An Inquiry into the Use of Intercoder Reliability Measures in Qualitative Research

When compared to quantitative approaches, qualitative approaches are relatively newer to the engineering education research community (Borrego, Douglas, & Amelink, 2009). As the community grapples with the recent emergence of these approaches, they continue to engage in discussions about how to establish rigor and quality in qualitative work (Kellam & Cirell, 2018; Walther et al., 2017). Despite the epistemological and methodological diversity in qualitative engineering education research, many scholars (e.g., Authors, 2018; Koro-Ljungberg & Douglas, 2008; Walther, Sochacka, & Kellam, 2013) agree that multiple procedures are needed to establish quality throughout the qualitative research process, from the initial conceptualization of a study through its publication.

We affirm that multiple methods can and should be used to establish quality and rigor in qualitative research. In this paper, however, we focus exclusively on one common method of establishing quality: demonstrating intercoder reliability. The purpose of this paper is threefold. First we describe how intercoder reliability has been conceptualized in social science research writ large, and how it has been applied to engineering education research specifically. Second, we situate intercoder reliability within the landscape of larger epistemological and methodological considerations, and we raise questions about its appropriateness and use in the context of qualitative studies. Third, using our own qualitative multiple case study research as an illustrating example, we outline considerations for researchers who seek to establish and communicate research quality through the use of intercoder reliability measures.

Intercoder Reliability

Intercoder reliability—a term which has been used interchangeably with other terms such as interrater reliability, interrater agreement, interjudge agreement, or intercoder agreement (Cho, 2008; Lombard, Snyder-Duch, & Bracken, 2017)—refers to the extent to which two or more independent coders make the same decisions when applying the same coding scheme to a dataset or a subset of a dataset. Our search in the *Journal of Engineering Education* (see below for more details) indicated that the term 'interrater reliability' is the most common among researchers who publish in that journal and report on inter-judge agreement.

Despite several assertions that intercoder reliability and interrater reliability are essentially the same, we propose a distinction between the two terms. According to the Merriam-Webster dictionary, the word *rate* has several definitions, including "a fixed ratio between two things," a "quantity, amount, or degree of something measured per unit of something else," or "relative condition or quality." Accordingly, we apply the term *interrater reliability* to instances when two or more coders assign a number to data in an evaluation of its quality. As an example, we imagine a scenario in which two readers use the *Engineering Design Process Portfolio Rubric* (Abts, 2011) to evaluate the quality of a student's presentation and justification of a problem and solution requirements (component one of the rubric), using a score from 0-5 according to the rubric.

We envision *coding* as being different from *rating.* Like *rate, code* has different definitions in dictionaries and in qualitative coding guides (Saldaña, 2015; Thornberg & Charmaz, 2014).

However, we understand coding to include descriptions that are not necessarily related to quality and that do not necessarily have a numerical assignment. For example, in our previous research (Authors, 2018) we identified forms of capital that high school students mobilized toward solving an engineering design problem. Our codes included *Social Capital: Peer*, defined as "other high school students who provided ideas or information regarding potential design elements" and *Embodied Capital: Literacy Practices,* defined as "locating, interpreting, and/or producing texts relevant to the design." In these cases, we did not attempt to evaluate and quantify the quality of the high school students' capital, but rather to describe and theorize it. Thus, we argue that *coding,* rather than *rating,* is a more appropriate term in this instance.

Because the two terms (intercoder and interrater) are used interchangeably in much of engineering educational research literature, we include research literature that uses either term throughout this paper. However, we prefer the term *intercoder reliability* in the context of much of qualitative research because this term more fully encapsulates the possibility for inductively-generated descriptions that identify, illuminate, and describe—rather than evaluate and quantify—phenomena.

Both intercoder reliability and interrater reliability—that is, both the assignment of descriptive qualitative codes *and* evaluative numeric scores—can be reported as percentage agreement, which identifies similarities in the application of codes without accounting for chance. Alternatively, they can be calculated using one of the more than 30 indices or measures of intercoder reliability (Cho, 2008), of which variations of Cohen's kappa (Cohen, 1960, 1968; de Vries, Elliott, Kanouse, & Teleki, 2008) are the most common.

