Adapting a Small Group Communication Quality Assessment to New Contexts

Lauren Jodi Van Scoy[1], Whittney Darnell[2], Tara Watterson[2], Vernon M Chinchilli[1], Emily J Wasserman[1], Daniel Wolpaw[1], Britta Thompson[1], Margaret Hopkins[1], Allison M Scott[2] & Rebecca Volpe[1]

[1] Penn State University, Pennsylvania, USA

[2] University of Kentucky, Lexington, Kentucky, USA

Correspondence: Lauren Jodi Van Scoy, Associate Professor of Medicine, Humanities and Public Health Sciences, Penn State University, Pennsylvania 17033, USA.

Received: May 11, 2018 Accepted: May 30, 2018 Online Published: June 4, 2018

doi:10.5430/irhe.v3n2p61 URL: https://doi.org/10.5430/irhe.v3n2p61

This work was supported by the Woodward Endowment for Medical Sciences Education from Penn State University. Dr. Van Scoy receives funding from the Parker B. Francis Career Development Award from the Francis Family Foundation.

Abstract

Objectives: Small group learning is a well-established medical education strategy for cultivating essential communication skills. Yet, how best to measure communication quality in these groups remains understudied. The objective of this study was to adapt a communication methodology (Communication Quality Analysis) to medical education small group setting.

Methods: This was an observational study of Preclinical Medical Humanities small group discussions. Groups were recruited by convenience sampling. Audio-recordings of 12 sessions (3 groups; n=22 students and 3 facilitators) were transcribed and analyzed using Communication Quality Analysis. Three coders assessed communication quality by assigning numeric scores based on how well participants accomplished communication goals within five domains: content, engagement, relationship, emotion and identity. Dialogue was coded every five minutes for each domain, resulting in 2,658 data points for analysis. Coder reliability was assessed using intra-class correlations. Variance components were assessed using a generalized linear model.

Results: CQA was successfully adapted to the small group education context. High inter-rater reliability was established for each of five communication quality domains (ICC range 0.875 to 0.98). Variability in scores the relationship and identity scores was based primarily on the duration of class (ie. 5 minutes into class versus 35 minutes into class). Variability in the content, emotion, and engagement scores was based primarily on the the participant (who was speaking). Considerable variability in domain scores was observed between participants, suggesting that the assessment is sensitive enough to detect nuanced differences between participants.

Conclusions: Our study shows that CQA is reliable when adapted to medical education small groups.

Keywords: communication theory, communication quality, medical humanities, small group learning, undergraduate medical education

1. Introduction

Small group learning is a well-established educational strategy that is used in a majority of U.S. medical schools (Kinkade, S., 2005) as well as in medical schools around the world (Christopher, D. F., Harte, K., & George, C. F., 2002; Schmidt, H. G., Vermeulen, L., & van der Molen, H. T., 2006). Small group learning improves understanding and retention of knowledge, while also facilitating development of important interpersonal skills, including listening, reflection, questioning, and coping with uncertainty. (Kinkade, S., 2005; Koh, G. C., Khoo, H. E., Wong, M. L., & Koh, D., 2008; Edmunds, S., & Brown, G., 2010; Ferris, H. A., 2015) Small group learning also helps to facilitate development of high quality communication skills that are key clinical competencies for clinicians. (Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A., 2013)

