To appear, *Journal of the Learning Sciences*

# Confusing claims for data: A critique of common practices for presenting qualitative research on learning

David Hammer, Tufts University
Leema K Berland, University of Wisconsin

## Abstract

We question widely accepted practices of publishing articles that present quantified analyses of qualitative data. First, articles are often published that provide only very brief excerpts of the qualitative data itself to illustrate the coding scheme, tacitly or explicitly treating the coding results as data. Second, articles are often published that treat inter-rater reliability solely as a matter of justifying the coding scheme, without further attention to the variance it makes evident in the process of coding. We argue that authors should not treat coding results as data but rather as tabulations of claims about data, and that it is important to discuss the rates and substance of disagreements among coders. We propose publication guidelines for authors and reviewers of this form of research.

## Introduction

Fifteen years ago, Michelene Chi (1997) published “Quantifying qualitative analyses of verbal data: A practical guide.” She wrote it as a practitioner, expressing clearly and directly how to gather and select data, devise and use coding schemes, and understand the results. Perhaps most valuable, she provided a straightforward “how the sausage is made” account of the challenges and pitfalls of this work. For many of us, reading and rereading that article is to see our own experiences and difficulties articulated. Most of her account applies to many forms of qualitative data, not just verbal, and the article has had a strong impact on the learning sciences. It is frequently cited in methodology sections, and it is a standard assignment in qualitative methods courses.

Our purpose here, working from Chi’s account, is to propose guidelines for the *presentation* of research that quantifies qualitative data. We write as readers of this research, motivated by the prevalence of two problematic but widely accepted practices—“accepted” in the tangible sense of publication.

First, articles often provide only very brief excerpts of the qualitative data to illustrate the coding scheme. Schoenfeld (1992) challenged this practice in an article proposing “standards for novel methods.” His standards included

- Describe the method in sufficient detail that readers who wish to can apply the method.
- Provide a body of data that is large enough to allow readers to (a) analyze it on their own terms, to see if their sense of what happened in it agrees with the author's, and (b) employ the author's method and see if it produces the author's analyses (Schoenfeld, 1992, p. 181).

Many articles do not meet these standards. Instead, the typical practice is to describe the coding scheme and to illustrate it with examples, then to present the results of researchers’ codings as the data for the article, often in tables that summarize numerical results in the various categories.[1] When articles give only brief illustrations of the coding they are effectively—and often explicitly—treating the results of coding as data itself.

Second, articles generally treat inter-rater reliability (IRR) solely as a matter of justifying the scheme and the coding results, without considering the variance it makes evident. Thus an article might report that IRR was 93%, explain that disagreements were resolved through discussion, and proceed to treat the

[1] We believe this will be familiar enough to readers that we do not need to call out particular examples.

coding results as definitive for further quantitative analysis. This, in effect, asks readers to accept the results of coding not only as data, but as *error-free data*.

We present our arguments regarding these practices in three sections. First, we address the question of what constitutes "data" in this form of research, to argue it is the set of qualitative records researchers examine in their analyses. We then review what takes place in coding, following Chi (1997), to argue that the synthesis of a coding scheme is the development of a "novel method" (Schoenfeld, 1992), and its application a process of making claims about the qualitative data. We then offer our proposals for guidelines in the presentation of research that quantifies qualitative data analysis.

## What is data?

Schoenfeld (1992) and Chi (1997) spoke of data as the qualitative records. Chi focused on verbal data, typically transcripts from videotapes, but as she noted the approach she was describing could apply to other forms of "messy data," meaning "such things as verbal explanations, observations, and videotapings, as well as gestures" (p. 271). Schoenfeld specifically wrote about "data in the form of videotape," but much of his discussion, and the standards he proposed, would apply to other forms of qualitative data as well. Regardless, in both of these accounts, "data" means the qualitative records, not the results of coding. Below we review the process of coding, following Chi's (1997) account, but before we do that it is important to address the question of what privileges the qualitative records to be designated as data.

