

Inter-Rater Agreement in Analysis of Open-Ended Responses: Lessons from a Mixed Methods Study of Principals¹

Will J. Jordan and Stephanie R. Miller
Temple University

Objectives

The use of open-ended survey questions in social science is appropriate when researchers investigate complex issues lacking a fixed, predetermined set of categories. The existing literature affords some guidance in applying systematic methods for analyzing written responses generated through open-ended survey items and other qualitative data collection procedures (Bernard, 1994; Carey, Morgan and Oxtoby, 1996; Miles and Huberman 1994; and Patton 1990). There is general consensus among scholars that taking explicit steps to increase rater agreements has methodological benefits and increases the scientific credibility of a study (Creswell, 2003). We contend that rigorous inter-rater reliability analysis might be necessary to guard against the introduction of subjective bias in the coding and analysis of qualitative data. This paper describes the analytic procedures we used to empirically measure the degree of consistency between multiple coders of qualitative data in a large-scale study of elementary and middle school principals. A methodological and pragmatic challenge we faced was that agreement is reached through an interpretive process, the diverse lenses of multiple researchers. While obtaining consensus among raters increases the likelihood of yielding credible or “true” findings, we recognize that such interpretations might differ from the authentic interpretations of the respondents we study (Morse, 1997).

Perspectives

Whenever researchers set out to investigate social or educational phenomena, our research designs inevitably contain some amount of error. Error can result in either an overestimation or underestimation of a respondent’s actual value on some metric or construct. There are many sources of error influencing observed values such as the survey instrument itself, characteristics of the respondents, conditions of survey administration, and inter-rater agreement. Qualitative researchers are quite concerned about error associated with rater judgments (Carey, Morgan and Oxtoby, 1996). When coding open-response items, researchers make interpretations or judgments based on substantive criteria outlined in a conceptually organized codebook and score the data accordingly. Examining inter-rater agreement is important because it provides an estimate of the amount of error associated with researchers’ interpretations. Understanding and reporting such error, we believe, will add to the trustworthiness of a study’s findings.

¹ The work reported in this paper was supported by a grant from the RETA program of the National Science Foundation, grant # EHR 0335384. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the Foundation. We would like to thank the additional researchers of the TMI Study for their contributions to the work reported here: Steve Benson, Lynn Goldsmith, Greta Johnson, Barbara Scott Nelson, Kristen Reed, and Amy Shulman Weinberg at the Education Development Center; and Dan Burke (retired) and Apriel Hodari at The CNA Corporation.

Data

The data for this paper was drawn from the *Thinking about Mathematics Instruction study (TMI)*. Sponsored by the National Science Foundation, the TMI study is a five-year, large-scale investigation of principals' "leadership content knowledge" in mathematics (Stein and Nelson, 2003). Borrowing from Shulman's (1987) concept of pedagogical content knowledge in which a teacher's subject matter knowledge is viewed as central to teaching and transformed through her or his pedagogical skills, and into content knowledge useable for teaching, this study suggests that principals' mathematics content knowledge is likewise critical to leadership practices. Mathematics knowledge is therefore transformed through principal's epistemological beliefs about teaching and learning to produce knowledge useful in providing intellectual leadership for instruction throughout the school.

The TMI study employs a mixed-method, quasi-experimental research design. There are two cohorts of principals in the sample. Cohort 1 (n = 96) has a baseline in spring 2004 and Cohort 2 (n = 381) has a baseline in spring 2005. The sub-sample for this analysis consists only of the responses from the pre-survey administered to Cohort 2. After filtering, cleaning the data, and recoding variables, open-responses for a total of 328 principals were examined and reported in this paper. Some cases were excluded from the analysis because of missing data and others because they were used in the calibration sessions.

Methods

Survey Instrument and Raters. An in-depth principal survey was administered to obtain information on beliefs about the teaching and learning of mathematics. Raters scored open-responses based on a scenario of a 4th grade classroom. The scenario, borrowed from Schifter and her colleagues (1998), depicted a conversation between the teacher and her students about what it means to divide a larger number by a smaller one.