Several federal funding agencies (e.g., Hallgren, 2012) have communicated that ICR or IRR measures are a key component of rigor in empirical research. For example, the Institute of Educational Sciences (IES) (2017) specified the types of "inter-assessor agreement reporting" that a study must include in order to be included in the What Works Clearinghouse. According to IES guidelines, two independent reviewers should code a minimum of 20% of data points across all phases and cases in a rigorous study. If the research team reports percentage agreement, then the minimum acceptable value is 0.80. If they use a statistic that accounts for chance, "the minimum kappa or correlation is 0.60" (p. 6).

Though scholars tend to agree that there are many approaches for establishing research rigor, many people in the educational research community rely heavily on intercoder reliability (ICR) measures as especially crucial to determining the rigor of a study. Kolbe and Burnett (1991) articulated this reliance on ICR in their statement that, "Interjudge reliability is often perceived as the standard measure of research quality. High levels of disagreement among judges suggest weaknesses in research methods, including the possibility of poor operational definitions, categories, and judge training" (p. 248). Other scholars have echoed this sentiment. For example, in writing of survey research, Cho asserted that adequate ICR is a "critical component…without which the interpretation of the content cannot be considered objective and valid" (p. 345). This sentiment has been echoed by others scholars when writing of ICR's importance to inductive analytic methods such as content analysis (Lombard, Snyder-Duch, & Bracken, 2002) or constant comparative analysis (Olson, McAllister, Grinnell, Walters, & Appunn, 2016).

In the context of engineering educational research specifically, intercoder reliability statistics have been reported in conjunction with the analysis of many different datasets, such as states' academic standards (Carr, Lynch, & Strobel, 2012); teachers' written responses to students' oral discussions of engineering problems (Aguirre-Muñoz & Pantoya, 2016); transcripts of interviews with university research mentors (Ahn & Cox, 2016); transcripts of high school students' think-alouds while addressing engineering design tasks (Mentzer, Becker, & Sutton, 2015); and published empirical studies (Bodnar, Anastasio, Enszer, & Burkey, 2016). In addition to having been applied to several types of datasets, ICR has been used in the context of research designs with different theoretical frameworks. Table 1 indicates some of the ways in which intercoder reliability has been used in the context of engineering educational research specifically. This table is not intended to be exhaustive but rather to illustrate some of the ways in which intercoder reliability has been used in conjunction with a range of research questions, study designs, and theoretical frameworks in engineering educational research.

Table 1. Studies in the *Journal of Engineering Education* that use terms 'interrater' or 'intercoder' reliability.

Study	**Context of IRR**	**Information about Interrater Reliability**
Ahn & Cox (2016)	The research team used constant comparative analysis to analyze interview transcripts in a qualitative portion of a mixed-method study.	"The IRR test produced Cohen's kappa values of 0.90 (95% CI, 0.89-0.95) and .85 (95% CI, 0.60 to 1.00) between the researcher and each of the two collaborators."
Bodnar, Anasasio, Enzer, & Burkey (2016)	The research team determined whether articles should be included in a systematic review.	"The interrater reliability between the first two reviewers was .74 as measured by Cohen's kappa."
Carberry & McKenna (2014)	The research team developed a rubric to code each student's response to identify what type of models they mentioned and how they use models in design.	"Changes to the rubric were made to establish 100% interrater reliability between the two raters."
Kilgore, Atman, Yasuhara, & Morozov (2007)	The research team coded students' written responses to an engineering problem posed to them (the Midwest Floods problem).	"The researchers, coding separately, achieved substantial agreement for both the frame of reference and the physical location codes, with kappa values of .748 and .746 respectively."
Kong, Douglas, Rodgers, Diefes-Dux, & Madhavan (2017)	In the "qualitative study of student team projects," the research team used constant comparative analysis to analyze student work products, specifically their	**"**The kappa values were found to be 100% for the definition category, 93% for the evaluation category, and 84% for the comparison category."

	graphical user interfaces.	
Koretsky, Kelly, & Gummer (2011)	The authors conducted a content analysis to contrast the survey responses of undergraduates who attended a virtual laboratory versus those who attended a physical laboratory.	**.**93, .85, and .89 Cohen's Kappa score for three different laboratories offered under each of the two conditions.
Mentzer, Becker, & Sutton (2015)	The authors coded the engineering design thinking of 59 high school students' think alouds as they participated in an engineering design task.	The authors reported the interrater reliability, as indicated by Cohen's kappa, for each individual code, which ranged from .80 to .95. They also reported the average interrater reliability of all codes.

