Validity in Quantitative Content Analysis

Liam Rourke
Terry Anderson

Over the past 15 years, educational technologists have been dabbling with a research technique known as quantitative content analysis (QCA). Although it is characterized as a systematic and objective procedure for describing communication, readers find insufficient evidence of either quality in published reports. In this paper, it is argued that QCA should be conceived of as a form of testing and measurement. If this argument is successful, it becomes possible to frame many of the problems associated with QCA studies under the well-articulated rubric of test validity. Two sets of procedures for developing the validity of a QCA coding protocol are provided, (a) one for developing a protocol that is theoretically valid and (b) one for establishing its validity empirically. The paper is concerned specifically with the use of QCA to study educational applications of computer-mediated communication.

The primary role of networked computers in higher education has shifted from presenting structured, preprogrammed learning materials to facilitating communication. In turn, the role of educational technology researchers has expanded to include the role of communication researcher. In the late 1980s, studies began to appear that incorporated new perspectives, new methods, and new techniques. One of the most promising was *quantitative content analysis* (QCA).

Berelson (1952) defined QCA as "a research technique for the systematic, objective, and quantitative description of the manifest content of communication" (p. 18). In this context, description is a process that includes segmenting communication content into units, assigning each unit to a category, and providing tallies for each category. Bullen (1998), for instance, studied participation and critical thinking in an online discussion by counting the number of times each student contributed to the discussion and by assigning each contribution to one of three categories of critical thinking.

By 1999, enough of these types of studies had been conducted in the field of educational technology that we were able to prepare a literature review (Rourke, Anderson, Garrison, & Archer, 2001). We used published reports to illustrate many of the basic issues of QCA, including objectivity, reliability, units of analysis, types of content, research designs, and ethics. Our intent was to summarize the experiences of researchers like us who were struggling to apply this unfamiliar technique in a disciplined and efficient manner. We found the technique promising but chided researchers on the rigor of their reports, particularly on the lack of reliability data. Those who persisted with the technique found their own way to these conclusions, and today

reviewers are more critical and reports more informative.

Even so, demonstrating the reliability of data collection is only one premise in a researcher's ultimate argument toward validity. It is hoped that argument will persuade readers that the inferences drawn from a QCA procedure are supported by empirical evidence and theoretical rationale. When QCA is used to tally the occurrence of wholly manifest content (e.g., counting the number of messages posted by a particular student), the argument is straightforward. When QCA is used to draw inferences about constructs (e.g., assessing the level of critical thinking in a computer conference transcript), the argument is not so clear-cut. In this article we review procedures for making a sound validity argument in the latter case.

The article is divided into three sections. The first section begins with the observation that QCA is a form of testing and measurement but notes that the procedures of test development codified in the psychometric literature are given meagre consideration in QCA research. The second section describes the process of constructing a coding protocol that is theoretically valid, or as Sheppard (1993) says, reasonable. This section draws on Crocker and Algina's (1986) presentation of essential steps in test construction. The third section follows Messick's (1989) discussion of several types of empirical studies that can be conducted to establish the validity of inferences derived from a testing procedure.

QCA AS TESTING AND MEASUREMENT

Description Versus Inference

A survey of QCA studies in the educational technology literature shows that they often involve making inferences about constructs. In the example cited earlier, Bullen's (1998) QCA culminated in conclusions about the students' level of critical thinking. This goes well beyond the purpose for which the standards of QCA were originally developed. Originally, media researchers such as Berelson (1952) used the technique in a more modest role to describe the surface content of communication. This is evident in Berelson's definition, which specified QCA's purely descriptive role. Studies of computer-mediated communication (CMC) that report the frequency with which students post messages (Bullen, 1998; Weiss & Morrison, 1998) or the average number of words in messages (Parson, 1996) adhere to Berelson's definition.

It is through these types of studies that the test validity question peculiar to QCA coding protocols emerged: Does the procedure describe what it purports to describe (Krippendorf, 1980; Riffe, Lacy, & Fico, 1998)? In this context, the researcher is providing something closer to data than interpretation, and the data speak for themselves (Kaplan, 1964).

When researchers use QCA to make inferences about constructs, the data are no longer speaking for themselves. To make the transition from description to inference, a richer definition of test validity is required, such as the one proposed in Messick's (1989) landmark chapter:

> Validity is an integrated evaluative judgment of the degree to which theoretical rationales and empirical evidence support the adequacy and appropriateness of interpretations and actions based on test scores or other methods of assessment. (p. 13)

The concepts included in this dense definition are particularly important when the frequency of certain types of content are used to offer interpretations about constructs such as higher-order learning, social processes, or critical thinking. For instance, using Henri's (1991) protocol, observers count (among other things) the number of times students in a computer conference *formulate a proposition that proceeds from previous statements* and then offer inferences about students' cognitive and metacognitive skills. Bereiter and Scardemalia (1987) questioned the validity of this type of procedure when they studied the cognitive processes that underlie composition:

> It might seem that the way to begin an explanation of our research is by showing pieces of writing that exemplify *knowledge telling* and *knowledge transforming*. That would be misleading, however. *Knowledge telling* and *knowledge transforming* refer to mental processes by which texts are composed, not to texts themselves. You cannot tell by reading this chapter whether we have engaged in problem-solving and *knowledge-transforming* operations while writing it or whether we have

simply written down content that was already stored in memory in more or less the form presented here. (p. 13)

