International Society of
the Learning Sciences

Quantified Qualitative Analysis: Rubric Development and Inter-rater Reliability as Iterative Design

Kathryn S. McCarthy, Joseph P. Magliano, Jacob O. Snyder, Elizabeth A. Kenney
kmccarthy12@gsu.edu, jmagliano@gsu.edu, jsnyder14@student.gsu.edu, ekenney3@student.gsu.edu
Georgia State University, Department of Learning Sciences

Natalie N. Newton, Cecile A. Perret,
nnnewton@asu.edu, cperret@asu.edu
Arizona State University, Department of Psychology

Melanie Knezevic
mknezevic2@student.gsu.edu
Georgia State University, Department of Learning Sciences

Laura K. Allen
laura.allen@unh.edu
University of New Hampshire, Department of Psychology

Danielle S. McNamara
danielle.mcnamara@asu.edu
Arizona State University, Department of Psychology

Abstract: The objective in the current paper is to examine the processes of how our research team negotiated meaning using an iterative design approach as we established, developed, and refined a rubric to capture comprehension processes and strategies evident in students' verbal protocols. The overarching project comprises multiple data sets, multiple scientists across (distant) institutions, and multiple teams of discourse analysts who are tasked with scoring over 20,000 verbal protocols (i.e., think aloud, self-explanation) collected in studies conducted in the last decade. Here, we describe the iterative modifications, negotiations, and realizations while coding our first subset comprising 7,559 individual verbal protocols. Drawing upon work in design research, we describe a process through which the research team has negotiated meaning around theory-driven codes and how this work has influenced our own ways of conceptualizing comprehension research, theory, and practice.

Introduction

Learning from text involves a negotiation of meaning between multiple entities, including authors, instructors, and students. Students are often faced with the challenges of understanding and learning from text that exceed their prior domain knowledge and reading skills (e.g., Goldman & Bisanz, 2002; Moje & Speyer, 2014). Through a more robust understanding of these learning processes, researchers can develop activities and interventions that can help students to better learn from the challenging texts that they encounter. Thus, we aim to understand the various constructive processes that learners use when building and negotiating the meaning of texts. In our work, our overarching goals are to understand these processes and to develop adaptive technologies that provide individualized support for students as they learn from text.

One approach to understanding readers' comprehension processes is to have them produce verbal protocols as they read. Students may be asked to "think aloud" by reporting whatever thoughts come to mind. They may also be asked to "self-explain" in which they explain the text to themselves as they read. Prompting students to generate these verbal protocols as they read can be a means of enhancing learning, but such prompts also serve as a window through which researchers can better understand the processes that students engage in when they read (e.g., Chi, 1997; McNamara, 2004; Pressley & Afflerbach, 1995). Coding and quantifying these responses allows researchers to examine how different processes and strategies relate to students' individual differences (e.g., interest, prior knowledge, reading skill; e.g., Coté et al., 1998) and to learning outcomes (e.g., Magliano et al., 2011) as well as response to interventions (McNamara, 2004).

For the past few decades, members of our team have gathered approximately 20,000 verbal protocols in our various studies of learning from and with text. Our current objective is to code (or recode) these reader responses using a common rubric agreed upon by an interdisciplinary group of researchers and discourse analysts.

Qualitative approaches lend a rich understanding of discourse (e.g., Gee, 2014; Pressley & Afflerbach, 1995). To attempt to scale these approaches, many researchers quantify qualitative aspects of discourse to statistically relate these features to other factors such as individual differences or experimental variables (e.g., Chi, 1997; Coté et al., 1998). This coding process relies on rubrics designed to capture a theoretical construct of interest and reaching reliability between raters (Chi, 1997). Typically, a rubric is designed by scientists or experts, and raters adapt their codes to increase reliability, iteratively matching to the conceptual rules set by the expert and to each other. The rubric is more or less the stable entity, wherein examples are added and some wordings may be modified, but few substantive changes are made. In some cases, discussions can lead to non-trivial refinements in how the constructs are conceptually and operationally defined, but this seems rare, or at least opaque. The debates, discussions, and decisions surrounding the operationalization of constructs (if they occur) are generally left unreported when the main objective is to reach quantifiable reliability between raters. In many papers, this complex process is often distilled down to a single sentence (e.g., "*raters were trained to criterion and then coded the remainder of the data*").

