of 30 is considered the minimum. In addition in parametric statistics, the smaller the sample or cell size, the more difficult it is to get statistical significance.[4] The probability that a significant result will be obtained if a real difference exists (which is called the power of the test) depends largely on the total sample size. Based on the size of the sample, you can compute the power of the test. Alternatively, if you know what power you want the test to have, you can compute the needed sample size to get that power.[5]

Tip: If there are known or expected differences by subgroup that could skew the overall findings, then disaggregate by those subgroups.

Tip: Be aware that there can be heterogeneity within subgroups. For example, while people who are visually impaired, hearing impaired, and learning disabled are all classified

as having disabilities, the differences among them are very large and it might be appropriate to disaggregate by different categories of disability.

Tip: Do preliminary analysis of subgroup differences in areas of importance to the study. This can help inform disaggregation and aggregation decisions.

Rationale: There may be some cases where disaggregation is not needed. As Jolly explained, "when you mix ammonia and Clorox in a room, everyone gets sick. At that level it is not necessary to disaggregate."[6] There are other cases where project/program impact may vary for different subgroups and disaggregation is needed. For example, since women students tend to exhibit lower skills in some spatial areas, such as 3-D rotation, that are important to success in engineering,[7] projects/programs tying improved spatial skills to increased retention in engineering should disaggregate their data by sex to determine if there are statistically significant interactions between sex and impact of program participation in a spatial skills training program. Other possible areas that need to be considered in decisions about disaggregation include:

- Class, race, ethnic, and gender differences in diagnosis of learning disabilities;[8]
- Race and ethnic differences in retention in undergraduate STEM academic programs;[9]
- Differences in retention to degree for two year college transfers;[10]
- Race and ethnic differences in time to STEM degree;[11]
- Differences in STEM preparation including:
    - course taking;
    - achievement;
    - participation STEM programs;
    - work experience.

Tip: Provide a rationale for the decisions made regarding which demographic categories are aggregated and which are disaggregated.

Rationale: For statistical reasons and often for confidentiality, some aggregation of data needs to be done even though information will be lost in each aggregation. For example, when aggregation is done across disability groups, the ability to determine if a project/program has different impacts on people with learning disabilities and people with mobility impairment is lost. Another example shows the aggregation of Native Americans with other groups because of their small numbers means that little is known about Native Americans and STEM.

While evaluators must assume responsibilities for capturing and correctly interpreting within-group variability for the groups under study,[12] types of disaggregation must be both meaningful and viable. If, for example, there is an interest in trend data, aggregation across years is not appropriate. If there is reason to think there might be different trends for different subgroups it is important to disaggregate by those subgroups. Since there are gender differences in some spatial skills, if you are interested in the impact of a project/program to improve spatial skills, then it is important to disaggregate by gender. Based on the questions to be answered, it might

be more appropriate to aggregate across subdisciplines, across institutions, across years, or across some racial/ethnic or disability categories.

Statistical Significance

Tip: Do not report results as "approaching statistical significance."

Rationale: Approaching statistical significance means that a result did not meet the predetermined level of significance, which is usually .05, but was close. In some ways it is a "nicer way of admitting that your results support the null hypothesis".[13] While it is tempting to report a result as approaching significance, it is problematic. A decision was made regarding the acceptable standard and now, based on the results, that decision is being modified to fit the data. This is a violation of the assumptions behind inferential statistics. The definition of "close" is problematic since it is based on what was found rather than any scientific rationale. If results are approaching significance it might be possible to collect data from more participants until the results are significant or until the significance gets worse. Another option would be to replicate the study.[14]

Tip: Limit the number of statistical tests that you do and have a rationale for each test that you do.

Rationale: Since statistical significance testing is based on a probability, the more statistical tests you do, the more likely you are to get statistically significant results that are incorrect.[15] This cumulative error rate means that it is inappropriate to do 100 tests and then report as significant the five comparisons reaching the .05 level because of the dangers of capitalizing on chance. The major exception would be if the tests were intended to be exploratory and would be used only in a subsequent independent study to help generate hypotheses.