Several years after Chi's piece, Hall (2000) published the provocative essay "Videorecording as Theory," in which he challenged the compelling illusion that videorecordings are "objective or theory neutral data." He described the myriad of choices, sometimes invisible, involved in generating video: What direction to point the camera; how tight the zoom; where to place the microphones; what portions of video to analyze, and what portions to present? All of these choices reflect the researchers' expectations and interests, and in this sense, Hall argued, video recordings are "both technology and theory laden." It is misleading to think they are simply information collected from the world. Rather, he argued, video records are *constructed* by researchers.

Certainly Hall's arguments apply to other sorts of qualitative records as well, and they raise serious difficulties for Schoenfeld's standards. They imply, for one, that readers who do not share the theoretical presumptions that went into the researchers' construction of their video data would need *other* data, in order to "analyze it on their own terms." More central to our purposes here, they show how blurry the line is between data and claims: What in principle distinguishes the records of the phenomena, which Hall shows are constructions by researchers, from the next level of construction in the coding of those records?

We agree with Hall in all respects. At the same time, it is important to recognize that there is no such thing as "theory neutral data" in any field. As such, the burden on research in the learning sciences is similar to the burden in all sciences: Authors must explain and defend the construction of data to make the case that the records are adequate and appropriate to inform the research questions.

For illustration, consider an experiment to determine whether air has mass. Scientists place a balloon on a scale to measure its mass, and record the number. Then they inflate the balloon, place it on the scale again, and record the number. They repeat this for multiple balloons. If the masses they recorded for the inflated balloons are always larger than for the deflated balloons, they conclude air has mass. Clearly, the set of measurements is the scientists' data, but it is "technology and theory laden." What is the sensitivity of the scale? How were the balloons inflated? (If by mouth, perhaps the added mass was from moisture.) Were there air currents in the room? How did the researchers account for buoyancy? How did they account for the pressure in the balloons? How did they account for the effects of static electricity?

There are certainly important differences between the "natural sciences" and "social sciences." Cicourel (1964), for example, argued that research in sociology is inevitably reflexive; researchers are people studying people. In this respect, however, there is similarity: All data is the result of a myriad of choices, often invisible, about how to construct records of phenomena. Researchers using records claim, tacitly or explicitly, that the records are worthy of study, but they can only support that claim with arguments of plausibility or principle, based on their theoretical framework, findings from other studies, or common experience. The phenomena themselves are not available.[2] This, we contend, is a clear difference between records as claims and codes as claims: Codes as claims have a persistent material basis.

That is to say, the qualitative records are the focus of extended attention and effort by the researchers, as Chi (1997) described. In this way, they serve as the "given material" for analysis, and in that literal sense they are the data: The word "datum" is from the Latin for "something given." They are also claims, in Hall's sense, about phenomena that are no longer accessible, and, as such, once those claims are made, the corpus is determined, and the analysis proceeds from there. The construction of records and the coding of those records both involve researcher decisions that become claims in publications. The latter can—and therefore should—be supported by material evidence collected within the study.

In sum, we argue, it is appropriate to consider the records of phenomena that researchers examine as the data for the study. In most cases, however, it is not appropriate to consider researchers' codings of those records as data, and we turn to argue that point now.

## What happens in coding?

The practice we are challenging, of providing numerical summaries of coding results but only brief examples of qualitative data, treats coding as analogous to the *reduction* of quantitative data. For example, scientists studying climate change might have the raw data of temperature readings from thousands of weather stations. They might then choose to average the readings for weather stations that are near each other, which would result in a smaller set of numbers they would use as the data—the given material—they analyze and report in manuscripts. The process can be mathematically systematic, such that other scientists given the same raw data could replicate it precisely, and this would help readers accept the reduced set as the data for the study.

There is great practical motivation for treating coding as data reduction. The problem is that the processes of devising and applying a coding scheme are not mathematically systematic. They involve researcher choices and judgment, in developing the scheme, applying it to data, as well as in deciding whether and when to iterate the process. We discuss the development and application of the coding scheme and the implications for publication in this section.