Our objective in the current paper is to examine the *process* of how our research team negotiated meaning as we established, developed, and iteratively refined a rubric to capture these important comprehension processes. We describe a negotiation of meaning, but one different from that faced by the readers - one underlying the process of coding these protocols: a negotiation between the scientists (who developed the rubric) and the discourse analysts tasked with reaching agreement in their assessments of the nature of the processes revealed within the protocols. We examine how these interactions resulted in a deeper understanding of the student responses and the larger research goals.

The current project

There are two unique aspects of the current project. The first is its scope and scale. In most situations, a few hundred verbal protocols are scored by a pair or a small group of raters who are co-located and the data are coded within a small window of time. Our project, by contrast, included multiple data sets, multiple scientists across three (distant) institutions, and multiple teams of discourse analysts. Moreover, due to COVID-19, these negotiations have occurred predominantly in virtual space. A second unique aspect of this project was exploring the development and refinement of the rubric as both an iterative design task to create a functional educational tool and a learning opportunity in and of itself.

While the larger research project draws from cognitive theories of text and discourse comprehension (see McNamara & Magliano, 2009, for a review), the current work around the development and refinement of the rubric takes inspiration from both instructional design and design-based research. Although agreed-upon definitions of design-based research remain elusive, central considerations of DBR is that research is iterative, collaborative, and highly sensitive to its context (e.g., Barab, 2014). Of particular interest in the current work is the way in which expert researcher/scientists and novice rater/discourse analysts co-constructed meaning within small rating teams and across the larger project as they worked through the coding process. Establishing reliability often requires in-depth discussion about which codes are relevant to the research question and how those codes are operationalized. Modifying the rubric requires mutual understandings between the scientists and the discourse analysts. We iteratively adapted the rubric and our mutual understandings to increase not only reliability between the raters, but also to capture the unique perspectives of the raters. This "push and pull" between reliability and validity, wherein one objective is to maximize reliability between raters, and the other is to maintain some level of validity intrinsic to the fundamentally qualitative nature of the analysis, was a driving force in this process.

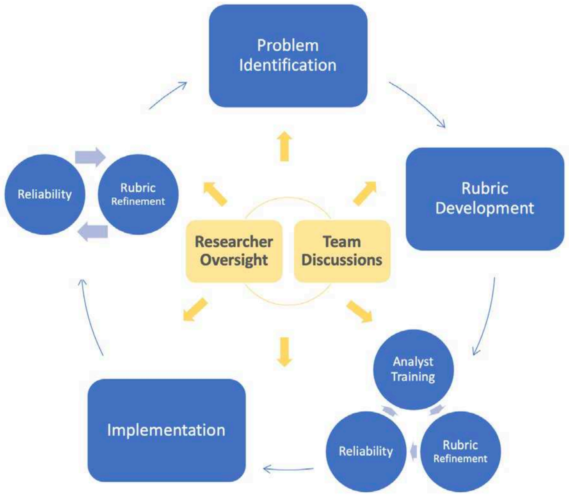
Context and data sources

The larger project relies on verbal protocol data from a series of large-scale studies. These data were collected from different regions of the US, included students reading at, above, and below grade levels, native and non-native English speakers, and students enrolled in both traditional college courses and those assigned to developmental education. Although the texts and tasks varied from study to study, the general procedure for collecting verbal protocols was the same. Students read a text and were prompted to type a verbal protocol at various target sentences (predetermined locations in which connections and explanations were likely to be fruitful for comprehension). Such typed protocols are easier to collect than spoken protocols and yield similar properties in terms of the strategies that readers demonstrate during reading (Muñoz et al., 2006). Some protocols were produced under the instruction to think aloud while others were instructed to self-explain. This varied both within and across studies. After reading and generating verbal protocols, students answered either multiple-choice or open-ended comprehension questions. In addition to the comprehension task, students also completed a variety of individual difference tasks including standardized comprehension tests, foundational reading skills tests, and various measures of interest such as working memory. In some studies, verbal protocols were collected in a single session, while in other studies, the verbal protocols were collected before and after reading comprehension interventions. The research team included four experienced researchers (the PI and Co-PIs of the project) and two teams of graduate students and research assistants spread across multiple research sites. A third team was brought into discussions, but due to practical limitations, especially in light of COVID-19, this team left the project. However, we wish to acknowledge their contributions to our thinking and the various revisions to the rubric.