As indicated by this table, ICR is a prevalent method of establishing rigor in engineering educational research. Though intercoder reliability is often used to establish rigor in quantitative research, it has also been recommended as a method for establishing rigor in qualitative research in engineering education as well. For example, in their outline of procedures for establishing quality in interpretive research, Walther, Sochacka and Kellam (2013) wrote, "Another way to improve the dependability of the interpretation of data is coding by several researchers to achieve interrater reliability" (p. 650). This assertion echoes the recommendations of popular guidebooks for qualitative research in social science research, such as Saldaña's (2015) *Coding Manual for Qualitative Researchers,* which indicates that 80% agreement between two coders is one method for establishing the trustworthiness of qualitative research. Given the widespread use of ICR and its centrality to the establishment of rigor—at least, according to many scholars in engineering educational research and beyond—we intend for this paper to "incite discourse" (Freeman, deMarrais, Preissle, Roulston, & St. Pierre, 2007) or further discussion regarding its appropriateness by raising pertinent questions and issues related to its use in different contexts.

Unpacking the Use of Intercoder Reliability

Certainly, we are not the first to raise questions related to intercoder reliability. Several scholars (Hallgren, 2012; Lombard, Snyder-Duch, & Bracken, 2002; Maxwell, 2010) have critiqued the ways in which ICR is conducted and reported. Many of these critiques focus on the appropriateness of different statistics in the context of particular studies. For example, Cohen's kappa, though popular, is not appropriate when more than two people code the data or when the distributions of codes fall under one category at a much higher rate than another (Hallgren, 2012). As another example of a critique based on statistical analyses, Salminen et al (2018) criticized several interrater agreement statistics because they were "characterized by subjectivity" (p. 1) and because they tended to be more sensitive to the number of categories while being less sensitive to the number of items.

With respect to these debates, we do not intend for this section to outline whether or when a particular statistic is appropriate for a particular coding scheme and procedure, but to question

whether it is epistemologically consistent at all with the assumptions about knowledge behind much of qualitative research. To begin this query, we take Baillie and Douglas's (2014) definition of epistemology as "the assumptions we are making about the nature of knowledge and what counts as evidence, with the aim of formulating or refining scientific research questions" (p. 2). The epistemology of *positivism*—which has historically been used and accepted in engineering educational research—is based on the assumption that reality can be categorized, quantified, and accurately measured through objective methods (Lather, 2007).

By contrast, other epistemologies are grounded in other assumptions. Post-structuralism, for example, is based on the assumption that categories are not natural, given, or unproblematic; instead, it seeks to interrupt or deconstruct binaries (e.g., male/female) or categories (e.g., poor spatial skills; mid-level spatial skills; good spatial skills) that are hallmarks of positivistic research (Lather, 2007). Social constructionism, another example of an epistemology, is based on the assumption that there is not a one-to-one relationship between reality and a study's findings because all human instruments, experiences, and perceptions are historically situated and mediated through linguistic and cultural tools; in other words, our reality is socially constructed. Accordingly, the positivistic metaphor of research as a mirror—an accurate and unbiased 'reflection' of a phenomenon—is not possible under this latter epistemology. This does not mean that researchers can never draw sound conclusions, but rather, there are multiple knowledges rather than a single Knowledge that can be generated from a dataset, which itself is a product of cultural and linguistic mediation (Willig, 2001).