In this argument, it becomes apparent that assessing the surface characteristics of written composition—something the writing teacher does—and measuring the cognitive processes that underlie written composition—something a cognitive psychologist might want to do—are two different things. Bereiter and Scardemalia (1987) rejected the possibility of learning about one through descriptions of the other. For the content analyst engaged in a purely descriptive study, this argument is inconsequential. For the analyst engaged in an inferential study—those using Henri's (1991) procedure for instance—this argument is fatal. Bereiter and Scardemalia would assert that one cannot say anything about students' cognitive or metacognitive skills based on how many times they *formulate a proposition that proceeds from previous statements.*

Psychometricians would disagree. Much of test theory and the day-to-day practice of testing and measuring is the attempt to make accurate judgments about unobservable constructs based on observable behavior: Candidates' potential for success in graduate school is gauged through their scores on paper-and-pencil aptitude tests; experimental subjects' attitudes are assessed using researchers' questionnaires; and applicants' suitability for jobs is predicted using personality measures, for instance.

Test theory accepts Bereiter and Scardemalia's (1987) argument as fundamental, and prescribes that if there is a gulf between that which one wishes to study and that which is directly observable, then some sort of correspondence between the two must be established before inference begins. The standards of testing and measurement corollary to Messick's (1989) definition of validity set out the steps for accomplishing this. To begin our discussion of these steps, we will first show that QCA is a form of testing and measurement.

Testing, Measurement, and QCA

A *test* according to Crocker and Algina (1986) is "a standard procedure for collecting a sample of behavior from a specified domain" (p. 4). Their definition is deliberately general because the class of things encompassed by the term *test* is diverse. It subsumes an assortment of procedures and aims that range from standardized achievement tests to teacher-made multiple-choice quizzes, from published personality inventories to researchers' questionnaires, and beyond.

One specific form of testing that Crocker and Algina (1986) depicted is "a standard schedule and a list of behaviors that can be used by an observer who codes behavior displayed by subjects in a naturalistic setting" (p. 4). This depiction alone provides a fairly complete characterization of QCA. It is precisely how data were collected in the two examples that have been presented previously. Bullen (1998), for instance, provided observers with a list of 16 behaviors, such as *defining terms and judging definitions,* which they were to look for in the messages that students posted to an educational computer conference. Similarly, the cognitive skills section of Henri's (1991) coding protocol consists of 18 behaviors that coders seek in the paragraphs within students' postings to an online educational discussion (Hara, Bonk, & Angeli, 2000). Each of the elements and processes in Crocker and Algina's depiction of testing is apparent in these two examples. The naturalistic setting in both examples is the online class discussion. The lists of behaviors are the 16- and 18-item lists that are purportedly indicative of critical thinking and cognitive skills as these constructs manifest themselves in online discussions. These lists are provided to the observers or, as they referred to in QCA, *coders* or *raters*; and the standard schedules that accompany these protocols are the individual messages or paragraphs within the messages in the online discussion. Note that physical or syntactical units of analysis (e.g., conference messages or paragraphs within the messages) replace temporal units (e.g., 30-sec intervals) when observation moves from the face-to-face classroom to the computer conference transcript. Clearly, there is room for QCA coding protocols under the general definition of tests and within the specific process that Crocker and Algina portrayed.

Add to this the corollary process of *measurement*-assigning numbers to properties of objects

or events, according to rules, such that the numbers reflect differences in the amount or type of the variable present in different objects (Lord & Novick, 1968; Rogers, 1999; Stevens, 1946). Bullen's (1999) report illustrates how this takes form in QCA:

> [The messages posted by] the students were sorted into one of three categories of critical thinking. The categories and corresponding scores were as follows: (3)-extensive use of critical thinking skills (2)-moderate use of critical thinking skills (1)-minimal use of critical thinking skills. (p. 14)

The objects or events that Bullen focused on were the messages posted by undergraduate students to their computer conference. The rules that were used to regulate the assignment of numbers to these messages included three things: (a) an overarching theory of critical thinking (Norris & Ennis, 1989) with which observers were familiarized, (b) a 16-item set of behaviors indicative of how three levels of critical thinking manifest themselves in text-based online discussion, and (c) an incremental numbering system corresponding to the hierarchical conceptualization of critical thinking. The manner in which numbers were assigned to the students' messages reflected differences in the types of critical thinking, and subsequent frequency counts of each of the categories reflected differences in the amount of critical thinking present across discussion weeks and between students.

Together, the complementary processes of testing and measurement provide a reasonably complete and accurate description of what the content analyst does. The generic definition of test and the depiction of one specific form reveal that QCA, as it is conceptualized and practiced, sits firmly within the rubric of testing and measurement.

This relationship has not been explicated in the past perhaps because it would have been unreasonable to bring the complex machinery of test theory to an operation that was purely descriptive. However, now that content analysts, particularly those in the field of educational technology, consistently use the technique in an inferential capacity, it is advantageous to position it in the appropriate psychometric context. This conceptualization could sensitize researchers to the tendency of data collection and analysis procedures to drift from the descriptive into the inferential, and it could alert them to the significance of this shift. Some content analysts have been unmindful of this distinction, or they have been mindful but have not known how to proceed in a valid manner. Locating QCA under the rubric of testing and measurement provides a well-articulated model of how to move, in a defensible manner, from frequency counts of directly observable behavior to insights about the complex constructs that they allegedly signify. The basic elements of this model are presented in the next section.