In the current work related to the design and iterative refinement of the coding rubric, we relied on multiple data sources. We collected copies of each iteration of the rubric (including tracked changes and comments from version-to-version). The raters coded the verbal protocols in excel spreadsheets. Raters were asked to provide their codes as well as any notes or questions they had regarding specific responses. Thus, we had these excel files for each rater's initial independent pass and a second pass "corrected" after discussion. We also had combined excel files in which we compared rater scores and logged discussion notes. Additionally, we examined rater and researcher notes from the weekly meetings that occurred during initial training as well as email threads that include discussions around specific examples.

Design cycle

This work was inspired by instructional design (e.g., Silber, 2007) and design-based research (Design Based Research Collective, 2003; Easterday et al., 2014). This paper reflects our initial design cycle from preliminary problem conception through a first implementation, evaluation, and refinement of the coding rubric. Critically, in addition to the larger "design loop", we also focus on important rapid, iterative cycles that occurred throughout the process. These cycles involved both quantitative and qualitative analysis of the rubric and inter-rater agreements. Our design cycle appears in Figure 1. In the following sections, we outline each phase of the research in more detail to describe the process and how the discussions within rater groups and across the larger team led to refinements in the rubric as well as the team's understanding of the theoretical underpinnings and implications of the codes.



The Design Framework reflects how the research team engaged in multiple rounds of coding, development, discussion, and refinement.

Figure 1. Design Framework

Phases 1 and 2: Problem identification and rubric development

Although these data sets have been scored by human raters in previous studies, the protocols have been scored by a number of different raters using similar, but not identical rubrics. Thus, in order to be able to draw meaningful conclusions and to make comparisons between datasets, the research team agreed that we needed to develop a common rubric. While our team has expertise in quantified qualitative protocol analysis, the scope of this project brought to light interesting theoretical and practical issues. Our data sets include different age groups and skill levels, different types of texts (history, science), reader tasks (e.g., self-explain, think-aloud), and different methodological aims across studies. Thus, we were tasked with developing a rubric that was (1) sensitive to different texts and tasks, but general and *flexible* enough to be applied across a number of contexts and (2) reflected strong enough criteria (e.g., rules) for consistent operationalization, but not so rigid that the codes no longer reflected authentic human judgments.

Our ideation process revolved largely around developing an initial rubric. Our initial rubric was adapted from a number of extant rubrics that have been designed based on theories of discourse comprehension (see McNamara & Magliano, 2009). These theories assume that readers engage in a number of comprehension processes and strategies that support the construction of a coherent and elaborated mental model. The rubrics used in this area of research tend to include three broad categories of strategies: paraphrasing, bridging, and elaboration. *Paraphrasing* is a process in which students reproduce content from the texts. *Bridging* is a process of establishing how the current sentence is related to prior discourse context and provides the primary basis for achieving coherence in a mental model. Finally, *elaboration* is a process that involves integrating information from prior knowledge with information provided by the text. Bridging and elaboration are *inferential processes* that are critical for mental model construction. It was also important to us, theoretically, to not categorize each protocol as *either* paraphrase, bridging, or elaboration, but rather the extent to which these processes were apparent in students' protocols. Thus, our rubric included three broad dimensions: presence of the strategy, nature of the strategy, and an overall quality score. Table 1 shows the questions that each of the codes were designed to address. Additional codes were included to capture other common behaviors (e.g., inaccurate statements, life events, direct copy and pasting).