### Developing a coding scheme

Chi (1997) described the process of developing a coding scheme as an "interactive top-down and bottom-up process. It is interactive in the sense that one should be open to additional modifications as one becomes familiar with the verbal data" (p. 291). Other accounts of qualitative analyses (Charmaz, 2005; Glaser and Strauss, 1967; Marton, 1988) similarly describe iterative processes for developing coding schemes that represent the data. Loosely speaking, these processes include positing or inferring some initial analytic scheme, trying to apply it to data, discovering inadequacies and returning to expand or refine the categories of their scheme.[3]

[2] Often in the natural sciences it is possible and important to reproduce phenomena, but for any particular experiment all that remains are the records.

[3] Our emphasis here is on the publication of this form of research; we recommend readers interested in methodology consult Chi's article and these other accounts.

Inadequacies in coding schemes often become evident through the IRR processDeveloping a coding scheme requires researchers to articulate definitions of categories well enough that others could interpret them and recognize them in the data; mismatches among coders drives further inspection of the data, discussion about the categories, and revisions to the scheme. Along the way, there are new insights into the data, consideration of many informal mini-hypotheses, and productive debates, and one reason to provide more qualitative data is to help readers benefit from those insights.

For example, studying students' written explanations and argumentation, Berland and Reiser (2009) began with a coding scheme based on existing literature, with epistemic categories of *claims*, *evidence* and *logical inferences*. However, they found that in many instances, they could not reliably distinguish *evidence* and *logical inferences* in the students' writing. The researchers shifted their attention to developing a new scheme for coding written products, with two categories: (1) evidence and inference were distinct and (2) evidence and inference were not distinct. Separately they had coded for persuasive statements, something they found they could do reliably. The new coding scheme showed a correlation: Written explanations in the first category (i.e., those that clearly differentiated evidence and inference) were more likely to contain persuasive statements than those that were in the second category. This led the authors to conjecture that students' efforts to be persuasive supported them in being more explicit when supporting their claims, including with respect to differentiating between their evidence and logical inferences. In this way, as they discuss, Berland's and Reiser's research focus emerged out of challenges with developing a coding scheme, and the scheme itself was an important aspect of their findings.[4]

In this way coding schemes are epistemic forms (Collins & Fergusson, 1993) in the learning sciences—they are "structures that guide inquiry" (p. 25). They represent methods for analyzing the data, and as such they mediate researchers' interactions with the data and with each other. In the end, a coding scheme itself represents a kind of finding, a claim that the data can be interpreted in this particular way (Marton, 1988), and that this sorting is meaningful and productive. For this reason, details and difficulties associated with its development are of interest to readers.

## Applying a coding scheme

Given a coding scheme, applying it to data involves more choices and judgment. As Erduran, Simon and Osborne (2004) put it, describing the challenges of applying their coding scheme to student discourse, "there is inevitably a process of interpretation to be made and that some of that process is reliant on listening to the tape and hearing the force of the various statements..." (p. 922). There is a great deal of intellectual action, for the individual coders trying to decide borderline cases, among coders comparing and trying to reconcile disagreements, and for the group as a whole in deciding whether the persisting challenges in applying the scheme warrant iteration back to its revision. Often this last decision is made with respect to a threshold of inter-rater reliability the researchers consider acceptable.

On large datasets, the independent coders will code a shared sub-set of the data, determine a sufficient level of IRR—possibly using statistical measures—and then divide-and-conquer the remaining data. On smaller datasets, multiple coders might code the entire dataset, report their percent-agreement, and resolve disagreements through discussion. Either way, researchers typcially treat an above-threshold IRR as a warrant for coding, and they proceed to present their findings based on the numerical results without further consideration validity or the implications of disagreements. This, we argue, is problematic, in two respects.