DEVELOPING A THEORETICALLY VALID PROTOCOL

The steps to developing a theoretically valid protocol, are:

- Identifying the purpose of the coding data
- Identifying behaviours that represent the construct
- Reviewing the categories and indicators
- Holding preliminary tryouts
- Developing guidelines for administration, scoring, and interpretation of the coding scheme

Identifying the Purpose of the Coding Data

The first step in developing a coding protocol is to identify the purpose for which the coding data will be used. To continue with a previous example, Henri's (1991) first question should be What do I want to do, say, or infer, after I have counted the number of times that students in a computer conference, for instance, *ask relevant questions?* There are two general purposes for any kind of testing: making decisions and conducting research. Questions and hypotheses, focusing mainly on the types of communication that occur in CMC and their influence on learning, have been explored using QCA. Henri's (1991) instrument, for instance, yields informa-

tion on the interactive, participative, social, cognitive, and metacognitive dimensions of communication in computer conferences.

Identifying the purpose of the data informs decisions about scaling, score interpretation models, and the types of validity evidence that are required. QCA protocols used in CMC studies typically use nominal scales of measurement. In these scales, numbers are used to categorize segments of transcripts with the numbers reflecting nothing about the segments other than they are different. Using Gunawardena, Lowe, and Anderson's (1997) scheme for coding social knowledge construction, messages are coded as 1 (*statement of opinion*), 2 (*statement of agreement*), 3 (*corroborating example*), 4 (*clarifying detail*), or 5 (*problem definition*). *Statement of opinion* (1) does not mean less social construction of knowledge than *statement of agreement* (2); the difference between 1 and 2 and 3 and 4 does not represent an equal difference in level of social knowledge construction; and 0 does not represent the absolute absence of social knowledge construction.

The purpose of the data also influences the selection of score interpretation models. Criterion-referenced score interpretations focus on the classification of test takers—pass-fail, master-nonmaster—in conventional testing. Norm-referenced interpretations focus on relative comparisons of test takers—average, above average, below average. The former model is prevalent in applied decision-making contexts—selection, placement, and so forth, and requires established criteria and justified cut-scores with which one can judge mastery. So far, criteria have not been proposed for the issues that educational technologists study with QCA (e.g., interaction, participation, procession through the problem-solving process).

Norm-referenced interpretations characterize the whole of QCA studies in our domain. Generally, one transcript or one segment of a transcript is positioned relative to another. For instance, Chou (2001) compared levels of learner-learner interaction in synchronous versus asynchronous communication modes, and concluded, in a norm-referenced fashion, that there was a higher percentage of social-emotional interaction in synchronous mode than in asynchronous mode. It is not logically necessary to use nominal scales or norm-referenced interpretation models; it is simply conventional. However, there are implications for both of these decisions and therefore they should be made thoughtfully.

The purpose of the coding data also determines the types of validity evidence that need to be gathered. In keeping with the previous discussion, Cronbach (1990) distinguished between using a test to describe and to make decisions about a person. Deciding which candidate should receive a scholarship or which applicants should be admitted to a graduate program is consequential and necessitates a thorough investigation of several elements of an assessment procedure, including its relevance, utility, social consequences, value implications, and construct validity. In the context of basic research, in contrast, the researcher may neglect some of these criteria without placing hopeful students in any peril. In this latter situation, Messick (1989) and Cronbach (1971) encouraged test developers to direct finite resources into a systematic investigation of construct validity. More will be said about this type of validity in subsequent sections.

Identifying Behaviours That Represent the Construct

Once the purpose of coding data has been determined, the next step is to identify behaviors that represent the construct. The goal in this step is to ensure that a coding protocol neither leaves out behaviors that should be included, nor includes behaviors that should be left out. This is particularly important in QCA because the technique is essentially observational; therefore, in an operationist and behaviorist sense, the construct, at one level, comes to be defined by observable behavior. The precariousness of this enterprise can be illustrated with a simple example. One construct that has received continued attention from CMC researchers is participation, which, in QCA studies, is regularly defined by a single representative behavior—posting a message to a conference. If this specific behavior is interesting in itself, perhaps in a human-computer interaction study, then this operational definition is sat-

isfactory. However, this is not the case in a typical analysis in the field of educational technology. Here, participation is usually regarded as a dependent or independent variable, and relationships are sought between that and other variables, such as moderator behavior (e.g., Anderson, Rourke, Garrison, & Archer, 2001), learner characteristics (e.g., McLean & Morrison, 2000), achievement (e.g., Richardson & Swan, 2003), satisfaction (Gunawardena & Zittle, 1998), and higher-order learning (e.g., Bullen, 1998, Gunawardena et al., 1997; Henri, 1991). To understand how participation relates to these factors, researchers may want to add at least one other representative behavior to the conventionally narrow definition. Sutton (2001) provided some evidence of one possible reason why. In his study of vicarious interaction, he found that many students who do not post messages in the online discussions learn adequately by observing and actively processing the interactions between other students. This mode of observation and active processing may be considered a legitimate type of participation depending on the aims of a study. Operationally defining a construct with a single representative behavior could lead to a distorted understanding of existing relationships.

Some of the procedures used to achieve construct representativeness in QCA protocols include literature review and qualitative forms of content analysis. Other methods used outside of QCA, such as the critical incident technique (Flanagan, 1954) and protocol analysis (Ericsson & Simon, 1993), may also be helpful. Conducting literature reviews is a common method for identifying representative behaviors. For instance, Curtis and Lawson (2001) wanted to examine student interaction in computer conferences from a collaborative learning perspective. Therefore, they referred to the extensive body of literature on collaborative and cooperative learning, including three decades of research by Johnson and Johnson (1979; 1986; 1989; 1992a; 1992b; 1994a; 1994b; 1996) books by Slavin (1991) and Sharon and Sharon (1992), and several meta-analyses (Johnson, Johnson, & Stanne, 2000; Johnson, Johnson, & Maruyama, 1983). Within this body of literature, they found a mature theory of collaborative learning, syntactical and operational definitions of the construct, and indicators that required only minor modification before they could be used in their QCA coding scheme.