Though we name but a few epistemologies here, qualitative research is richly diverse in terms of the theoretical frameworks that are used. It is widely agreed that theoretical frameworks should be consistent with, and informed by, the larger epistemologies, or assumptions about the construction of knowledge, and perspectives adopted by the researchers. Following Case and Light (2011), we affirm that methods sections of published research studies are not simply decontextualized lists of procedures aimed at uncovering objective truths, but instead methodologies that should be conceptualized as part of "a theoretical justification for the methods used in a study" (p. 187). Kellam and Cirell (2018) describe the act of writing a research article as a "back-and-forth recursive process" wherein the theoretical frameworks, methods, and results sections all "talk to one another" so as to create a "unique theory-methods package" (p. 358). Seen under this light, methods sections specify theoretically-informed actions that should be consistent with the study's epistemology and theoretical framework.

Koro-Ljungberg and Douglas (2008) made a similar argument in their early analysis of qualitative research in engineering education when they critiqued much of this research based on their finding that it lacked "epistemological consistency" (p. 163). Accordingly, the extent to which a qualitative study demonstrates "quality" should not be necessarily be conceptualized in terms of whether its procedures (e.g., ICR measures) demonstrated "rigor" as defined under positivism, but instead whether the major sub-components of the study were theoretically aligned. Along this vein, Baillie and Douglas (2014) asserted that qualitative researchers in engineering education unknowingly enact tenets of positivism in their research. In their words:

"Commonly in engineering education work, engineers adopt a positivist epistemology because it feels more 'rigorous,' sometimes without even being aware that they are doing so" (p. 2).

When this statement is applied to discussions of ICR, we argue that qualitative researchers, who often ostensibly embrace non-positivist epistemologies that inform diverse theoretical frameworks, might more fully consider whether ICR measures are theoretically consistent with their stated epistemologies and frameworks. At the same time, while raising this question, we do not seek to establish a simplistic and unhelpful binary between qualitative and quantitative research (Hammersley, 1992) by implying that quantitative research is "positivist" and qualitative research is "interpretivist," a term based on the assumption that researchers' subjectivities influence their interpretation of locally-constructed social realities (e.g., Walther, Sochacka, & Kellam, 2013). Rather than establishing this binary, we think it might be helpful to consider positivism and interpretivism along continua or spectra, in which ICR measures might be helpful in the context of some qualitative studies but inconsistent in the context of others. To further raise questions about the use of ICR, we next describe our own qualitative work in engineering educational research and we describe our discussions and considerations surrounding ICR in our attempts to ensure quality in our own qualitative research.

Intercoder Reliability and Quality: Reflections on a Qualitative Multiple Case Study

To contextualize our discussion of ICR measures and quality, we begin with a brief description of our own ongoing qualitative work: a multiple case study whose purpose is to identify the literacy practices associated with engineering. The research participants are eight engineers who work in a variety of disciplines (e.g. aerospace, biological, chemical, civil, computer, electrical, environmental, mechanical), at different stages in the product life cycle, and at different engineering firms. We (a literacy researcher, a registered professional engineer and engineering education researcher, and an engineering education doctoral student) are currently observing each engineer for six months; interviewing them monthly; and conducting retrospective and concurrent analyses as they read and write texts in order to identify the interpretive frameworks they use when reading and writing. By identifying these frameworks and modifying them in developmentally appropriate and culturally responsive ways for a K-16 audience, we hope to broaden participation in engineering by advancing authentic learning environments in which diverse students are supported in complex engineering activity.

The epistemology that informs this study is constructivism, in which researchers "write reality as observed" (Lather, 2007, p 163) using methods such as observation, interviews, and emergent design. Our theoretical framework, which is informed by this larger epistemology, is New Rhetorical genre studies (RGS), which foregrounds textual genres within communities of practice. According to theories of RGS, genres are typified social actions that shape and are shaped by the larger goals, activities, and values of the community (Bazerman, 1988; Devitt, 2009; Miller, 1984). We sought to establish data generation and analysis procedures that were consistent with the study's epistemological stance and theoretical framework. In our research team discussions and our attempts to ensure a quality study, we considered using ICR. Using this study as an illustrating heuristic, we turn to a larger discussion of ICR as it relates to quality, rigor, validity, and reliability in the following section.