While the benefits of small group learning are well-documented, there is a paucity of research on how best to measure and assess the quality of the communication in these medical education groups. (Christopher, D. F., Harte, K., & George, C. F., 2002; Schmidt, H. G., Vermeulen, L., & van der Molen, H. T., 2006) Currently, there are no gold standard measures of small group communication quality. Traditional assessments of small group communication involve learner self-assessment, learner satisfaction, subjective facilitator assessments, and/or checklists of communication behaviors. (Schmidt, H. G., Vermeulen, L., & van der Molen, H. T., 2006; Koh, G. C., Khoo, H. E., Wong, M. L., & Koh, D., 2008; Edmunds, S., & Brown, G., 2010; de Jong, Z., van Nies, J. A., Peters, S. W., Vink, S., Dekker, F. W, & Scherpbier, A., 2010; Roter, D. L., Hall, J. A., Kern, D. E., Barker, L. R., Cole, K. A., & Roca, R. P., 1995) Shortcomings of these assessments, particularly those related to communication quality, are well-described. (Edmunds, S., & Brown, G., 2010; Ferris, H. A., 2015; Scott, A., 2014) Most importantly, these communication assessments lack grounding in communication theory or any conceptual framework. (Scott, A., 2014) Self-assessments, the most commonly used modality, are particularly problematic considering the robust data showing that clinicians have limited ability to accurately self-assess. (Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L., 2006) As such, we propose the use of a well-grounded communication framework, Multiple Goals Theory, (Caughlin, J. P., 2010; Scott, A. M., & Caughlin, J. P., 2014; Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) as a means to assess communication quality in small group learning in medical education.

Multiple Goals Theory defines high quality communication as occurring when conversants balance three conversational goals simultaneously: "task goals" (accomplishing a particular task), "relational goals" (affirming or validating relationships with others), and "identity goals" (managing one's self-presentation, ideas or agendas). (Caughlin, J. P., 2010) Low quality communication occurs when goals are not aligned or when one goal is pursued at the expense of the others. Grounded in this theory, Conversation Quality Analysis (CQA) operationalizes it into an objective, quantitative assessment of communication as it occurs. (Scott, A. M., & Caughlin, J. P., 2014; Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017; Scott, A. M., & Caughlin, J. P., 2012) Because Multiple Goals Theory is broad in its definition of communication quality, CQA has the unique potential to be applied in a wide variety of contexts. We have previously shown that CQA is a reliable and valid approach to the assessment of group communication about end-of-life issues, (Scott, A. M., & Caughlin, J. P., 2014; Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017; Scott, A. M., & Caughlin, J. P., 2012) but it has not yet been applied to other contexts, such as medical education.

The purpose of this study was to adapt the CQA methodology for use in the assessment of small group learning in medical education. To do so, we studied the communication quality of small groups during a pre-clerkship Medical Humanities course. We hypothesized that CQA can be reliably adapted to this context and would result in a reliable measure of communication quality in this setting. We also examined whether the assessment approach was capable of discriminating students' level of performance.

2. Methods

2.1 Recruitment and Setting

This study took place at Penn State University College of Medicine. The Institutional Review Board approved this study. Nine small groups (consisting of first year medical students) were offered participation in the study based on continuity of small group facilitator throughout the year. Groups were enrolled in the study only if informed consent was obtained from all students in the small group and the faculty facilitator(s). Of the 9 groups approached, 6 provided consent and we analyzed 3 groups' transcripts using Communication Quality Analysis (which resulted in 2,658 data points). The three groups were selected based on those who had consistent faculty facilitators throughout the sessions as intended in the course. We opted to record six groups' sessions in case groups forgot to start/stop audio recorders and to account for potential group dropout from the study, although this did not occur. We chose to study only three groups for this pilot study because three was sufficient to provide us with enough experience and data points to assess feasibility and reliability of adapting the method to this context.

2.2 Description of Medical Humanities Small Group Sessions

The Medical Humanities course consists of 13 weekly sessions that involve a 50-minute large group plenary immediately followed by 50-minute small group breakout sessions consisting of 7-8 students and one faculty facilitator. Weekly class topics were: Culture of Medicine; The Wellness-Illness Continuum; Empathy; Suffering; Addiction; Spirituality; Disability; Caregiver Issues; Death and Dying; and Joy in Medicine. Each session has a written set of 4-5 learning objectives that are distributed to students and facilitators. All small group sessions were audio recorded for the 3 enrolled groups. In order to get a broad sampling of topics, four different sessions from each

of the 3 groups were transcribed, resulting in a total of 12 transcripts covering 10 topics.