Codebook. The *Math-In-Use Scoring Rubric (MUSR)* was designed to assess the extent to which respondents focused on mathematical ideas contained within or extrapolated beyond the scenario using their own math knowledge. The codebook contained three scoring indices: (1) *Math-In-Use*, (2) *Correctness* and (3) *Mathematics Consistency*. *Math-In-Use*, the primary variable, was a 5-point item (A to E), where "no math" was scored on one end (A), and high level and detailed descriptions of math (E) at the other. *Correctness* was a 3-point item measuring whether the math in the written response was "correct, incorrect or unscored." *Mathematics Consistency* was a 3-point item measuring whether respondents discussed more mathematics in the final two open-response segments, in comparison to the prior segments.

Coding. All eight TMI study researchers were trained on the codebook in fall 2005. After training was completed, the researchers were randomly assigned to code data in rotating groups of three. After coding independently, raters met in triads to reach consensus on scores. The sessions involved a three-step process. First, the group facilitator separately recorded the scores each rater assigned independently. Next, the facilitator identified cases where there was complete agreement among the group and cases where unanimity was lacking. If there was three-way agreement, then the codes were entered into the database, and the session was complete. If agreement was not reached independently, raters engaged in debate to reach consensus for each case. Finally, if consensus could not be reached, the team forwarded the data to a group of senior researchers for a final ruling, called arbitration. About 3% of the cases were arbitrated.

Calibration meetings were held early on during the coding period in an effort to improve raters' agreement by increasing homogeneity of interpretations through feedback. As a collective, we met in calibration meetings to discuss ideal types and disagreements. The purpose of such meetings was to build uniformity on rules for coding items (for example, explaining the meaning of a "C" for Math-In-Use).

Measuring Reliability. As mentioned above, three raters were used to code each open-response. It becomes statistically harder to obtain consistently high agreement as the number of raters increases, however, attempting to do so puts the study in a better position to accurately interpret the "true" meaning of the data, the ideas intended by the principals themselves. The rationale is that multiple researchers might attend to different aspects of a response as they apply the rules of codebook and the outcome of their collective interpretations stand a better chance of reflecting the respondent's intended meaning.

There are several methods for estimating inter-rater reliability in a study (Cohen, 1960; Fleiss, 1980; Creswell, 2003), and this paper combines two — Fleiss's kappa statistic (1971) and the alternative chance-correlated coefficient (AC1). These statistics were applied because they allow for the estimation of agreement among three raters who classify subjects into more than two distinctive categories. Both statistics were calculated using a SAS macro called INTER_RATER.MAC. This macro was used because statistical software packages are designed only to calculate agreement between two raters, which is the more common form of inter-rater reliability analysis.

Reliability coefficients cannot be calculated with 100% certainty but researchers can estimate them. Tests of reliability are used to determine whether repeated measures of the same instrument will produce the same results every time. Inter-rater reliability is an estimate of the consistency between two or more raters who score the same data. According to Gwet (2002), the reliability estimate quantifies the distance of scores assigned by a group of raters to the same subjects; the closer the scores assigned by each rater, the higher the reliability. In this analysis, Fleiss' (1981) interpretations of reliability coefficients is used, where $\kappa < 0.40$ suggests "poor agreement," $0.40 \leq \kappa \leq 0.75$ means "good agreement," and $\kappa > 0.75$ is viewed as "excellent agreement."

Results

Kappa Statistic. The results of the estimation of the kappa statistic for the Math-In-Use suggest agreement among raters in interpreting participants' discussion of mathematics in the scenario was good, $\kappa = 0.6745$, $p \leq .001$. The degree to which raters were able to independently classify responses at the lower end of the distribution (Category A, No Math) was excellent, $\kappa = 0.8607$, $p \leq .001$. For the B and C categories, the agreement among raters to independently categorize the responses was good, $\kappa = 0.5921$ and 0.6411 respectively, $p \leq .001$. The findings for "Correctness" revealed that overall there is a high level, "excellent," agreement among raters in identifying the accuracy of mathematics written in the open responses, $\kappa = 0.7270$. Furthermore, the agreement among the raters in detecting correct and unscored responses was high, $\kappa = 0.7610$ and $\kappa = 0.7644$, $p \leq .001$, respectively. The results also indicate that the agreement among the raters in classifying responses as incorrect was good, $\kappa = 0.3325$, $p \leq .01$. The results for the Mathematics Consistency showed that the agreement among raters on this item was also good, $\kappa = 0.5476$, however, the coefficient was not significant.