The Relationship Between Quality and ICR Measures

To identify the relationship between research quality and ICR measures, it is useful to first begin a conversation around possible definitions of quality. To begin our discussion of quality, we return to Kolbe and Burnett's (1991) assertion that "Interjudge reliability is often perceived as the standard measure of research quality" (p. 248). Though quality is defined in diverse ways throughout theoretical and methodological literature, we believe this assertion does not consider various facets of *quality*, which can be conflated with *rigor* (Moss et al., 2009). Following Floden (as quoted in Moss et al., 2009), we assert that:

> A judgment of quality, for example, may include assessing whether or not a study addresses a question of broad interest and social significance. A study might be rigorous in the sense that it uses a design that guards against many threats to validity, yet be of low quality because the question it addresses is trivial (p. 505).

Following Floden, we consider quality and rigor to be related, yet distinct. In engineering educational research, we understand the term *quality* to include the social significance of the study; its pragmatic potential to lead to long-term impacts that benefit society as a whole and underrepresented groups in particular (Authors, 2018); its proactive responsiveness to research participants and relevant stakeholders (Gutiérrez & Penuel, 2014); theoretical consistency across all aspects of the study (Koro-Ljundberg & Douglas, 2008); transparency, including in documentation and communication (Walther, Sochacka, & Kellam, 2012); ethical in its conduct and implications (Walther, Pawley, & Sochacka, 2015); *as well as* a carefully-planned research design that responds to the research questions, whereby the generation of data enables the researchers to make supported claims. Although rigor is bound up in all aspects of a study—from its level of cultural responsiveness to communication with internal and external stakeholders throughout the research process—our definition of rigor is narrower than our definition of quality. Specifically, we understand *rigor* to mean that a study's claims and implications have been carefully supported with data, and that alternative explanations have been considered and addressed throughout the research design.

Validity and reliability have historically been perceived as requisites for establishing rigor, especially in qualitative content analyses. In writing of ICR measures in content analyses, Krippendorff (2004) explained that an instrument is considered to be *valid* to the extent that it captures what it sets out to capture. When applied to inductive analyses, codes are considered valid to the extent that they capture the phenomenon that they represent. Schreier (2012) asserted that comparisons of coding processes across persons—which can include establishing and communicating ICR measures—is one method for ensuring validity.

Using our study as an example, we intend to question a potential for overreliance on ICR as a requisite for establishing validity (and by extension, rigor). Because we sought to identify engineers' literacy practices in our multiple case study, we attempted to establish *validity* in our codebook through familiarizing ourselves with previous empirical literature on engineers' reading and writing practices and using that literature to inform our initial codebook. Moreover, we also sought the input of people with different expertise, including multiple literacy experts as

well as industry practitioners from each discipline represented in our study design. Specifically, our interdisciplinary research team drafted initial codes, which were shared with literacy experts, industry experts, and engineering educational research experts. These experts included practicing engineers from the disciplines being studied (e.g., a computer software engineer and aerospace engineer). Though our study is not over, we will further establish validity through sharing our results, including our frequency counts and codebooks, with the research participants themselves and asking for their feedback on our codes and our overall analyses. Through these methods, we hope to generate codes that resonate with the perspectives of relevant stakeholders, including the research participants themselves.

Through this iterative process, we developed codes that more closely captured the perspectives of those within different engineering professions. For example, we (the literacy specialist and two engineers with a background in mechanical engineering) used terms in our initial codebook that did not reflect the specialized language or textual genres of software engineering. Through consultation with an industry expert, we were able to develop a second draft of a codebook, complete with illustrating examples, which more fully resonated with the software engineer's emic perspective of software engineering. Even if we (the literacy specialist and mechanical engineering specialists) had 100% agreement regarding our initial codes, this level of agreement would not have ensured our codes were *valid* in the sense that we used language or definitions that cohered with those used by the software engineering community. Thus, we assert that IRC should not be used as a proxy for quality on even on such a narrow indicator of quality as *validity of codes*. Instead, many different procedures, including consultation with the participants themselves where possible, can be used to ensure that the codes incorporate relevant perspectives and reflect, to the greatest extent possible, the socially-constructed realities of those involved.