Tip: If statistical significance is found, check for effect size and, as needed, use a website that does effect size computations.[16]

Rationale: A finding of statistical significance means that the null hypotheses—the hypotheses of no difference—has been rejected. By itself, it does not mean that the findings are important or meaningful. It is important to determine the size or magnitude of the effects found. Effect sizes can be computed for group differences and for correlations.[17] One well known effect size is Cohen's *d*. Cohen defined effect sizes as "small, $d = .2$," "medium, $d = .5$," and "large, $d = .8$," but warned of the risks inherent in defining the terms in as "diverse a field of inquiry as behavioral science."[18] Larger effect sizes indicate that not only is a difference statistically significant but the likelihood is that the difference is meaningful as well.

Tip: Be sure that the statistical tests being used are the right ones for the data. This may involve using an online tool such as the one developed by the UCLA: Statistical

Consulting Group[19] or hiring a statistical consultant to provide assistance in the selection of appropriate tests.

Rationale: Data can be analyzed in multiple ways, each of which could yield legitimate answers. There are a number of factors that determine whether an analysis is appropriate including the number of dependent (or outcome) variables; the nature of the independent (or predictor) variables; whether your dependent variable is an interval variable, ordinal, or categorical variable[20]; and whether it is normally distributed.[21] Use of an inappropriate statistic can lead to inaccurate results.

Coding/Rating Open-Ended Responses

Tip: Make the coding/rating procedures as anonymous as possible. Have participants put their names and any other identifying information at the end of their responses or on a separate page so coders/raters don't see it.

Rationale: Knowing the gender, race/ethnicity, and even the first name of the person whose work is being rated has been shown to impact ratings of:

- open ended responses;
- research work;
- faculty evaluations;
- resumes; and
- course work.[22]

People were found to rate statements as less true when they were spoken by non-native speakers[23]; while academics rated job applications for lab managers and instructors with male names higher than the identical applications with female names, (although there were no differences in their ratings of tenure applications[24]).

Tip: Have more than one coder/rater code the responses and check for inter-coder/rater reliability. If the reliability is not high enough, have coders/raters discuss the rationale behind their ratings and recode.

Tip: Unless the coding is being based on grounded theory, develop and test the coding/rating protocol in advance of doing the data analysis.

Rationale: While some question whether it is possible to generate reliable codings/ratings of open-ended responses, others feel that inter-coder/rater reliability is a useful concept in settings characterized by applied, multidisciplinary, or team based work. Establishing high inter-coder/rater reliability is an attempt to reduce error and bias.[25]

Qualitative Data Validations

Tip: Have participants review a summary of the results for credibility. For example, after a focus group discussion, the facilitator can quickly summarize the major takeaways from the discussion, which participants can, at that time, validate.

Tip: Thoroughly describe the evaluation context and the assumptions that were central to the evaluation.

Tip: Describe any critical changes that occurred in the project/program and related areas (e.g., the hiring of a new high level staff member) and how these changes affected the way the evaluation team approached the study.

Tip: Have another evaluator take a "devil's advocate" role and actively search for and describe any negative instances during the data collection or in the data that contradict reported prior observations.[26]

Rationale: Criteria for judging the quality of quantitative research focus on validity and reliability; however some qualitative researchers, most famously Guba and Lincoln[27], argue that there are different criteria for judging the quality of qualitative research. These are credibility, transferability, dependability, and confirmability. While there has been debate among methodologists about the value of an alternative set of standards for judging qualitative research, since qualitative research cannot easily be considered an extension of the quantitative paradigm into the realm of nonnumeric data, it is important to consider these standards in the assessment of qualitative analysis.[28]

Comparison Groups

Tip: Make comparisons across more and less effective projects/programs. Factors common across effective programs may also be common across ineffective projects/programs.

Rationale: If only exemplary projects/programs are included, it is not possible to determine if “effective” characteristics identified by the research are indeed unique to these positive outliers. Comparison groups of other projects/programs are needed to see what is unique to effective projects/programs. For example, one study comparing effective schools to typical schools found both were using similar curriculum.[29]

Tip: When using existing data sets for comparison groups, determine the process that was used to categorize participant race/ethnicity and disability status and then use a similar process to categorize your participants to ensure data are comparable.

Rationale: If data sets use different methods to categorize race/ethnicity and disability status, cross data set comparisons will not be valid. For example, in the Survey of Earned Doctorates (SED) a person can define themselves as non-Hispanic, Mexican/Chicano, Puerto Rican, Cuban, or Other Hispanic as well as by race.[30] In the Integrated

Postsecondary Education Data System (IPEDS), one can identify as non-Hispanic or Hispanic and by race. However, aggregate data reported to IPEDS has Hispanics of any race as one category and only non-Hispanics are reported by race.[31] Also, SED[32] asks about five broader areas of disability while IPEDS[33] reports on 11 areas of disability.