2.3 Communication Quality Analysis (CQA)

CQA is a previously published technique for assessing the quality of conversation for each member of a group. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) Briefly, CQA involves listening to audio of a conversation while reviewing transcripts and then assigning numeric scores of communication quality for five quality domains derived from the three goals: 1) task (engagement and content domain); 2) relational (relationship and emotion domain); and 3) identity (identity domain) Dialogue from each group member has a score assigned in five-minute increments for each of the five domains. Two authors (LJV, AMS) with experience using CQA trained two coders (TW, WD) using previously published methodology. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017)

2.3.1 Adaptation of the CQA Codebook to Medical Humanities Small Groups

To adapt the CQA codebook to the medical humanities small group context, three coders (LJV, WD, TW) first reviewed 2 transcripts and identified exemplars within the data that represented each of the five domains. Each coder revised the original end-of-life CQA coding dictionary to apply the definitions to the current data. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) These definitions were then brought to the group and a consensus definition was defined. Once the group reached consensus for the domain definitions, exemplars were selected for each domain and were rated on a 7 point Likert scale as illustrated in Figure 1.

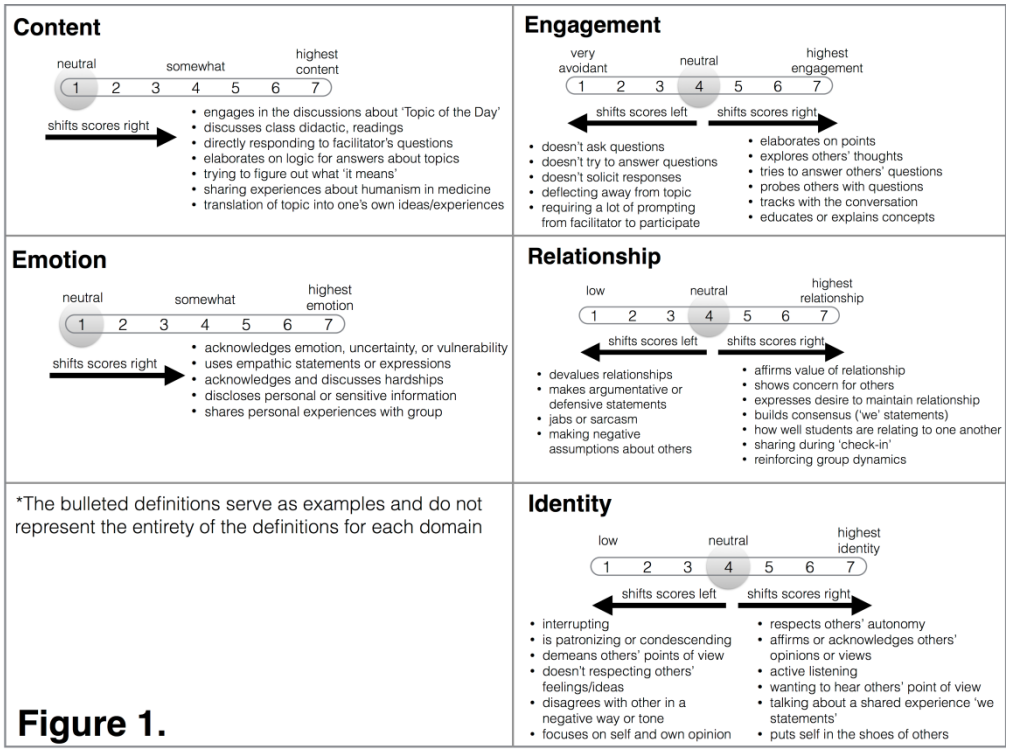

Figure 1. The CQA Coding Process

Coders begin scoring dialogue at the neutral score for each domain (represented by the grey dot). As participants engage in dialogue in accordance with the codebook (bulleted text), scores are increased or decreased. Scores are recorded every five minutes for each of the five domains.