Qualitative content analysis has also been used to gather behavioral indicators. Mason (1991) had experts examine a large sample of computer conference transcripts in order to catalog the types of communication in which students engage. She then reduced these to a parsimonious set of categories and indicators. Her typology included eminently codable behaviors such as *use of personal experience related to course themes* and *reference to appropriate material outside the course package.* Most QCA researchers move iteratively between these two methods—literature review and direct observation of transcripts—to develop relevant and representative coding indicators.

Beyond the domain of QCA, other procedures have been developed to identify representative behaviors. In the field of industrial and organizational psychology, the critical incident technique (Flanagan, 1954) is widely used to discover behaviors that contribute to success in specific situations. To analyze a domain using the critical incident technique, a researcher first asks people familiar with the context to describe particularly effective behaviors, that is, critical incidents. The researcher then identifies themes represented by the incidents, and these themes are used as the basis for indicators. In CMC studies, students and teachers could be interviewed in an effort to identify critical types of communication that enhance learning. These data could then be used to justify their inclusion in a coding protocol.

The preceding techniques are useful in analyzing manifest content or, as Potter and Levine-Donnerstein (1999) explained, content that resides on the surface of communication. However, CMC researchers often want to analyze latent content, or constructs that are signified by the communication, for example, cognition (Garrison, Anderson, & Archer, 2001; Henri, 1991). To identify behaviors that are representative of these types of constructs, it seems appropriate to refer to the techniques developed by cognitive psychologists. Researchers such as Snow, Federico, and Montague (1980) and

Embretson (1983) argued that the process of test construction should be informed by an understanding of the cognitive activities that are constituents of a task.

Techniques such as computational and mathematical modeling, chronographic analysis, and protocol analysis have been used to this end. Bereiter and Scardemalia (1987), in fact, used the latter two techniques in their program of research on cognition during composition. Measuring the start-up times of students confronted with various writing tasks, and analyzing students' think-aloud protocols enabled them to offer evidentiary statements about the cognitive processes of students. In a CMC context, students could be asked to think aloud as they read and responded to messages in computer conferences in order to identify representative behaviors.

There are few examples of these methods being applied in the educational CMC content analysis literature, but there are many examples of researchers proceeding without them. Garrison et al. (2001) developed a coding protocol to assess *cognitive presence* in computer conferencing, which they defined as the ability of learners to construct meaning through sustained communication. The indicators that constituted the protocol were based on the authors' four-stage model of critical thinking and included items such as puzzlement, brainstorming, connecting ideas, and defending solutions. When coders applied the indicators in an empirical study, they were unable to find any evidence of one of the four stages, coded a third of the transcript as *other*, and coded the remaining two thirds into the final two stages. This left the authors unable to interpret whether the data reflected shortcomings in (a) the coding protocol, (b) the instructional design of the course, or (c) the medium, the lack of (d) cognitive presence, or a combination of (e) all these factors and more.

Cronbach (1990) criticized the process by which constructs are translated into observable behaviors as largely private, informal, and undocumented. Our personal experience with the development of coding protocols corroborates his critique. "Such an approach," argued Crocker and Algina (1986), "results in a highly subjective and idiosyncratic operationalization of the construct" (p. 67). Worse still, it can result in data that are uninterpretable or rival interpretations that are more plausible than those offered by the content analyst.

Reviewing the Categories and Indicators

Most of the steps in this section are connotative of content validity, that is, procedures for ensuring that the categories and indicators in a QCA protocol adequately represent a performance domain or construct. Establishing content validity is largely a subjective operation that relies on the judgment of experts. This process ranges in levels of formality; however, a responsible developer will want a group of experts to evaluate the provisional coding categories and indicators to determine their relevance and representativeness. The term *expert* in this context refers to someone who has experience with QCA, knowledge of the construct, and familiarity with the context in which the coding protocol will be used. If several judges are available, some variation of a Delphi technique (Dalkey & Helmer, 1963) can be used, in which members of the group evaluate the indicators, compare their evaluations, and discuss deviant evaluations.

Holding Preliminary Tryouts

The next step in this process is a preliminary tryout of the coding protocol. There are several examples of this in the educational technology literature. In fact, this step accounts for most of the published QCA studies. Except for Fahy's (in press; 2002a; 2002b; 2001), Henri's (1991), and the Community of Inquiry's (2002) coding schemes, we know of no protocols that have been used in multiple studies.

Even rigorously developed coding protocols will be subject to unforeseeable shortcomings. During each successive use of Anderson et al.'s (2001) coding scheme for teaching presence and Rourke et al.'s (1999) scheme for social presence, substantial changes were made: (a) indicators that were not being used were abandoned, (b) indicators that were unreliable were reworded

or discarded, and (c) indicators that were conceptually misaligned were moved to more appropriate categories. Similar commentary has accrued with Henri's (1991) instrument although no changes have been formally adopted.

Developing Guidelines for Administration, Scoring, and Interpretation of the Coding Scheme

Based on the work that has been conducted in the previous steps, developers are in a position to amass guidelines for administration, scoring, and interpretation of the protocol. Thorough QCA reports, such as those presented by Jonassen and Kwon (2001), contain information concerning intrarater and interrater agreement, training procedures for coders, and samples of coded transcripts. The frequency with which journal editors are requesting this information is increasing, and if researchers are eager to have their protocols used and a set of results replicated, it is requisite information.