First, there is the familiar point that high IRR does not imply validity. Asked to code an oval as a circle or a square, independent coders would reliably choose circle, but nobody thinks it is a circle. Validity in qualitative research requires alignment among observations (or data), measurement instruments (coding schemes, in this case), and theoretical paradigms (Kirk & Miller, 1986, p. 21-23). At the time of writing this essay, the second author is experiencing a situation in which her coding results do not align with the observations thereby causing her to question the validity of the coding scheme: She has attained a

[4] In other respects, the article followed the same widely-accepted practices we are calling into question.

relatively high IRR (ranging from 75-85%, depending on the particular code), analyzing student discourse, and patterns in that coding suggest an interesting finding. She does not, however, see evidence to support that finding in the data—the videos and transcripts. While the coding scheme seems promising simply on the basis of reliability, she does not believe it represents the data.

Second, the IRR process itself provides evidence to challenge working from the numerical results without qualification. If IRR is 85%, then the researchers disagreed on 15% of the codes. It is important to recognize that this implies some of the 85% agreement occurred by chance: There were close calls the coders happened to make in the same direction. As such, the IRR results should continue to inform the research, including with respect to subsequent interpretations of generalizability. Few articles spend time discussing the substance of borderline cases, and we are not aware of any that fold IRR into calculating statistical significance of coding differences. By limiting the role of IRR to warranting the coding scheme, researchers treat the numerical counts as error-free, even when the IRR results themselves give evidence regarding variation.

The numerical results of quantified analyses of qualitative data are, and should be treated as, tallies of claims about that data. Of course it is not practical for authors to support each claim, for each coding, with evidence in the data within an article. The burden on authors instead is to convince readers to accept their process of interpreting the data, in the aggregate, and that should include acknowledgement and discussion of uncertainties in the numbers.

## Guidelines for presentation

We have identified three levels of claims, in research that quantifies analyses of qualitative data (Chi, 1997), and each level needs to be addressed in publications.

First, as Hall (2000) argued, there is the claim that the data— the qualitative records—are worth studying for insight into the phenomena of interest. There is no "objective or theory neutral" data (Hall, 2000), and we agree with him that it is essential for researchers to describe their construction of data as part of supporting the claim that it can provide insight into phenomena of interest.

Second, there is the claim that the coding scheme is a meaningful, productive structure for analyzing that data. There must be empirical support for that claim in the qualitative data. Precisely as Schoenfeld (1992) argued, readers need to see a "body of data that is large enough" for them "to see if their sense of what happened in it agrees with the author's."

Moreover, we suggest that the presentation of the data must offer discussion of borderline cases, to give readers a sense of limitations in the method, where the categories may be difficult to discern. To that end, the data should go well beyond the best cases researchers generally use to illustrate categories; the examples should show the complexity of applying the codes, including instances in which the independent coders disagreed and explaining how agreement was reached.

Third, there is the series of claims about elements of the data, the coded records, and the support for each of these claims would again be in the data. It is not generally possible or appropriate for researchers to provide specific evidence within a manuscript for each application of a code, so what needs support in an article is how much to trust the aggregate quantities. This, certainly, overlaps substantially with supporting the scheme itself, and much of the support for the one is support for the other: If there is enough data, then readers can "employ the author's method and see if" their codings agree (Schoenfeld, 1992).

The final matter we suggest needs to be addressed in presentation is a sense of how much the numbers of codes in the different categories could vary, from one analysis of the data to another. Readers can get a qualitative sense of that by trying to code data themselves, and articles should make use of the information from tests of inter-rater reliability not only as part of the methodology sections but within the analyses and conclusions as well.

These, then, are all considerations we propose for authors and reviewers, in the preparation and assessment of manuscripts, such that readers can benefit from the rich intellectual work that took place for the group, in developing and applying the coding scheme. Beyond manuscripts, it is now possible for more extensive dissemination of data in online supplements to articles. While it is not possible or desirable for authors to present evidence to support each claim, code-by-code, within a research article, it might be possible to make that data available for interested readers. This would have the added advantage for the community of secondary analyses that could support or challenge the original claims, or, perhaps, open entirely new lines of inquiry (Baecker, Fono & Wolf, 2009; MacWhinney, 2009).