Table 1: Questions driving the three dimensions of interest for paraphrasing, bridging, and elaboration

Dimension	Paraphrase	Bridging	Elaboration
Presence of Strategy	How much of the target sentence is captured? *(none; some; most)*	To what extent is information from previous text present? *(none, little, some, much)*	To what extent does the reader bring in outside information to make sense of the text? *(none, little, some, much)*
Nature of Strategy Use	-	Are connections made to ideas in previous sentence (*local bridge*) or other previous sentences (*distal bridge*)?	Is the outside information relevant to understanding the text?
Overall Quality	What is the overall *quality* of the response? *(Poor, Fair, Good, Great)*		

Phase 3: Training and rapid iterative refinement

The initial data set (Magliano et al., 2020) included verbal protocols from 597 college students. Students wrote verbal protocols for two texts. The history text (Louis XVI) included six think aloud locations and the science text (Erosion) included seven think aloud locations. In total, the data set comprised 7,559 individual verbal protocols. The verbal protocols were left segmented "as is" in the sense that the totality of what a student wrote in response to a given target sentence was coded as a single response.

We randomly selected a subset of 20% of the verbal protocols for training and refining the rubric. This subset of protocols was randomly selected at the participant by text level and then divided into three training phases. The research sites engaged in both intra-team and inter-team discussions. Given time constraints, Team 1 was working full time, while Team 2 was working part time. Raters were encouraged to make notes in the coding files and to bring issues to group discussions. Evaluating and discussing common disagreements led to additions and refinements to the rubric. Major modifications included: a) the addition of a metacognitive *monitoring* code reflecting occasions when students expressed clarity or confusion, b) addition of codes that reflected the nature of a students' paraphrases (*lexical change*, *syntactic change*), c) the simplification of an *accuracy* code from four levels to a dichotomous *misconception* code, and d) modifications of language related to presence and quality of elaborations.

After the third round of coding to obtain reliability, Team 1 demonstrated good reliability on overall (0-3) score (k = .71) and the presence scores (paraphrase presence = .74, bridge presence = .73, elaboration presence = .77). Reliability estimates on the nature of the verbal protocols scores were often lower, but acceptable (syntactic change = .64; lexical change = .81; bridge contribution = .73, elaboration relation = .64). Additional codes (e.g., too short, copy/paste) ranged from .73 to full agreement.

Phase 4: Implementation

Once the first team of raters had established sufficient reliability, the remaining responses were divided amongst the raters, with a random set of 10% of their total assigned responses for post hoc reliability calculations. The raters sent their scores at the end of each week. Coding this data set took approximately 4 months to complete (with some delays due to COVID-19). At the same time, the second team was continuing to establish reliability and to confer with the first team about ongoing disagreements. There were no major changes made to the rubric during this time.

Phase 5: Evaluation and refinement

In most projects, coding the remainder would reflect the "end" of this process. However, given our overarching objective to understand the nature of the protocols, we continued into a second phase of evaluating reliability. Once the first team of raters completed scoring the data set, we examined reliability across the randomized overlap to explore rater drift. The reliability calculations illustrated that most codes were markedly lower than the initial benchmarks, with overall score k = .46 and other scores ranging from k = .48-.73. In order to identify systematic inconsistencies between the raters, we used confusion matrices to identify common mismatches across the raters. Table 2 shows a confusion matrix for the presence of paraphrasing for one of the training phases. The diagonal (green) shows where raters agreed, and the other cells indicate discrepancies. The confusion matrices allowed the research team to identify systematic differences across raters (indicated here in red). These instances became the basis for discussion that led to insight into the construction of meaning by students and is reflected in the rubric.

Table 2. Example confusion matrix for paraphrasing.