The definitions and scored exemplars formed the initial coding dictionary. After each of the three coders used this dictionary to individually code 15% of data, intra-class correlations were calculated. Conflicts in coding were resolved by group adjudications, and the codebook was refined to reflect the consensus reached in this group

discussion. When intra-class correlations were >0.70 for each domain, the remainder of the dataset was coded independently by all three coders.

2.3.2 Assigning CQA Scores

Table 1 provides examples of low, neutral and high scoring exemplars from discussions for each of the five domains.

Table 1. Examples of low, neutral and high scoring conversation (excerpts)

Domain	#	Quote	Participant	Additional Domains*	Score	Justification
** Quotations selected to exemplify scores of 1,5,7 for each domain in the first column. Additional domain scores are shown to exemplify additional goals within each quote for illustrative purposes.*						
Content		[*F-3-4 describes helping a dog tied up in the rain during norming/check-in]* F-4-4: What kind of dog was it? F-3-4: A little Australian cattle dog. M-1-4: It was nice, I mean, it was very friendly. GF-4: I'm just trying to picture my dog tied up outside. She would not like— F-3-4: And during the storm yesterday, too. [lots of comments: Oh yeah, that sucks]	All	**Content**	**1**	Not discussing class topics
				Relationship	5	Sharing personal experience during check-in; group consensus building
	2	M-1-4: What are the rules legally regarding that *(referring to an obligation to provide medical care to a terrorist injured in his attack*]. Can you refuse to get involved or if you are called in—or what are the rules regarding that?	M-1-4	**Content**	5	Talking about legal aspect of providing care rather than empathy per se, but it is still on the class topic (empathy).
				Engagement	6	Asking probing questions
		GF-4: It would depend on kind of where you are and what you're practicing, so if you're in the emergency room…you can't refuse there…but if he's now stabilized and you're trying to send him somewhere for follow-up care, perhaps. It would vary.	GF-4	**Content**	**5**	Answering questions related to the topic
				Engagement	6	Elaborates and explains answers
	3	M-2-4: I think that's a really interesting topic (*referring to another student's comment*)…most of my time was spent in an emergency room and it can be very shocking when you see the first patient like die in front of you but that kind of desensitizes you when it happens. A lot of the stuff that you're talking about, like joking in OR and stuff … it's very interesting to me that some people, that's their way of dealing with stress is like they aren't really, they empathize	M-2-4	**Content**	**7**	Overtly discussing the class topic–empathy; internalizing previous experiences related to empathy
				Engagement	7	Elaborating on ideas and opinions in depth
				Identity	7	Affirming other person's contribution to conversation, seeking to understand others' points of view, providing alternative explanations for others' behavior (ie. facesaving)