A useful element in interpreting the results of a QCA study would be a pool of normative data. Unfortunately, no such pools exist. Kamin, O'Sullivan, Younger, and Deterding (2001) used QCA to study the critical thinking of medical students during problem-based learning. Their discussion pointed to the problem of interpreting results in the absence of normative data:

> Additional work is needed to determine the magnitude of critical thinking ratios that should be obtained in a typical 3rd-year medical students' problem-based learning group. Critical thinking could possibly be higher in other situations or this could be as high as it gets. Other researchers studying classroom discourse found critical thinking ratios between .44 and .87, but they sampled undergraduate students, not medical students or problem-based learning groups. (p. 33)

At the culmination of their study, Kamin et al. (2001) were able to describe their construct for the group they studied, but without normative data, they could not offer a meaningful interpretation of the data in a larger context.

The final step in test construction that Crocker and Algina (1986) discussed is designing and conducting validity studies for the final form of a coding protocol. This step signals the conclusion of the first stage of establishing validity in QCA studies and the first section of this article. In this section we translated some of the essential processes of responsible test development into the less paradigmatic testing and measuring context of QCA. Addressing these steps during the development process will increase the durability, persuasiveness, and validity of QCA studies.

Having completed the above steps, the content analyst has an instrument that is plausibly valid. In the next section, we present several research designs for gathering empirical evidence of the validity of a QCA protocol.

GATHERING EMPIRICAL EVIDENCE FOR VALIDITY

A rigorous and systematic process of construction yields a set of indicators that are reasonable (Sheppard, 1993). However, the appropriateness of inferences made from the protocol remains hypothetical until it is demonstrated empirically. One example of this, discussed earlier, is the problem that vicarious interaction (Sutton, 2001) presents for the validity of participation protocols. Another example is Bereiter and Scardemalia's (1987) cautious attitude toward writing samples as evidence of cognition. A further example is found in our work on social presence (Rourke & Anderson, 2002). Originally, we developed a scheme for coding the purely social elements of computer conferences, based on the belief that social communication is an important antecedent to critical discourse. However, interviews indicated that this interpretation was not universal among the students who were participating in computer conferences. As one participant said:

> While I was not inhibited from commenting in general, I was reluctant to bring up points of dispute. The environment became much more social than useful in the exchange of ideas. I grew tired of the niceties of online protocol and wished that other participants would just get to the point. (Rourke & Anderson, 2002)

The problem this comment reveals is not with

the power of the coding scheme to provide a tally of salutations, compliments, and other indicators in the protocol. It is with the inference that what the raters have observed, categorized, and counted is communication that supports critical discourse. Further investigation is required to warrant such an interpretation.

Messick (1989) discussed several types of investigations that should be conducted to establish the validity of any test. Of these, three are particularly germane to the development of coding protocols whose purpose is to collect descriptive data and generate inferential information in research contexts. These are:

1. Correlational analyses
2. Examinations of group differences, and
3. Experimental or instructional interventions.

In this section we will examine all three. To illustrate the discussion, we will draw primarily on examples from our own work with which we are most familiar and which provides some of the few available examples.

Correlational Analyses

The first type of study Messick described is correlational analyses. In this type of study, the content analyst attempts to demonstrate that measurements of the construct made through QCA are consistent with measurements of the construct made through other methods. Rourke and Anderson (2002) performed this type of study on their social presence instrument. Working from a definition of social presence as the ability of learners to project themselves socially and affectively into a mediated community of inquiry, they constructed a coding scheme consisting of 15 representative behaviors (Rourke, Anderson, Garrison, & Archer, 1999). The purposes of their correlational study were to (a) explore the relevance of the indicators, (b) test that the frequency of the indicators was meaningful (an implicit assumption of QCA), and (c) assess social presence using an alternate method. To accomplish these goals they asked students to rate the frequency of the each of the indicators. They also asked students to assess the social presence of their conference using a more traditional method, a 5-item semantic differential scale anchored by the positive adjectives *warm, friendly, trusting, disinhibiting, close,* and *personal* (Gunawardena & Zittle, 1998; Short, Williams, & Christie, 1976). The authors found that correlations between the frequency of the 15 indicators and the students' ratings of social presence were weak (r = 0.4, approximately). Furthermore, significant correlations were observed only between a subset of the indicators and a subset of the social presence dimensions represented on the semantic differential scale.

These results point to difficulties in at least two areas of validity. First, content representativeness—the coding scheme was not measuring all of the dimensions of social presence. Second, content relevance—some of the indicators did not correlate with any of the dimensions of social presence. Ultimately this presents a challenge to the appropriateness of the inferences that one would want to make from the coding data; that is, that observing, categorizing, and counting the occurrence of the 15 indicators would allow one to infer how well students could project themselves socially and affectively into a mediated community of inquiry.

We have also conducted studies in which multiple methods of data collection were used to corroborate the results of a QCA (Rourke & Anderson, 2002). To explore the validity of Anderson et al.'s (2001) teaching presence coding scheme, they combined a QCA with interviews and a questionnaire. The questionnaire items were derived explicitly from the teaching presence indicators. For example, the indicator *drawing in participants* was transposed into the questionnaire item: This week's discussion leader was effective at drawing in participants. Similarly, the interviews were structured to further address the same indicator. Data from each of the methods were then triangulated. We were relieved to find that most of the indicators the raters were observing, categorizing, and counting were perceptible to the students, were performing their hypothesized function, and were differentially effective based on their frequency. We were also able to determine that many of the segments of the transcript that were rated high by the QCA were the same ones that were rated

high by the students who had participated in the conference.

