Judgment in Quantitative Research

Carl J. Huberty

Introduction

The use of subjective judgment is commonplace in the lives of all people. In particular, it is often used by mathematics education quantitative researchers (whether or not it has been so indicated). What is to be addressed in this article is the use of subjective judgment in research that involves statistical/quantitative methods. [The use of subjective judgment in research that involves qualitative methods is another story for another day.]

What is *judgment*? Often used interchangeably are *personal judgment, subjective judgment, personal informed judgment*, and *subjectivity*. A definition has been advanced by Yates (1990): "A *judgment* is an opinion about what is (or will be) the status of some aspect of the world" (p. 6). Meyer and Booker (1991) define *expert judgment* as "data given by an expert in response to a technical problem" (p. 3).

Judgments made with respect to the four aspects of quantitative research discussed below involve making *decisions*. Such decisions often call for the use of *common sense*—once defined as the set of prejudices acquired by age eighteen—in combination with informed judgment. Abelson (1995, pp. 176-178) briefly discusses the role of common sense in making decisions. Highly related to the role of judgment in quantitative research is the role of personal values (see Pedhazur & Schmelkin, 1991, pp. 159-160, 207, 209).

It may be of some interest to note that informed/expert judgment is a topic of at least four fairly recent books (Cooksey, 1996; Lad, 1996; Meyer & Booker, 1991; Yates, 1990). The comprehensive book by Cooksey (1996) provides a detailed, authoritative discussion of the (cognitive) theory and applications of methods of judgment in quantitative research paradigms. Early in their book, Meyer and Booker (1991, pp. 4-5) discuss three needs of using what they term expert judgment: (a) to provide estimates, (b) to forecast future events, and (c) to integrate or interpret existing data. Some history on the development of the subjectivity of science, especially with respect to probability, is discussed by Lad (1996, pp. 19-37). Yates (1990) presents judgment as a "partner" with decision making; in sum, he states:

> This book is about human decision making, particularly in the presence of uncertainty. Special emphasis is placed on one of the most significant contributors to decision behavior -- judgment. Shortcomings in judgments are a prime example of decision errors, which are specific behaviors that are responsible for failed decisions. (p. 11)

What will be discussed in this article is the use of subjective judgment with respect to four aspects of quantitative research: (1) design, (2) preliminary analyses, (3) general analyses, and (4) specific analyses. Scattered throughout the discussions will be references that support such use, and other references that suggest support for objectivity as opposed to subjectivity in the research process. The article is concluded with a section including comments related to judgment and textbooks, experience, and disagreements as well as some relevant quotes.

Judgment in Design

Design issues involving judgment include the selection of variables and analysis units to be studied. Given a research question of interest, decisions may need to be made regarding the grouping variables and the levels of such. If a new "treatment" or two are to be studied, not much judgment will be needed to decide on the treatment levels. If, however, something like cognitive ability is to be considered as a grouping variable, some judgments will clearly have to be made as to the number of levels and what the cognitive ability cutoffs are. A specified research question may also suggest what

Carl J. Huberty is a professor in the Department of Educational Psychology at the University of Georgia. He teaches statistical methods courses from the intro course up through multivariate methods, does design and data analysis work on grants, and writes manuscripts dealing with statistical methods and ideas. His email address is chuberty@coe.uga.edu

response variable(s) to include in the study. This decision would also call for some judgment, as would the decision of how to measure the response variable(s).

How to select a sample of analysis units would also involve some judgment. Does one use representative sampling, simple random sampling, convenience sampling, or another sampling plan? Sometimes cost restrictions impose some judgment limitations on the sampling plan. How about the sample size(s)? If research costs are not a big issue, then statistical power may play a role in the decision-making. To employ power tables, one has to decide on the seriousness of statistical decision errors and on an expected effect size (see Brewer, 1991). Such judgments are not trivial, and may require some experience. A brief discussion of the relationship between experience and judgment is given in a subsection of the last section of this article.

Judgment in Preliminary Analyses

It is assumed at this point in the research process that all data have been collected. A first step in the analysis phase is to "look at your data." Now, for what does one look? One thing to look for are missing data. In a multiple response variable context a judgment would need to be made as to what is a "large" percent of analysis units on which measures on one response variable are missing. That issue aside, a decision would have to be made regarding the imputation method used to fill in the data gaps (see, e.g., Rencher, 1998, pp. 23-27; Roth, 1994). Another thing to look for is analysis units with outlying scores. Determining if a score is "far out" is a judgment call. [There is no formal, specific, general, agreed-upon empirical definition of an outlier to be used with all data sets!] One also has to decide whether to delete identified outliers, or to delete them one at a time, or to delete them simultaneously.