As indicated by its name, intercoder reliability measures are intended to indicate *reliability,* even though they have been conflated with quality, rigor, and validity. In other words, ICR measures are intended to indicate that similarly-trained coders can apply the codes in the codebook to similar datasets in similar contexts, and they should be able to achieve similar ICR scores. There may be instances in qualitative research in which this type of information is valuable as one method of establishing reliability. In other words, there may be instances in which this information is deemed to be theoretically consistent with the overall study, and (perhaps more pragmatically) qualitative researchers may feel that ICR measures communicate rigor to potentially positivistically-oriented reviewers and readers.

After deciding that ICR is an appropriate method for establishing reliability, qualitative researchers should choose the right statistics for the dataset, which includes a consideration of number of coders, number of categories of codes, and distribution of codes (Lombard, Snyder-Duch, & Bracken, 2002). However, even assuming the appropriate statistic is chosen, we found in our own discussions of ICR that many factors influence ICR measures, and these factors are very rarely reported in published research. We use our own study as a heuristic to illustrate some of these factors.

We sought to code multiple facets of the texts that engineers read and wrote, including the genre of the text (e.g., manual, regulations) and the purpose for which the engineer used the text (e.g.,

to fix errors, to document actions). Several studies of the engineering educational research studies in Table 1 did not describe or even mention segmenting processes, we found that our method of segmenting tremendously influenced all aspects of our data analysis—from number of codes reported (frequency counts) to our ability to get two readers to see similar phenomena in the dataset. At the same time, however, we recognize that many data may come "pre-segmented," in the sense that it is already divided into meaningful units, such as brief responses to survey questions, so although we raise segmenting as a methodological concern, we realize it may not be salient to all qualitative research projects.

We used a qualitative software coding package, Dedoose, to help us track our coding processes. When analyzing field notes from observations, we realized immediately that in order for two researchers to draw similar interpretations of the data, we had to first segment the field notes into chunks. Otherwise, one researcher might have highlighted five lines in the raw data (the field notes), while another researchers might have highlighted four lines, which would have influenced our ability to reach a statistically acceptable level of agreement. In accordance with our coding system and research purpose, we intended for each chunk to indicate that the engineer interacted with one text. Thus we segmented our field notes according to *text,* with a new chunk occurring each time a new text was read or written.

Though segmenting or chunking may seem simple, in our case, it was actually quite complex because we had to generate and agree on a definition of *text,* which evolved as we learned more about the engineers' literacy practices. Moreover, segmenting was also complex because the engineers frequently consulted multiple texts over the matter of a few seconds, for example, by repeatedly referencing multiple webpages and PDF documents (Code for Acceptance of Construction Quality of Steel Structures; International Standards for People's Republic of China) while writing their own document. Thus, some of our segments were a few words, while others were several paragraphs.

Given that engineering educational research has reported ICR measures on diverse datasets, and that ICR measures are inextricably tied with segments or units that are coded, we argue that in order to increase transparency, researchers who report ICR should also report on segmenting processes. This reporting can include answering questions such as: How were the data segmented into smaller units? and Who segmented the data? (e.g., multiple people in agreement or just one person), and How did the research team establish agreement in relation to data segmentation?

Other procedures that can improve the transparency of ICR include selecting data that are representative of the dataset, not randomly selected data, when determining ICR measures for a subset of the data. In our case, "representative of the dataset" would include codes from each of the eight engineers from field notes in which they engaged in 'typical' activity. We quickly learned this lesson when we randomly selected segmented field notes from an observation for two of us to independently code, and the observation that we selected turned out to be anomalous in the sense that it represented an activity that an engineer rarely engaged in and thus very few literacy-related codes were assigned. As another point of consideration, those who embrace ICR recommend reporting ICR statistics for each code, and not reporting one ICR statistic for the

dataset as a whole. We found this was not always done in the engineering education research literature we studied.

In the end, when discussing quality in the context of our particular study, we determined that we would attempt to establish rigor through the careful incorporation of multiple relevant viewpoints, and through having two coders read through the entire dataset, applying codes, and resolving all disagreements through discussion. In our view, this method more fully reflected our epistemological assumptions that knowledge construction is culturally and socially mediated and should reflect the perspectives of those involved to the greatest extent possible. Though we reached this conclusion, at the same time, we embrace the diversity of methods in qualitative research and realize there are cases when ICR measures may be theoretically consistent with the study, in which case they should be applied and reported in appropriate and transparent ways.