Group Differences

Coding schemes for constructs such as critical thinking, group problem solving, or social communication are essentially embodiments of what their authors think the ideal process would look like. This belief gives rise to a type of study in which an ideal group engaging in the target process is compared to a group that is much less ideal. One hopes the instrument distinguishes between them appropriately. There are two forms of this type of study, cross-sectional and longitudinal. In a cross-sectional study, an effective protocol for coding social communication should distinguish between a cohort group in the final year of their program and a zero-history group. In a longitudinal study, the same protocol should be able to distinguish between the initial weeks and the final weeks of a computer conference. Demonstrating this ability is an important early step in validating a QCA protocol.

Anderson et al. (2001) were able to provide this type of evidence for their teaching presence instrument. Composed of the categories *direct instruction, instructional design,* and *facilitating discourse,* their instrument was able to distinguish between weeks of discussion led by students and weeks of discussion led by the instructor. As would be hypothesized by their model, instructor-led discussion contained more evidence of direct instruction and instructional design while student-led discussion contained more evidence of facilitating discourse (Rourke & Anderson, 2002).

This particular study, however, is not a paradigmatic group differences test, nor are we aware of other more definitive studies in the educational technology literature. Because it is important to understand this key form of test validation, we have selected an example from outside the domain. Willms (1978, as cited in Rogers, 1999) developed a test to assess knowledge of mental retardation. Once a set of items had been amassed, Willms administered the provisional instrument to four intact groups who he hypothesized would have varying but predictable degrees of success. His predictions were confirmed when education students studying mental retardation scored highest, community college students enrolled in upgrading classes scored lowest, and the remaining two groups—high school tutors working with mentally retarded students and education students studying statistics—scored in the middle.

Experimental and Instructional Intervention

In the group differences scenario, naturally occurring criterion groups are identified that are expected to differ with respect to the construct being measured (e.g., teachers vs. students on moderating ability). In experimental or instructional intervention studies, researchers deliberately manipulate the groups or their environment to induce an alteration in the construct under study. An attempt is made to modify behavior in theoretically predicted ways to determine whether the coding protocol is sensitive to these changes. For example, Jonassen and Kwon (2001) studied students' problem-solving skills using a content analysis protocol developed by Poole and Holmes (1995). If valid, such an instrument should be sensitive to instructional interventions such as training students in the problem-solving process or providing expert assistance while students are engaged in problem-solving tasks. The following year, Jonassen and Kwon (2002) conducted such a study. Undergraduate economics students were divided into small teams and asked to resolve economics problems online. To communicate with each other, some teams used a standard computer conferencing system while others used an electronic argumentation–scaffolding system. Using Poole and Holmes's coding protocol, Jonassen and Kwon were able to identify clearly where the scaffolding system had been used and where it had not.

Correlational analyses, group differences studies, and experimental or instructional intervention studies provide empirical evidence that a coding protocol describes what it purports to describe and that the inferences made from the

coding data are warranted. Without this evidence, inferences remain speculative. If protocol developers are convinced of the obviousness or appropriateness of their behavioral indicators, they will not be averse to an empirical evaluation of the coding protocol's utility. If the proper steps have been taken during the development stage, evaluations of the provisional instrument should not result in dramatic surprises or disappointments.

CONCLUSION

In a previous paper, we constructed a table in which the criteria of QCA constituted the columns and published studies the rows (Rourke et al., 2001). Studies that epitomized the use of a particular criterion were inserted in the appropriate cells. If we prepared a similar table for this paper, most of the cells would be empty. Examples of QCA research in which a coding protocal is developed methodically and validated systematically are rare. Consequently, the qualities that define QCA, distinguish it from qualitative forms of content analysis, and endow it with its appeal as a research technique fade away. Reeves (1995) chastised educational technologists for failing to establish the validity of their measurement procedures, and he regarded this continual failure as evidence of a "research malaise of epic proportions" (¶ 23). We have argued merely that attention to test validity will strengthen the claims of QCA studies and increase their information yield.

Our argument is relevant to the content analyst engaged in a purely descriptive study, but it is directed specifically at those who include inferential processes in the collection, analysis, and interpretation of their data. Some researchers continue to use QCA in the manner in which it was originally conceived. They systematically identify, categorize, and count the objective elements of communication and provide audiences with a summary of this data. The procedure is sound, the analysis leaves little room for counter interpretation, and the results of descriptive studies are valuable, especially when they concern relatively new educational phenomena such as the use of CMC in teaching and learning.

Others have tried to deepen our understanding of these phenomena by using procedures that expand the traditional conception of QCA. They too begin by identifying and categorizing directly observable behaviors. These behaviors, however, are not of interest in and of themselves. They are taken as signs, evidence, or indicators of an underlying construct. Drawing conclusions about underlying constructs based on frequency counts of the surface content of communication is a complicated analytical process, though it is rarely recognized as such.

Seeking a model to guide this complicated process, we have drawn attention to a fact that is often overlooked: QCA is a form of testing and measurement. This ontological positioning connects content analysts to an assortment of well-defined procedural tools that will help them make inferences and interpretations that are theoretically and empirically defensible. Among these tools are procedures for amassing a set of legitimate behavioral indicators and a set of empirical studies designed to test their usefulness.