		with the patient and I know it because I've worked with them so often, and you know the person, but their way of dealing with the stress of being in the ER… is kind of just joking around. And I understand that.		Emotion	5	Acknowledging emotion, but not explicitly elaborating on his own emotions about the topic
				Relationship	5	Sharing personal experience with the group but not explicitly talking about relationships within the group
Engagement	1	*[Check-in question: who would you want to play you in a movie?]* GF-5: You're all looking at me. I don't know what to say. I think I'll take a pass on that.	GF-5	**Engagement**	**1**	Doesn't try to answer question or participate
	2	F-4-4: Yeah, I don't really have anything else to add. Doing fine. On top of stuff, I guess, hopefully.	F-4-4	**Engagement**	**2**	Answers the question but doesn't elaborate
	3	F-1-4: Do you, facilitator, know or remember anyone from medical school who now has totally changed [*i.e. become less empathetic and more jaded*] as an attending or anything?	F-1-4	**Engagement**	**7**	Asking a direct question to the facilitator in order to further discussion, probing or exploring others points of view.
				Content	7	Asking a direct question about the topic of the day– empathy
Emotion	1	F-2-4: I think it comes back to the idea of being professional and just knowing that as a physician, our first and foremost priority is for the patient, like medically, like to treat them, if they're dying is to save their life.. even though they may be a criminal that did something we don't agree with…[we have to] save them [although] they may still go out and do like things that we don't agree with or things that are illegal but I don't think as physicians, we're the judge of … I don't think that's specifically one of our roles.	F-2-4	**Emotion**	**2**	Discussing a sensitive topic while advocating for separating one's self from one's emotions
	2	GF-4: I don't know that anybody has totally changed. I think though that there are probably different times in your life perhaps that you're not at your best self so I know during internship, I was probably, there were times I know I was not at my best self because you are exhausted and there are a lot of pressures so I think that that does play into it.	GF-4	**Emotion**	**5**	Acknowledging vulnerability, some self-disclosure
	3	M-1-4: I've actually been on the	M-1-4	**Emotion**	**7**	Self-disclosure that

		opposite end of that as the patient where something was seriously wrong and the doctor kind of assumed that I was just seeking attention. I was in really, really, really bad shape. And it's definitely scary being like that and trying to get something across and the doctor's not really listening.				includes describing feelings, fears, and information that makes the person potentially vulnerable to negative evaluation
	4	M-1-4: …if I know someone is a child molester or something like I'd have a very, very tough for me to set aside my personal feelings … I don't know how I'm going to be able to balance that out. I'd love to say I can just do my job, but there's definitely something more than just doing a job that makes you a physician, and I wouldn't be able to have empathy for them.	M-1-4	**Emotion**	**7**	Acknowledging one's limitations, displaying vulnerability, discussing hardship
				Content	7	Discussing in depth and personalizing the topic of the day
Relationships	1	GF-5: My experience has been doctors, it's unbelievable how doctors will turn a blind eye. I can't tell you how many experiences, I, I for years have been involved with impaired physicians … it's incredible how bad we are, how bad you guys are, how bad I am in terms of we'll see a colleague in the hallway, we smell alcohol on his breath or he's making a series of mistakes and no one says anything. We just all kind of walk on by and just [unclear] until a patient dies, the hospital gets sued, and you know, then. I don't know why we do that. But we tend not to confront one another.	GF-5	Relationships	1	Making negative assumptions about others in the room, critical or demeaning of others
	2	GF-4: Yeah, so we're over the jet lag and had a day to try to catch up and [unclear] a little bit of catch up. Familiar concept. Play catch up even as attendings. M-1-4: So it doesn't get any better? [group laughter] GF-4: Different kind of catch up, you find out.	All	**Relationships**	**5**	Building group consensus
	3	F-1-4: Sure. It doesn't matter. I think it went well. I like, I think the stories from people's personal experiences help a lot and I definitely think like people can express how they are really thinking here, so I think it went well.	F-1-4	**Relationships**	**6**	Reinforcing group dynamic