Conclusions

In writing this methodological paper, it was not our intention to make data analysis more onerous or proscriptive. Instead, we intended to raise the point that ICR measures can be overemphasized as essential for quality in both qualitative and quantitative research, when in fact at best they represent a small sub-component of rigor: reliability. Even then, more considerations are needed (e.g., information about segmenting) in order to determine whether ICR measures are a fair indicator that other research teams could apply the codebook to similar datasets with similar results. Moreover, in many qualitative studies that embrace non-positivistic epistemologies, ICR may not be theoretically consistent at all. In fact, ICR measures are probably not theoretically consistent with interpretive research if they represent the *sole* method for establishing the quality of a codebook, as opposed to other possible methods such as eliciting and incorporating feedback from relevant stakeholders; and providing clear examples and definitions of codes for readers in manuscripts so they can make their own determinations as to the quality of the codebook.

Aware that this paper is but one interpretation and viewpoint of research, we hope that it can incite discourse by encouraging other qualitative researchers to reflect on their own epistemological and theoretical stances and their implications for establishing both quality and rigor. As contextualized within this larger discussion, qualitative researchers in engineering education can critically consider whether and how ICR measures can be incorporated into their own work.

References

Abts, L. (2011). *Engineering design process portfolio scoring rubric.* Retrieved from: http://teams.mspnet.org/media/data/EDPPSR.pdf?media_000000008449.pdfv

Ahn, B., & Cox, M. F. (2016). Knowledge, skills, and attributes of graduate student and postdoctoral mentors in undergraduate research settings. *Journal of Engineering Education, 105,* 605-629.

Author, 2018a.

Author, 2018b.

Baillie, C., & Douglas, E. P. (2014). Confusions and conventions: Qualitative research in engineering education. *Journal of Engineering Education, 103*(1), 1-7.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science.* Madison, WI: University of Wisconsin Press.

Bodnar, C. A., Anastasio, D., Enszer, J. A., & Burkey, D. D. (2016). Engineers at play: Games as teaching tools for undergraduate engineering students. *Journal of Engineering Education, 105,* 147-200.

Borrego, M., Douglas, E. P., & Amelink, C. T. (2009). Quantitative, qualitative, and mixed research methods in engineering education. *Journal of Engineering Education, 98*(1), 53-66.

Case, J., & Light, G. (2011). Emerging research methodologies in engineering education research. *Journal of Engineering Education, 100*(1), 186-210.

Cho, Y. I. (2008). Intercoder reliability. In P. J. Lavrakas (Ed), *SAGE encyclopedia of survey research methods* (pp. 345-346). Thousand Oaks, CA: SAGE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin, 70,* 213-220.

Devitt, A. (2004). *Writing genres.* Carbondale, IL: Southern Illinois University Press.

deVries, H., Elliott, M N., Kanouse, D. E., & Teleki, S. S. (2008). Using pooled kappa to summarize interrater agreement across many items. *Field Methods, 20,* 272-282.

Freeman, M., deMarrias, K., Preissle, J., Roulston, K. & St. Pierre, E. (2007). Standards of evidence in qualitative research: An incitement to discourse. *Educational Researcher, 36*(1), 25-32.

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher, 43*(1), 19-23.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol, 8,* 23-34.

Hammersley, M. (1992). Reconstructing the qualitative-quantitative divide. In M. Hammersley (Ed.), *What's wrong with ethnography? Methodological explorations* (pp. 159-173). London: Routledge.

Institute of Educational Sciences. (2017). *Reviewer guidance for use with the Procedures Handbook (version 4.0) and Standards Handbook* (version 4.0). Retrieved from: https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_reviewer_guidance_103017.pdf

Kellam, N., & Cirell, A. M. (2018). Quality considerations in qualitative inquiry: Expanding our understandings for the broader dissemination of qualitative research. *Journal of Engineering Education, 107*(3), 355-361.

Kilgore, D., Atman, C. J., Yasuhara, K., Barker, T. J., Morozov, A. (2007). Considering context: A study of first-year engineering students. *Journal of Engineering Education, 96,* 321-334.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18,* 243-251.