## Closing note

The press for quantification in the learning sciences is largely motivated by the productivity of quantification in the ”natural” sciences, as captured in the famous quotation of Lord Kelvin:

> In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. (Kelvin, 1891)

Since that time, quantification has proven valuable in other disciplines than physical science; certainly there is reason to expect it to serve the learning sciences.

Kelvin’s dictum underplays, however, the challenge of that “first essential step," of establishing those “principles of numerical reckoning and practicable methods for measuring some quality.” In the physical sciences as in the learning sciences, one cannot simply elect to quantify. Arriving at a useful quantification is a creative process, heavily theory and technology laden. But, much as video recording can create the illusion of objectivity (Hall, 2000), it is easy to lose track of the human judgment that went into their construction when one has numbers in hand (Porter, 1995). And, in any field, to settle on a quantification inappropriate for a given research question can be detrimental to a program of research, systematically misdirecting researchers‘ attention.

Thus, while we see quantification as a powerful tool for making sense of the whether, when, and how students learn, we worry that the field’s almost casual publication of coding results obscures the complexities of the work. We are not remotely arguing against research by these methods; we are arguing for ways to make its dissemination more substantive and productive.[5] The practices we have challenged make it unlikely that readers can understand and use—or learn from—the authors’ methodological achievements. As such, we are arguing for new publication standards. The community should come to expect authors to depict the complexity of their coding process and address the potential errors and insights therein.

## Acknowledgements

This essay draws significantly on Hammer & Louca (2008). In an earlier version it benefitted from feedback by Mike Stieff and Bob Mislevy, and in a more recent version by Michelle Wilkerson-Jerde.

[5] Nor, it may be important to note, are we arguing that qualitative research must eventually lead to quantification. To be sure, our own work (e.g. Berland & Hammer, 2012) is often “purely” qualitative.

## References

Baecker, R. M., Fono, D., & Wolf, P. (2009). Toward a Video Collaboratory. In R. Goldman, R. Pea, B. Barron, & S. Derry (Eds.), *Video Research in the Learning Sciences* (p. 461-478). New York: Routledge Taylor & Francis Group.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Charmaz, K. (2006). *Constructing grounded theory*. London ; Thousand Oaks, Calif.: Sage Publications.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the learning sciences6*, *6*(3), 271–315.

Cicourel, A. V. (1964). Method and Measurement in Sociology. New York: The Free Press.

Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, *28*(1), 25–42.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*, 915–933.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory; strategies for qualitative research. Chicago,: Aldine Pub. Co.

Hall, R. (2000). Videorecording as Theory. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 647–664). Mahway, NJ: Lawrence Erlbaum Associates.

Hammer, D. and Louca, L. (2008). Challenging accepted practices of coding. In V. Jonker & A. Lazonder (Eds.) *Cre8ting a Learning World: Proceedings of the 2008 International Conference of the Learning Sciences (*Vol. 3, pp. 307-8). International Society of the Learning Sciences.

Kelvin, W. T. (1891). *Popular lectures and addresses*. London, New York,: Macmillan and Co.

Kirk, D. J., & Miller, M. L. (1986). *Reliability and Validity in Qualitative Research*. Newbury Park, CA: Sage Publications, Inc.

Marton, F. (1988). Phenomenography – A research approach to investigating different understandings of reality. In R. R. Sherman & R. B. Webb (Eds.), *Qualitative Research in Education: Focus and Methods* (Vol. 21, pp. 143-161). London: Falmer Press.

MacWhinney, B. (2009). A transcript-video database for collaborative commentary. In R. Goldman, R. Pea, B. Barron, & S. Derry (Eds.), *Video Research in the Learning Sciences* (p. 537-546). New York: Routledge Taylor & Francis Group.

Porter, T. M. (1995). *Trust in Numbers : The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J.: Princeton University Press.

Schoenfeld, A. H. (1992). On Paradigms and Methods: What Do You Do When the Ones You Know Don't Do What You Want Them to? Issues in the Analysis of Data in the Form of Videotapes. *The Journal of the Learning Sciences*, *2*(2), 179–214.