Once the final data set is determined, more data examination is suggested. Descriptive information on the final data set is often desirable and informative. In additional to numerical information, graphical presentations are often very informative. One type of such presentation is a box-plot. The interpretation of box-plots calls for some judgments. No matter how experienced the researcher is, box-plot perception judgments may be erroneous (Behrens, Stock, & Sedgwick, 1990). Another type of presentation that involves additional researcher judgment is the smoothing of data point curves (Tukey, 1977, chaps. 7, 16).

The last example of a set of preliminary analyses that would call for researcher judgment pertains to assessment of requisite data analysis conditions. Judgments need to be made when considering many conditions (e.g., score independence across analysis units, score distribution form), but covariance matrix homogeneity will be used as an example. Statistical tests of covariance matrix equality are very powerful. Therefore, information in addition to statistical test information should be considered. The information of interest are two alternative indexes of *generalized variance*: the logarithm of the determinant of a covariance matrix, and the trace of a covariance matrix. So, in examining the equality of, say, three covariance matrices, one can eye-ball the equality of the three determinant logarithms and of the three traces. The eye-balling of equality of the two sets of numbers clearly calls for some researcher judgment. See Huberty and Petoskey (in press) for an example of this judgment.

Judgment in General Analyses

The analysis context in which nearly all of the comments are made in this article is that of classical methods, as opposed to Bayesian methods. For discussions of the use of subjectivity in connection with a Bayesian analysis see Pruzek (1997) and Wang (1993, pp. 153-166). The book by Jeffrey (1992), a Bayesian statistician/philosopher, contains 16 of his writings, one of which (used as the book title) pertains explicitly to judgments about probability statements. Giere (1997) discusses philosophical differences and similarities between the Bayesian point of view of the use of subjectivity and the classical point of view. In reviewing differing points of view of two books in particular, Giere (1997) states that classical methods "are no less arbitrary (or subjective) than Bayesian methods. In fact, ... classical methods are even more arbitrary (or subjective) than Bayesian methods" (pp. S182-S183).

From now on in this article I will focus on the so-called classical data analysis methods. In a very basic context, Brewer (1991) adeptly discusses the vast array of subjective judgments made when conducting and interpreting statistical testing and interval estimation.

With these methods one must decide on, and make judgments about, the magnitude of *p*-values, the magnitude of an effect-size index, and the confidence level to use. [There are at least two methodologists who take exception to the use of some researcher judgment in statistical testing. Conger (1984) contends that "alpha levels are to a great extent dictated by the general scientific community; little freedom of choice exists” (pp. 291-292). Frick (1996) concludes that "it is undesirable for the outcome of a statistical test to depend on subjective, arbitrary, or possibly biased choices by the experimenter" (p. 386).] A beginning decision in a statistical test situation pertains to the type of test to be conducted. That is, should a normal-based or nonparametric test be used? Although this decision is mentioned by Brewer (1991), subjective judgments in this decision-making process are discussed in some detail along with two other contexts by Stewart-Oaten (1995)—the two other contexts are multiple comparisons and sums of squares for (nonorthogonal) ANOVA tests.

Huberty and Pike (in press) review some of the pros and cons of using subjective judgment in the statistical testing process. They mention that the use of judgment was advocated by the statistical testing pioneers (R.A. Fisher, J. Neyman, E.S. Pearson). Pearson (1962) emphasized "a more intuitive process of personal judgment... as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities" (p. 396). More recent advocacies of the use of judgment in the statistical testing process are advanced by Cohen (1994) and Cortina and Dunlap (1997). A simulation by Marsh and Hau (1996) to study the evaluation of goodness of fit indexes for structural equation models provided some validation of a conclusion of Bollen and Long (1993): "...test statistics and fit indices are very beneficial, but they are no replacement for sound judgment and substantive expertise" (p. 8). This conclusion is questioned by Markus (1998).

Some judgments are research-context-specific. For example, for a very "large" sample size, extent of variable effect (as measured by a *p*-value, effect size index value, or any other numerical index value) may be different from that for a lesser sample size. As another example, the determination of a "real" effect may depend upon the substantive area of study (see, e.g., Rosnow & Rosenthal, 1988).

Judgment in Specific Analyses

As I indicated earlier, the potential for the use of subjective judgment is very high in statistical data analyses. Some specific analysis situations where judgments are called for will now be stated in the form of questions. In the single outcome variable context, examples of questions are:

- What contrasts should be examined in a *k*-group comparison study?
- What error sum of squares should be used with a nonorthogonal design?
- What is a "small" distribution tail-area (probability)?
- What is a "large" effect-size index value?
- Should analysis-of-variance main effects be examined in the face of real interaction effects?
- Should a normal-based analysis be conducted?
- In a two-group situation, should a directional alternative hypothesis be used?
- Should a concomitant variable be included?
- Is the sample size large enough to yield adequate statistical power?
- Should a blocking variable be employed?
- How are test tail area values adjusted for multiple testing?
- Is the X-Y relationship linear?
- What analysis approach should be used with repeated outcome variable measures?