Kong, Y., Douglas, K. A., Rodgers, K. J., Diefes-Dux, H., & Madhavan, K. (2017). A size and scale framework for guiding curriculum design and assessment. *Journal of Engineering Education, 106,* 431-453.

Koretsky, M., Kelly, C., Gummer, E. (2011). Student perceptions of learning in the laboratory: Comparison of industrially situated virtual laboratories to capstone physical laboratories. *Journal of Engineering Education, 100*(3), 540-573.

Koro-Ljungberg, M., & Douglas, E. P. (2008). State of qualitative research in engineering education: Meta-analysis of JEE articles, 2005-2006. *Journal of Engineering Education, 97*(2), 163-175.

Krippendorff, K. (2004). *Content analysis: An introduction in its methodology.* Thousand Oaks, CA: Sage.

Landis & Koch. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 378-382.

Lather, P. (2007). *Getting lost: Feminist efforts toward a double(d) science.* Albany, NY: State University of New York Press.

Lombard, M., Snyder-Duch, J., Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587-604.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2017). Intercoder reliability. In M. Allen (Ed.), *SAGE encyclopedia of communication research methods* (pp. 722-724). Thousand Oaks, CA: SAGE.

Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry, 16*(6), 457-482.

Mentzer, N., Becker, K., & Sutton, M. (2015). Engineering design thinking: High school students' performance and knowledge. *Journal of Engineering Education, 104*(4), 417-432.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70,* 151-167.

Moss, P. A., Phillips, D. C., Erickson, F. D., Floden, R. E., Lather, P. A., & Schneider, B. L. (2009). Learning from our differences: A dialogue across perspectives on quality in educational research. *Educational Researcher, 38*(7), 501-507.

Nelson, K. G., McKenna, A., Brem, S. K., Hilpert, J., Husman, J., & Pettinato, E. (2017). Students' misconceptions about semiconductors and use of knowledge in simulations. *Journal of Engineering Education, 106*(2), 218-244.

Olson, J. D., McAllister, C., Grinnell, L. D., Walters, K. G., & Appunn, F. (2016). Applying constant comparative method with multiple investigators and inter-coder reliability. *The Qualitative Report, 21*(1), 26-42.

Salminen, J. O., Al-Merekhi, H. A., Day, P., & Jansen, B. J. (2018). Interrater agreement for social computing studies. [Did not find full information for this paper)

Schreier. (2012). *Qualitative content analysis in practice.* Thousand Oaks: SAGE.

Sheshkin, D. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton, FL: CRC Press.

Streveler, R. A., & Smith, K. A. (2006). Conducting rigorous research in engineering education. *Journal of Engineering Education, 95,* 363-371.

Thornberg, R. & Charmaz, K. (2014). Grounded theory and theoretical coding. In U. Flitz (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 153-169). Los Angeles, CA: SAGE.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, CA: Academic Press.

Walther, J., Pawley, A. L., & Sochacka, N. W. (2015). *Exploring ethical validation as a key consideration in interpretive research quality.* Paper presented at the ASEE Annual Conference & Exposition, Seattle, WA. Retrieved from: file:///C:/Users/A01680927/Documents/Presentations/ASEE%202018/Walther%20et%20al%202015.pdf

Walther, J., Sochacka, N. W., & Kellam, N. N. (2013). Quality in interpretive engineering education research: Reflections on an example study. *Journal of Engineering Education, 102,* 626-659.

Walther, J., Sochacka, N. W., Benson, L. C., Bumbaco, A. E., Kellam, N., Pawley, A. L., & Phillips, C. M. L. (2017). Qualitative research quality: A collaborative inquiry across multiple methodological perspectives. *Journal of Engineering Education, 106*(3), 398-430.

Watson, M. K., Pelkey, J., Noyes, C. R., & Rogers, M. O. (2016). Assessing conceptual knowledge using three concept map scoring methods. *Journal of Engineering Education, 105*(1), 118-146.

Willig, C. (2001). *Introducing qualitative research in psychology: Adventures in theory and method.* Buckingham, UK: Open University Press.