Confounding Events

Tip: If there have been extreme external events prior to or during the study period, check for possible impact of those events in the analysis.

Tip: Explore possible alternative explanations for both positive and negative findings.

Rationale: Unexpected outcomes from such natural disasters such as 2005's Hurricane Katrina and 2012's Superstorm Sandy closed schools and in many cases meant students needed extensions of deadlines to complete their work.[34] While these are extreme events, there may be other events that may influence the data and skew the results so they could not be accurately applied to a program under more normal circumstances. For example, in difficult economic times, staying employed with no promotion could be a positive outcome for a training program.

---

[1] Nelson-Barber, S., LaFrance, J., Trumbull, E., & Aburto, S. (2005). Culturally-responsive program evaluation. In S. Hood, R. Hopson, & H. Frierson (Eds.), *The role of culture and cultural context: A mandate for inclusion, the discovery of truth and understanding in evaluative theory and practice* , (pp. 61–85). Greenwich, CT: Information Age Publishing.

[2] Luhby, T. (2012). Worsening wealth inequality by race. CNN Money http://money.cnn.com/2012/06/21/news/economy/wealth-gap-race/index.htm?hpt=hp_t2

[3] American Psychological Association. (2013). Disability & Socioeconomic Status. Retrieved from http://www.apa.org/pi/ses/resources/publications/factsheet-disability.aspx

[4] Mehta, C. R., & Patel, N. R. (2011). IBM SPSS Exact tests. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Exact_Tests.pdf

[5] Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials 2*, 93–113. ftp://maia.biostat.wisc.edu/pub/chappell/641/papers/paper28.pdf

[6] Jolly, E. J. (9/07/12). Personal communication.

[7] Metz, *S. S.*, Donohue, *S. &* Moore, *C. (*2012*).* Spatial skills: A focus on gender and engineering. In B. Bogue & E. Cady (Eds.), *Apply Research to Practice (ARP) Resources*. http://www.engageengineering.org/associations/11559/files/ARP_SpatialSkills.pdf

[8] Center for Disease Control. (2011). Percentage of Children Aged 5–17 years ever receiving a diagnosis of learning disability, by race/ethnicity and family income group - National Health Interview Survey, United States, 2007–2009. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6025a6.htm

[9] *Science and Engineering Indicators 2012*. (2012). Arlington, VA: National Science Foundation. http://www.nsf.gov/statistics/seind12/c2/c2s2.htm#s3
National Academy of Sciences (NAS), *Expanding Underrepresented Minority Participation: America's Science and Technology Talent at the Crossroads* 2011, Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=12984

[10] National Student Clearinghouse Research Center (Spring, 2012) Snapshot report: Mobility. http://www.studentclearinghouse.info/snapshot/docs/SnapshotReport6-TwoYearContributions.pdf

National Student Clearinghouse Research Center (Spring, 2012) Snapshot report: Degree attainment. http://www.studentclearinghouse.info/snapshot/docs/SnapshotReport8-GradRates2-4Transfers.pdf

[11] Bell, N. (2010, March). Research report on data sources: Time-to-degree for doctorate recipients. *Communicator*, 1–3. Washington, D.C.: Council of Graduate Schools. Retrieved from http://www.cgsnet.org/ckfinder/userfiles/files/DataSources_2010_03.pdf

Huang, G., Taddese, N., & Walter, E. (2000). *Entry and persistence of women and minorities in college science and engineering education* (No. NCES 2000601). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2000601

[12] Nelson-Barber, S., LaFrance, J., Trumbull, E., & Aburto, S. (2005). Culturally-responsive program evaluation. In S. Hood, R. Hopson, & H. Frierson (Eds.), *The role of culture and cultural context: A mandate for inclusion, the discovery of truth and understanding in evaluative theory and practice*, (pp. 61–85). Greenwich, CT: Information Age Publishing.

[13] PSYCHblog (2012). Approaching significance. http://psychbloglD.wordpress.com/2012/03/10/approaching-significance/

[14] Stassam (2012). Is the term "approaching significance" cheating? http://statssam.wordpress.com/2012/03/11/is-the-term-approaching-significance-cheating/

[15] Hopkins, W.G. (2000). *A new view of statistics*. http://www.sportsci.org/resource/stats/errors.html

[16] www.uccs.edu/~lbecker/

[17] www.uccs.edu/lbecker/effect-size.html

[18]Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates. p 25.