Alternative Chance-Correlated Coefficient. Though the analysis calculates and reports the results for both the conditional and unconditional standard errors, we focused on the conditional variance because inferences can be made to the actual sample of raters. The estimate of the *ACI* coefficient was typically larger than the *kappa* coefficient because of the sensitivity of *kappa* to the unequal trait prevalence in the population, but the pattern was similar. The results for Math-In-Use, $ACI = 0.7412$, $p \leq .001$, provided evidence that agreement among the raters in identifying participants' discussion of math concepts in the scenarios is strong. In addition, the agreement among raters in independently classifying responses in Category A was excellent, $ACI = 0.8920$, $p \leq .001$. For Categories "B and C", there is fairly high agreement among raters, $ACI = 0.6683$ and $ACI = 0.7344$ respectively, $p \leq .001$. Finally, consistent with the results of the *kappa* analysis, the agreement among raters in Categories D and E was not significant. The results for "Correctness" and "Math Consistency" followed a pattern similar to *kappa*, but the magnitude was greater.

Scientific Importance of the Study

This paper offers homegrown lessons to the educational researchers and social scientists about strategies used in a large-scale study on school administrators to rigorously examine the important issue of agreement among multiple raters of open-response items. The evidence suggests the existence of strong agreement among raters in categorizing written responses. The agreements between coders were greater on scores made at the lower end of the distribution. This has implications for substantive features of the study, specifically how principals use of rudimentary math concepts or avoidance of mathematics altogether may be easier to identify than more sophisticated uses of math. Two analytic approaches to investigating inter-rater reliability were employed in tandem and each indicated a relatively high consistency of agreement among raters.

We conclude that at least two inferences can be made about decisions made by the researchers involved in the study and about use of the codebook. First, based on results of *kappa* and *ACI* analysis, each of the raters appeared to have a relatively clear, common understanding of the codebook. This is reflected in both the estimates of reliability analysis. *MUSR* appears to be a useful codebook for collecting information on beliefs about the teaching and learning of elementary mathematics. Because the agreement among raters was high in most instances, this may suggest that the rubric itself is well defined and the categories are distinguishable. There were, overall, more instances of agreement than not, and inter-rater agreement increased overtime. Perhaps calibration sessions helped to cultivate common interpretations. Second, the research team took a conservative approach to coding by grouping raters in teams of three, rather than pairs. No doubt that the use of pairs would produce higher reliability estimates, in statistical terms, but adding an additional rater may increase the possibility of moving interpretive accuracy closer to authentic accuracy, or finding the true meaning of the data. While it is difficult to obtain consistently high agreement among three independent raters, attempting to do so places researchers in a study of this nature in a better position to accurately interpret the true meaning of the data, the meaning intended by the study participants.

References

- Carey, J. W., Morgan, M., and Oxtoby, M. J. (1996). Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research. *Cultural Anthropology Methods*, 8, 3, 1-5.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Second Edition. Thousand Oaks: Sage Publications
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.
- Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Second Edition. New York: Wiley.
- Gwet, K. (2002). An inquiry into the adequacy of inter-rater reliability assessment methods and the validity of associated standard errors. Retrieved April 11, 2006, from STATAxis Consulting Web site: <http://www.stataxis.com/articles/papers/Inquiry.pdf>
- Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis*. Second Edition. Thousand Oaks, CA; Sage Publications.
- Morse, J. M. (1997). Perfectly healthy, but dead; the myth of inter-rater reliability. *Qualitative Health Research*, 7, 4, 445-447.
- Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage Publications.
- Schifter, D., Bastable, V. and Russell, S. (1998). *Developing Mathematical Ideas: Number and Operations: Making Meaning for Operations*. White Plains NY: Dale Seymour.
- Stein, M. K., and Nelson, B. S. (2003). Leadership content knowledge. *Educational Evaluation and Policy Analysis*. Winter 2003, 25(4). 423-448.
- Shulman, L.S. (1987). Knowledge and teaching. *Harvard Educational Review*. 57, 1, 1-22.