In a multiple response variable context, decisions regarding many more questions need to be made. Some examples are:

- What is a good fit in structural-equation modeling?
- Are the group covariance matrices really different?
- What is the rank-order of the outcome variables in a descriptive discriminant analysis?
- What is the rank-order of the predictors in a predictive discriminant analysis?
- What is a "good" C_p index value in a multiple regression analysis?
- What is a meaningful definition of the linear composite of some response variables in a multiple correlation study? In a multivariate analysis of variance study? In a principal component analysis?
- What index of similarity is appropriate for a cluster analysis of a particular data set?

- What numbers should be used as prior probabilities of group membership in a predictive discriminant analysis?
- For a given data set, what is the "optimal" number of clusters?
- What response variables should be deleted in contexts of predictive discriminant analysis and multiple regression analysis?
- For a given data set, what is the "optimal" number of components/factors?
- For a given data set, what is the "optimal" number of pairs of composites for a canonical correlation analysis?
- To what extent is your sample representative of the intended population?
- What response variables are to be initially chosen?
- If an initial data reduction analysis is called for, what analysis should be used?
- How should unordered categorical response variables be scaled?
- Which approach should be used to impute missing response variable scores?
- When considering response variable deletion, should some response variables be "forced" in the final subset?
- Should the response variable measures be standardized for a cluster analysis?
- What is a meaningful description of resulting clusters?
- How is it determined if an analysis unit does not clearly belong to one group or another in a predictive discriminant analysis?
- How many linear discriminant functions should be retained for interpretation purposes in a descriptive discriminant analysis?
- What method of composite extraction should be used while exploring response variable structure?
- What is a "high" structure *r*?

Comments

Textbooks

It is surprising and disappointing that very few statistical methods textbooks encourage, or even mention, the use of subjective judgment in the quantitative research process. Five types of recent methods books were examined regarding the mention of subjective judgment: introductory (20 books), analysis of variance (12), multiple correlation and regression (13), applied multivariate (23), and books of readings (7). Save complete reading of each of the 75 books, each was checked by scanning the table of contents and the subject index. It was found that only two of the 75 books mentioned the use of subjective judgment (Hair, Anderson, Tatham, & Black, 1995, pp. 487, 505; Huberty, 1994, pp. 22-23, etc.). It is recognized that textbook indexes and tables of content are not all-inclusive when it comes to the use of terms/concepts.

Experience Base

In at least some quantitative research situations, one might think that researcher experience related to the study and analysis would aid in the judgment process. Experience is generally considered a positive contributor to making judgments and decisions (Yates, 1990, pp. 372-375). This very idea has been espoused by a number of writers/methodologists, for example:

- "...judgment is based on experience, previous research..." (Schaafsma & van Vark, 1979, p. 108).
- "In general we encourage researchers in the soft sciences to take into account experience and intuition in their inquires" (Wang, 1993, p. 158).
- "...there is no statistical substitute for the knowledge and experience of the researcher" (Murray & Dosser, 1987, p. 72).

At the same time, Payne, Bettman, and Luce (1998) conclude that "experience does not necessarily improve judgment" (p. 330).

Disagreements

Relying on judgment in the quantitative research process is not a negative. Of course, the researcher should admit that judgments were utilized and provide some rationale (based on experience, literature, common sense, intuition, etc.) for the judgments. This would *not* be an admission of guilt! Rather, it would be "setting the record straight." Sure, some readers will disagree with judgments made; and some of the disagreements may be justified. An example of a disagreement about the sampling methods used in the Third International Mathematics and Science Study reported in 1998 was given by Rothberg (1998).

Most judgment "errors" made in the quantitative research process pertain to disagreements between the writer and the reader on the quality of the methods used, on the selections made, and on the magnitudes of numerical indexes. We consider these as *disagreements* as opposed to errors. Lad (1996) claims that it is a "fact that reasonable people can and do disagree in their analyses and conclusions" (p. 1).