		Coder 2		
		0	**1**	**2**
Coder 1	**0**	**116**	20	0
	1	5	**50**	7
	2	1	10	**98**

Examining confusion matrices also revealed that elaboration had become overgeneralized to include *any* information that could not be found in the text and that this had resulted in inflated overall scores for statements that were unrelated to the semantic content of the text. Through discussion across the teams of raters, it was agreed that *elaboration apparent relation* code would be removed and to instead add a more general *nonsense* code to capture statements that were *irrelevant* or off-task. There was also an addition of an *evaluative statement* that seemed to emerge frequently in the history text. These statements tended to follow an "x should have done y" format. While such responses are on task, they did not align with notions of paraphrase, bridging, or elaboration in ways exhibited by other types of statements. As a result of these discussions and changes to the rubrics, raters' reliability increased on *elaboration presence* from k = .48 to .80 and the new codes showed good reliability (evaluative = .74; irrelevant = .83; monitoring = .78).

Lessons learned

Examination of our rubrics and meeting notes (as well as discussions related to the development of this paper) revealed a number of important lessons that we carry forward into our next cycle of refining and implementing this rubric on the next set of data. Our lessons-learned span issues pertaining to the construction of meaning, sensitivity to contexts, the iterative process of refinement, and how consideration of this task as design work led to important insights and valuable changes to our research more broadly.

Negotiating and co-constructing meaning

There were inherent challenges in developing operational definitions that allowed coders to draw upon a shared understanding of the processes of paraphrasing, bridging, and elaboration. These challenges included the negotiation of the underlying meaning of various constructs and the operationalization of these constructs.

Negotiating constructs

The co-construction of knowledge occurred in the context of differences in disciplinary knowledge, epistemological perspectives, and experience coding protocols for both the research scientists and coders. This occurred even in the context of the senior researchers on this project who have an extensive history of collaboration, share a common theoretical perspective, and have implemented similar coding protocols over the past 20 years (e.g., Best et al., 2004; Kopatich et al., 2019; Magliano et al., 2011; McNamara, 2004). While there was a shared understanding of the constructs delineated in Table 1, there were non-trivial differences in how they have been operationalized across studies. For example, both Magliano et al. (2011) and McNamara (2004) used rubrics that were intended to identify the presence of elaboration. However, Magliano et al. (2011) did not evaluate the extent to which elaborative processes reflected the process of knowledge building, whereas McNamara (2004) advocated sensitivity to identifying the role of elaboration in support of knowledge building. These differences were revealed in the early stages of refinement and training of coders, but required several rapid iterations to uncover. The lack of common ground on this issue was only discovered with intense qualitative discussions on coding disagreements between separate teams. These discussions laid bare not only gaps in our rubric, but also helped the researchers to co-construct the meaning of these definitions and how they related to comprehension processes.

Negotiating operationalization

There are tensions inherent to transforming qualitative natural language into quantitative scores. One objective is to reach reliability between raters. This process requires constructing a well-defined rubric, with examples and rules. There are two particular tendencies that we have observed. These two stances toward the coding task became apparent to us because one team of analysts approached this task from one direction, while the other team approached it from the other. On one side, we observed a tendency for raters to seek absolute rules and guidelines, (e.g., exact number of words necessary to qualify as a paraphrase, number of words beyond the target sentence qualify as an elaboration). We have found it challenging to convince coders that humans will have disagreements, but if one relies solely on 'counting' words, the element of human judgment is lost (i.e., we might as well have a computer count words). On the other side of the continuum, coders sometimes "read into" protocols rather than basing judgments on the explicit content of the protocols. Such a tendency is natural because we cannot stop the activation of our own prior knowledge. Thus, coders make inferences using their own knowledge while coding. However, the coders sometimes expressed notions that the student producing the protocol should be "given credit" for an attempt. This was particularly true for protocols that were relatively short (e.g., *Erosion are bad.*), had mechanical and grammatical errors that made them difficult to understand, or were ultimately idiosyncratic (e.g., *Too much water isn t [sic] also good for crops people and mostly, Houston. it floats every time it rains so the people are mostly homeless.*). As a consequence, there emerged a propensity to identify the presence of more sophisticated learning strategies than those reflected in the explicit content of the protocols. Establishing agreed upon ways of scoring protocols required our team(s) to reevaluate and refine their understanding of the task writ large and its role in addressing the larger research question.