[19] What statistical analysis should I use? UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/mult_pkg/whatstat/

[20] What is the difference between categorical, ordinal and interval variables? UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/mult_pkg/whatstat/nominal_ordinal_interval.htm

[21] What statistical analysis should I use? UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/mult_pkg/whatstat/

[22] Anderson-Clark, T., Green, R. & Henley, T. (2008). The relationship between first names and teacher expectations for achievement motivation. *Journal of Language & Social Psychology*, *27*(1), 94–99. doi:10.1177/0261927X07309514 (http://dx.doi.org/10.1177/0261927X07309514)

Correll, S. J., & Benard, S. (2006). Biased estimators? Comparing status and statistical theories of gender discrimination. In S. R. Thye and E. J. Lawler (eds.), *Advances in group processes* (vol 23, pp. 89–116). New York, NY: Elsevier.

Pellegrini, A. D. (2011). "In the eye of the beholder": Sex bias in observations and ratings of students' aggression. *Educational Researcher*, *40*(6), 281–286. doi:10.3102/0013189X11421983 (http://dx.doi.org/10.3102/0013189X11421983)

[23] Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. doi:10.1016/j.jesp.2010.05.025 (http://dx.doi.org/10.1016/j.jesp.2010.05.025)

[24] Moss-Racusin, C. A., Dovidio, J. F, Brescoll, V. L., Graham, M. J., & Handselsman, J. (2012). Science faculty's subtle gender biases favor male students. *PNAS Early Edition*. 1–6. http://www.pnas.org/content/early/2012/09/14/1211286109.full.pdf+html

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7/8), 509–528. doi:10.1023/A:1018839203698 (http://dx.doi.org/10.1023/A:1018839203698)

[25] Hruschka, D.J., Schwartz, D., St. John, D.C., Picone-Decaro, E, Jenikns, R. A. & Carey, J.W. (2004). Field Methods, 16, 3, 307-331. http://www.analytictech.com/mb870/readings/hruschka.pdf

[26] Trochim, W.M.K. (2006). Qualitative validity. Research Methods Knowledge Base. http://www.socialresearchmethods.net/kb/qualval.php

[27] Guba, E. G., & Lincoln, Y. S. (2005). "Paradigmatic controversies, contradictions, and emerging influences" In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage Handbook of Qualitative Research* (3rd ed.), pp. 191-215. Thousand Oaks, CA: Sage.

[28] Trochim, W.M.K. (2006). Qualitative validity. Research Methods Knowledge Base. http://www.socialresearchmethods.net/kb/qualval.php

[29] Clewell, B. C. & Campbell, P. B., with Perlman, L. (2007). *Good schools in poor neighborhoods: Defying demographics, achieving success.* Washington, DC: Urban Institute Press
[30] NORC. (2009). SED (Survey of Earned Doctorates) [survey instrument]. http://www.nsf.gov/statistics/srvydoctorates/surveys/srvydoctorates_2010.pdf
[31] Institute of Education Sciences (IES)/National Center for Education Statistics. (2007). Memo: Changes to race/ethnicity reporting to the integrated postsecondary education data system (IPEDS) as mandated by the U.S. Department of Education. Retrieved from http://nces.ed.gov/ipeds/news_room/ana_Changes_to_10_25_2007_169.asp
[32] NORC. (2009). SED (Survey of Earned Doctorates) [survey instrument]. http://www.nsf.gov/statistics/srvydoctorates/surveys/srvydoctorates_2010.pdf
[33] Raue, L., Lewis, L., & Coopersmith, J. (2011). NCES 2011–018, students with disabilities at degree-granting postsecondary institutions: First look [report and data]. http://nces.ed.gov/pubs2011/2011018.pdf
[34] Pappas, S. (October 31, 2012). Sandy wiped out NYU lab mice, dealing blow to medical research. *LiveScience*. http://www.livescience.com/24443-nyu-lab-mice-drowned-hurricane-sandy.html

Campbell-Kibler Associates, Inc.
80 Lakeside Dr
Groton, MA 01450
www.campbell-kibler.com

Science Museum of Minnesota
120 West Kellogg Blvd.
St. Paul, Minnesota 55102
www.smm.org