Some quotes

Selected quotes are now offered in conclusion:

- "...there is no theory of hypothesis testing that can be applied in a purely formal or mechanical way, without informed personal judgment....(J.) Neyman and (E.S.) Pearson reconsidered inductive inference as decision making, where statistical theory and personal judgment must be interlocked" (Gigerenzer & Murray, 1987, p. 16).
- "(C.F.) Gauss (1827), however, was disinclined to give any more directions, and in his answer he compared the situation to every day life, where one often makes intuitive judgments outside the reign of formal and explicit rules" (Gigerenzer et al. 1989, p. 83).
- "Scientific activity necessarily involves the personal judgment and belief of its participants, ..." (Lad, 1996, p. 12).
- "The good statistician must be prepared to use his subjective judgment where necessary to modify the results of a formal statistical analysis" (Chatfield, 1995, p. 119).
- "One of the tasks of logical positivism is to outlaw all speculative statements as meaningless. A consequence of this "scientific philosophy" is a disrespect of intuitive and subjective knowledge….human knowledge is not acquired by objective methods..." (Wang, 1993, p. 153).
- "Remember that throughout the process in which you conceive, plan, execute, and write up a research, it is on your informed judgment as a scientist that you must rely, and this holds as much for the statistical aspects of the work as it does for all the others. This means that your informed judgment governs the setting of the parameters involved in the planning ..., and that informed judgment also governs the conclusions you will draw." (Cohen, 1990, p. 1310)

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Behrens, J. T., Stock, W. A., Sedgwick, C. (1990). Judgment errors in elementary box-plot displays. *Communications in Statistics-Simulation, 19*, 245-262.

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park, CA: Sage.

Brewer, J. K. (1991). Subjectivity in statistical inference. *Florida Journal of Educational Research, 31*, 5-12.

Chatfield, C. (1995). *Problem solving: A statistician's guide.* London: Chapman & Hall.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Conger, A. J. (1984). Statistical considerations. In M. Hersen, L. Michelson, & A. S. Bellack (Eds.), *Issues in psychotherapy research* (pp. 285-309). New York: Plenum.

Cooksey, R. W. (1996). *Judgment analysis.* San Diego, CA: Academic Press.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*, 161-172.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379-390.

Giere, R. N. (1997). Scientific inference: Two points of view. *Philosophy of Science (Proceedings), 64,* S180-S184.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance.* New York: Cambridge University Press.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis.* Englewood Cliffs, NJ: Prentice Hall.

Huberty, C. J (1994). *Applied discriminant analysis.* New York: Wiley.

Huberty, C. J, & Petoskey, M. D. (in press). Multivariate analysis of variance and covariance. In H.E.A. Tinsley & S. Brown (Eds.), *Handbook of multivariate statistics and mathematical modeling.* Orlando, FL: Academic Press.

Huberty, C. J, & Pike, C. J. (in press). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology,* Vol. 5. Greenwich, CT: JAI Press.

Jeffrey, R. (1992). *Probability and the art of judgment.* New York: Cambridge.

Lad, F. (1996). *Operational subjective statistical methods.* New York: Wiley.

Markus, K. A. (1998). Judging rules. *The Journal of Experimental Education, 66,* 261-265.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64,* 364-390.

Meyer, M. A., & Booker, J. M. (1991). *Eliciting and analyzing expert judgment.* London: Academic Press.

Murray, L. W., & Dosser, D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34,* 68-72.

Payne, J. W., Bettman, J. R., & Luce, M. F. (1998). Behavioral decision research: An overview. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 303-359). San Diego, CA: Academic Press.

Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics, 33,* 394-403.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis.* Hillsdale, NJ: Erlbaum.

Pruzek, R. M. (1997). An introduction to Bayesian inference and its application. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 287-318). Mahwah, NJ: Erlbaum.

Rencher, A. C. (1998). *Multivariate statistical inference and applications.* New York: Wiley.

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 35,* 203-208.

Rothberg, I. C. (1998). Interpretation of international test score comparisons. *Science, 280,* 1030-1031.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537-560.

Schaafsma, W., & van Vark, G. N. (1979). Classification and discrimination problems with applications. Part Iia. *Statistics Neerlandica, 33,* 91-126.

Stewart-Oaten, A. (1995). Rules and judgments in statistics: Three examples. *Ecology, 76,* 2001-2009.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wang, C. (1993). *Sense and nonsense of statistical inference.* New York: Dekker.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

TME Subscriptions

Subscriptions will be required in order to receive the printed version of *The Mathematics Educator* starting with Volume XI, Number 1 (Winter 2001).

To subscribe, send a copy of this form, along with the requested information and the subscription fee to

The Mathematics Educator
105 Aderhold Hall
The University of Georgia
Athens, GA 30602-7124

I want to subscribe to *The Mathematics Educator* for Volume XI (Numbers 1 & 2).

Name ______________________________________

Address_____________________________________

____

____

Amount Enclosed____________
($6/individual; $10/institutional)