At the same time, emergent and persistent disagreements led us to reconsider our constructs. For example, our original coding scheme included four levels (0-3) of elaboration presence. Raters indicated that they were able to clearly understand the difference between a 1 (one or two *words* from outside the text) as compared to 2s and 3s (*ideas* from outside the text). However, they struggled to discern a 2 (an idea from outside the text that is vaguely conveyed) from a 3 (an idea from outside the text that is complete and clearly conveyed). As a result, we collapsed 2 and 3 into a single code that reflected any idea from outside the text. This inability to achieve reliability brings into question the extent to which this distinction is construct-relevant or theoretically important.

Multiple aspects of context

Those who have written about quantified qualitative analysis (e.g., Chi, 1997; Vogel & Weinberger, 2016) note that context is an important consideration for evaluating discourse. In our work, we further identified the need to consider multiple contexts simultaneously.

Tensions between sentence context and participant context

The purpose or quality of a given protocol is context dependent and each protocol exists within multiple contexts. For example, the protocol *Erosion are bad* is, in comparison to other reader responses, relatively poor in quality. However, this protocol nested within a struggling reader's other protocols, may reflect a breakthrough or at least preliminary evidence that the reader is making some sense of the text. Our different coding teams adopted different preferences for sorting and coding the protocols. Some of our raters found it easier to code protocols linearly by student. That is, the rater would code Student A's protocols in response to target sentences 1-9 before coding Student B's responses. These raters indicated that the context in which the student made the utterance helped to make sense of the students' intended meaning in ambiguous responses. By contrast, other raters found it easier to code by target sentence, such that they would code target sentence 1 for Students A-*n* and then code target sentence 2. These raters found that the comparison across multiple students talking about the same sentence made the distinctions between codes more apparent. Both approaches have obvious advantages. However, the end result was that the different approaches resulted in different ratings. Our solution was to adopt aspects of both. Our current approach is to first code by participant to leverage the knowledge that the moment-to-moment context provides. Once the rater has scored a number of participants (e.g., at the end of the week), the rater resorts the data file by target sentence, examines their own intra-rater reliability, and makes changes based on any discovered inconsistencies.

Tensions between operationalization and generalizability to text context

Our data sets include different texts from multiple genres, including science and history. Thus, our texts vary not only in terms of task, but also in terms of content, genre, and text-specific features. The operational definitions in our rubric needed to be broadly conveyed so that they could be reliably applied to multiple types of texts. This led to protracted discussions regarding text and sentence-specific patterns that impacted how students expressed their thoughts. For example, both texts within the training corpus included cause-and-effect relations. However, responses within the history text included comments about "character" intentionality, which were not evoked within the science texts; and vice versa, science texts evoked particular phrases and sentence structures that were not present in the history texts. These differences forced us to reconsider the extent to which a statement was semantically-related and relevant to the broader context of the text. Such concerns have led us into continuing discussions regarding the extent to which these processes are similar or different across tasks, contexts, and genres and what these differences say about comprehension more generally.

Iterative refinement within a larger design-based framework

The final set of insights pertained to the benefits of our design loop. The multiple rapid cycles of discussion and refinement without the larger design cycle were particularly helpful in a project of this scale. As illustrated in Figure 1, we have emphasized rapid cycles of evaluation and refinement at multiple points throughout the coding process. This flexibility in approach has allowed us to quickly respond to concerns and adjust to issues as they emerged. Those who are well-versed in instructional design and design-based (implementation) research are aware that these processes occur and can be crucial to the success of an intervention. However, cyclical, reactive iterations are often not considered within the context of coding verbal protocol responses.

In most verbal protocol coding projects, the objective is to capture the target construct, but the focus is on establishing and maintaining reliability between coders. Once reliability has been achieved, there is little consideration of how changes that may have occurred during training influence the larger project. By checking our reliability during the implementation phase, we were able to identify stark drops in reliability, which forced us to consider the ramifications of the issue, including its theoretical implications. Discussions amongst the team highlighted the value of the quantitative data as a means to drive qualitative understandings. The quantification of discrepancies between coders provided a basis for discussing how constructs were defined. Frequent use of reliability metrics and confusion matrices allowed us to more quickly identify and diagnose points of disagreement, confusion, and misconception. These quantitative metrics afford a valuable way to engage in rich discussion, allowing the team to be more responsive to issues of both reliability and validity as coding continued.