Researchers who are contemplating a QCA study may find this discussion disheartening; that has not been our purpose. Most book-length treatments of QCA begin by informing readers that "the first step in observational research is developing a coding scheme" (Bakeman & Gottman, 1997, p. 15) thus propelling investigators headlong into the tangle of issues discussed in this paper. Gall, Borg, and Gall's (1996) suggestion, located in the context of a broader discussion of research design issues, is immeasurably more appropriate: "Consider employing a coding system that has been used in previous research" (p. 359). Jonassen and Kwon (2001, 2002) took this approach in their studies of problem solving in CMC. Rather than postponing their study while they undertook an elaborate process of instrument development and validation, the authors proceeded directly with their investigation after selecting an instrument that had been developed by Poole and Holmes (1995).

Unfortunately, few researchers appear to be interested in conducting their studies with existing instruments. Those who do accomplish several things: They contribute to the accumulating

validity of an existing procedure, are able to compare their results with a growing catalog of normative data, and leapfrog over the instrument construction process. In our research program, we dedicated two years and a considerable proportion of our research funds to the development of three QCA protocols (Community of Inquiry, 2002). Only then were we able to begin using the protocols to investigate the phenomena that had originally captured our attention. In our case, the trade-off was justified because instrument development was a central focus of our proposal, and we and other researchers continue to use the instruments in current studies. This is an exceptional case.

The purpose of this discussion has been to prompt some reflection about what is required before one can make inferences from frequency counts of communicative behavior. The discussion is incomplete. We have talked about correlational analyses, but have not mentioned factor analysis (or path analysis, structural equation modeling, or hierarchical linear modeling). We have touched on protocol analysis and chronometric analysis but not computational or mathematical modeling. Of Crocker and Algina's (1986) 10 steps of test construction, we have discussed only 6. And we have drawn minimally from Messick's (1989) 102-p. chapter. There is a rich body of literature that researchers can refer to when developing a protocol and making inferences from coding data. Some preliminary considerations are outlined in this paper but much work remains. □

Liam Rourke [lrourke@ualberta.ca] is a Ph.D. candidate in the Department of Educational Psychology at the University of Alberta, Edmonton AB.

Terry Anderson [terrya@athabascau.ca] is Professor and Canada Research Chair in Distance Education at Athabasca University in Alberta.

REFERENCES

Anderson, T., Rourke, L., Garrison, D.R., & Archer, W. (2001). Assessing teaching presence in a computer conferencing environment. *Journal of Asynchronous Learning Networks, 5* (2). Retrieved, March 6, 2002, from http://www.aln.org/alnweb/journal/jaln-vol5issue2v2.htm

Bakeman, R., & Gottman, J.M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bereiter, C., & Scardemalia, M. (1987). *The psychology of written composition.* Hillsdale N.J: Lawrence Erlbaum Associates.

Berelson, B. (1952). *Content analysis in communication research.* Glencoe, IL: Free Press.

Bullen, M. (1998). Participation and critical thinking in online university distance education. *Journal of Distance Education, 13*(2), 1–32.

Chou, C. (November, 2001). *A model of learner-centered computer-mediated interaction for collaborative distance learning.* Paper presented at the Annual Meeting of the Association for Educational and Communications Technology, Atlanta, GA.

Community of Inquiry. (2002). Critical thinking in a text-based environment: Computer conferencing in higher education. Retrieved March 6, 2002, from http://www.atl.ualberta.ca/cmc

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Toronto: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Cronbach, L. (1990). *Essentials of psychological testing* (5rd ed.). New York:Harper & Row.

Curtis, D., & Lawson, M. (2001). Exploring collaborative learning online. *Journal of Asynchronous Learning Networks, 5*(1). Retrieved, March 6, 2002, from *http://www.aln.org/alnweb/journal/Vol5_issue1/Curtis/curtis.htm*

Dalkey, N., & Helmer, O. (1963). An experimental application of the delphi method to the user of experts. *Management Science, 9*(3), 458–467.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Ericsson, K., & Simon, H.(1993) *Protocol analysis: Verbal reports as data.* Cambridge, Mass: MIT Press.

Fahy, P. (2001). Addressing some common problems in transcript analysis. *International Review of Research in Open and Distance Learning, 1*(2). Retrieved, March 20, 2002, from the World Wide Web at http://www.irrodl.org/content/v1.2/research.html

Fahy, P. (2002a). Epistolary and expository interactions patterns in a computer conference transcript. Retrieved, March 1, 2002, from http://cde.athabascau.ca/softeval/reports/mag2-jde.pdf

Fahy, P. (2002b). Evaluating critical thinking in a com-

puter conference transcript: A comparison of two models. Unpublished manuscript.

Fahy, P. (in press). Use of linguistic qualifiers and intensifiers in a computer conference. *American Journal of Distance Education.*

Flanagan, J. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327–359.

Gall, M., Borg, W., & Gall, J. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Garrison, D.R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15(1).*

Gunawardena, C.N., Lowe, C.A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research 17*(4), 397–431.

Gunawardena, C.N., & Zittle, F. (1998). Social presence as a predictor of satisfaction within a computer mediated conferencing environment. *The American Journal of Distance Education, 11*(3), 8–25.

Hara, N., Bonk, C., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science, 28*(2), 115–152.

Henri, F. (1991). Computer conferencing and content analysis. In *Collaborative learning through computer conferencing* (pp. 117–136). Berlin: Springer-Verlag.

Johnson, D., & Johnson, R. (1979). Conflict in the classroom: Controversy and learning. *Review of Educational Research 49,* 51–70.