By contextualizing this work within a design cycle, we were able to capitalize on the notion of "closing the loop" (e.g., Liu & Koedinger, 2017). In much design work, tools and systems are built, deployed, and evaluated, but many fall short of subsequently capitalizing on the discoveries made during evaluation to improve the design moving forward. Closing the loop refers to making sure that the data derived from one evaluation is used to inform the next cycle of problem identification. In light of increased cross-institutional collaborations and the growth of large-scale and big data, framing coding procedures as design work affords a way to structure the development and refinement and, perhaps most importantly, the documentation of the iterative nature of rubric development. We continue to use these methods of discussion and iterative refinement as we progress through additional data sets. One question that remains for us is the extent to which it is appropriate and feasible to reopen closed loops. That is, as we progress through our data sets, to what extent might we revisit "completed" data sets when we have a new discovery that changes our rubric. We continue to consider these implications, including their costs and their benefits.

Conclusion

Learning scientists and educational researchers more broadly often rely on rubrics and inter-rater reliability to identify and evaluate learning processes and behaviors. Through exploring our own process of iterative development and refinement, we observed critical tensions between theory and real-world implementation, as well as reflexivity regarding our own assumptions about comprehension processes. We encourage those who engage in these types of coding tasks to similarly step back from the process and flexibly respond to breaks in the coding process as opportunities for knowledge co-construction and negotiation of meaning.

References

Barab, S. (2014). Design-based research: A methodological toolkit for engineering change. In *The Cambridge Handbook of the Learning Sciences, Second Edition* (pp. 151-170). Cambridge University Press.

Best, R., Ozuru, Y., & McNamara, D.S. (2004). Self-explaining science texts: Strategies, knowledge, and reading skill. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the Sixth International Conference of the Learning Sciences: Embracing Diversity in the Learning Sciences* (pp. 89-96). Mahwah, NJ: Erlbaum.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*(1), 1-53.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5-8.

Easterday, M. W., Lewis, D. R., & Gerber, E. M. (2014). Design-based research process: Problems, phases, and applications. Boulder, CO: International Society of the Learning Sciences.

Gee, J. P. (2014). *An introduction to discourse analysis: Theory and method*. Routledge.

Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: implications for understanding and learning processes. In *The psychology of science text comprehension* (pp. 19-50).

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* (pp. 307-325). American Psychological Association.

Kopatich, R. D., Magliano, J. P., Millis, K. K., Parker, C. P., & Ray, M. (2019). Understanding how language-specific and domain-general resources support comprehension. *Discourse Processes*, *56*(7), 530-552.

Liu, R., & Koedinger, K. R. (2017). Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains. *Journal of Educational Data Mining*, *9*(1), 25-41.

Magliano, J. P., Higgs, K., Santuzzi, A., Tonks, S. M., O'Reilly, T., Sabatini, J., Feller, D., Kopatich, R., Ray, M., & Parker, C. (2020). Testing the inference mediation hypothesis in a post-secondary context. *Contemporary Educational Psychology*, *61*, 101867.

Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacognition and Learning*, *6*(2), 131-154.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1-30.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of learning and motivation*, *51*, 297-384.

Moje, E. B., & Speyer, J. (2008). The reality of challenging texts in high school science and social studies: How teachers can mediate comprehension. *Best Practices in Adolescent Literacy Instruction*, 185-211.

Muñoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools. *Behavior Research Methods*, *38*(2), 211-217.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Routledge.

Silber, K. H. (2007). A principle-based model of instructional design: A new way of thinking about and teaching ID. *Educational Technology*, 5-19.

Vogel, F., & Weinberger, A. (2018). Quantifying qualities of collaborative learning processes. *International handbook of the learning sciences*, 500-510.

Acknowledgments

This project was funded by the Institute of Education Sciences through Grants R305A190063 and R305A180144. The opinions expressed are those of the authors and do not represent views of the IES.