Johnson, D., & Johnson, R. (1986). Computer-assisted cooperative learning. *Educational Technology 26*(1), 12–18.

Johnson, D., & Johnson, R. (1989). *Cooperation and competition: theory and research.* Edina, MN: Interaction.

Johnson, D., & Johnson, R. (1992a). *Creative controversy: Intellectual challenge in the classroom.* Edina, MN: Interaction.

Johnson, D., & Johnson, R. (1992b). Positive interdependence: Key to effective cooperation. In R. Hertz-Lazarowitz & N. Miller (Eds.), *Interaction in cooperative groups: The theoretical anatomy of group learning* (pp. 174–99). Cambridge, England: Cambridge University Press.

Johnson, D.W., & Johnson, F. (1994a). *Joining together. Group theory and group skills* (5th ed). Englewood Cliffs, NJ: Prentice Hall.

Johnson, D., & Johnson, R. (1994b). *Leading the cooperative school* (2nd ed.) Edina, MN: Interaction.

Johnson, D., & Johnson, R. (1996). Cooperation and the use of technology. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 1017–1044). New York: Simon and Schuster Macmillan. (1996).

Johnson, D.W., Johnson, R., & Maruyama, G. (1983). Interdependence and interpersonal attraction among heterogeneous and homogeneous individuals: A theoretical formation and a meta-analysis of the research. *Review of Educational Research, 53*(5), 5–54.

Johnson, D., Johnson, R., & Stanne, M. (2000). *Cooperative learning methods: A meta-analysis.* Retrieved, March 6, 2003, from http://www.co-operation.org/pages/cl-methods.html

Jonassen, D., & Kwon, H. (2001). Communication patterns in computer mediated versus face-to-face group problem solving. *Educational Technology Research and Development, 49*(1), 35–51.

Jonassen, D., & Kwon, H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development, 50*(3), 5–21.

Kamin, C., O'Sullivan, P., Younger, M., & Deterding, R. (2001). Measuring critical thinking in problem-based learning discourse. *Teaching and Learning in Medicine, 13*(1), 27–35.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* Scranton, PA:Chandler.

Krippendorff, K. (1980) *Content analysis: An introduction to its methodology.* Beverly Hills, CA: Sage.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mason, R. (1991). Analyzing computer conference interactions. *Computer in Adult Education and Training, 2*(3), 161–173.

McLean, S., & Morrison, D. (2000). Learners sociodemographic characteristics and participation in computer conferencing. *Journal of Distance Education, 15*(2). Retrieved, March 9, 2002, from http://cade.athabascau.ca/vol15.2/mclean.html

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed, pp. 13–103). New York: Macmillan

Norris, S., & Ennis, R. (1989). *Evaluating critical thinking.* CA: Critical Thinking Press and Software.

Paisley, W. (1969). Studying style as deviation from encoding norms. In G. Gerbner, O. Holsti, K. Krippendorf, W. Paisley, & P. Stone (Eds.), *The analysis of communication contents: Developments in scientific theories and computer techniques* (pp. 4458). New York: Wiley.

Parson, M. (1996). Look who's talking: A pilot study of the use of discussion lists by journalism educators and students. (ERIC Document Reproduction Service No. ED 400 562)

Poole, M., & Holmes, M. (1995). The longitudinal analysis of interaction. In B. Montgomery & S. Duck (Eds.), *Studying interpersonal interaction* (pp. 286–302). New York: Guilford. 1991.

Potter, J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*(3): 258–284.

Reeves, T. (1995). Questioning the questions of instructional technology research. [Online] Available http://www.hbg.psu.edu/bsed/intro/docs/dean/

Richardson, J., & Swan, K. (2003). Examining social

presence in online courses in relation to students' perceived learning and satisfaction. *Journal of Asynchronous Learning Networks 7*(1). Retrieved July 1, 2003 from the World Wide Web at http://www.aln.org/publications/jaln/v7n1/v7n1_richardson.asp

Riffe, D., Lacy, S., & Fico, F. (1998) *Analyzing media messages: Using quantitative content analysis in research.* Mahwah, NJ: Lawrence Erlbaum.

Rogers, W.T. (1999). *Error of measurement and validity.* Edmonton AB: Available from author.

Rourke, L., & Anderson, T. (2002). Social communication in computer conferencing. *Journal of Interactive Learning Research, 13*(3), 259–275.

Rourke, L., Anderson, T. Garrison, D.R., & Archer, W. (1999). Assessing social presence in asynchronous, text-based computer conferencing. *Journal of Distance Education, 14*(3), 51–70.

Rourke, L., Anderson, T., Garrison, D.R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education, 12(1),* 8–22.

Sharon Y., & Sharon S. (1992). *Group investigation: Expanding cooperative learning.* New York: Teacher's College Press.

Sheppard, L. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* London, U.K.: Wiley, 1976.

Slavin, R. (1991). *Student team learning: A practical guide to cooperative learning* (3rd ed.). Washington, DC: National Education Association.

Snow, R., Federico, P., & Montague, W. (Eds.). (1980). *Aptitude, learning, and instruction. (Vols. 1 & 2).* Hillsdale, NJ: Lawrence Erlbaum.

Sutton, L. (2001). The principle of vicarious interaction in computer mediated communication. *International Journal of Educational Telecommunications, 7*(3), 223–242.

Stevens, S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Weiss, R., & Morrison, G. (1998). Evaluation of a graduate seminar conducted by listserve. (ERIC Document Reproduction Service No. ED 423 868)