	4	F-2-4: This weekend was good. I'm doing well. I met my patient for the patient project with my partner on Sunday. That went really well. And yeah, everything's going ok, except I dropped my phone this morning— [loud chorus of, "oh, no"] F-2-4: --and now my screen's cracked. Yeah, that was the only thing, I was like, great— M-3-4: That's the worst. M-2-4: I did that last week. F-2-4: I know, I'm like debating if I should fix it or not. If it's even worth it. M-1-4: [unclear] 100 bucks. F-2-4: I looked on Apple, and it's like 100 bucks, to send it in. But—everything else is good though. M-1-4: You have to hope you don't slice your finger one time though. F-2-4: Yeah, thankfully I have a screen protector on…	**All**	**Relationships**	**6**	Building group consensus
	5	M-1-4: I actually really kind of. Couple things. I'm actually going have to start, probably going to be doing some thinking over the next couple of days late at night just about these kinds of things as I'm going to bed, kind of wrestling with some of these things that are kind of challenging my own inner beliefs [unclear]. Back of my mind kind of thinking for a while trying to resolve it. I just want to kind of thank people for making me grow as a person. I like how kind of open it is. Honestly, I have no problem with you interrupting me, if anything I feel that it makes it feel less formal.	M-1-4	**Relationships**	7	Affirming the value of the relationships within the group, explicitly appreciating others' contributions
				Emotion	7	Discussing personal feelings
Identity	1	M-1-4: I definitely think stress always brings out a person's true character. And-- M-3-4: [interrupting/interjecting] Right, I agree and like certain, maybe even like the retard comment, maybe that was off the hand, maybe that's not really speaking to that person's character because I do think that does happen sometimes, I mean, we all let something slip out there like we're like, 'oh man, like I just—'	All	**Identity**	**2**	Talking over one another, frequent interruptions; not a '1' because some acknowledgement of common ground

	M-1-4: [interrupting/interjecting] But-- M-3-4: --that shouldn't have happened. M-1-4: [interrupting/interjecting] But they're actually saying what they actually believe. M-3-4: [continuing] Right, I think it's different when it is-- M-1-4: [interrupting/interjecting] to say that, when something— M-3-4: --who the person is.				
2	F-3-4: But I like what you were saying about meeting them at a basic human level. Cause for me, my personal belief is like everyone is good, like deep down inside everyone is good and like maybe their life circumstances have been such that they act like a jerk, or that they want to blow up a plane with 250 people, and that's not something I can relate to or empathize with but I can empathize with like man, like something must have really happened to them that made them act like this and that, I think, would allow me to give them just that basic level of care.	F-3-4	**Identity**	**5**	Briefly validating classmate's point of view
			Engagement	7	Explaining her opinions, statements
3	F-1-5: I came up with that same question…after reading that same article. Since our role would be as a physician, ok, this woman has some form of OCD…and we might perceive that as, oh, there's something wrong, or we might perceive that as ill, but if they're happy with it, and they are content with it, and they're content with…biting her nails or chewing on her hair or something, or pulling her hair maybe, if she's happy with that, you know, where is the line of when do we treat it and when do we not.	F-4-4	**Identity**	*7*	Acknowledging others' points of view that may seem different than your own

2.3.3 Scoring Task Goals (Content and Engagement Domains)

The content domain was scored by rating the degree to which participants discussed relevant topics (ie. the class topic of the day, see Table 1). For the content domain, the 'neutral' score was set at '1' and raised incrementally depending on the degree to which participants discuss relevant topics and the degree to which they explore those topics. The neutral position was set at 1 because previous studies found that coders were more reliable rating the 'presence' of content rather than the 'absence of content'. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) The highest scores were assigned to participants who engaged in a broad and deep exploration of the topic of the day either by asking provocative questions or discussing the topic on a

sophisticated level (e.g. examining the topic from various perspectives).

It is customary at the beginning of humanities small group classes that students have a brief "check-in" during which they take turns sharing with the group 'how they're doing' or perhaps an event that has occurred recently. This dialogue was rated as content 'neutral' (unless it related to the 'topic of the day') even though it is considered an important part of the class expectations. However, these conversations were rated more highly in other quality domains (ie. emotion, relationship) based on the degree to which the participants disclosed sensitive information or shared with the group.

The engagement domain was scored by rating the degree to which individuals participate in the conversation (regardless of whether or not the dialogue is 'on topic'). Unlike the content domain, the neutral score for engagement was set at '4' because of the normative expectation for participants in these groups to, at a minimum, engage in courteous conversation. As participants elaborated on their ideas (whether related to the topic or not), probed others' ideas with questions, contributed to and/or tracked with the conversation, and explained their ideas, scores were increased. When individuals did not participate in the conversation, deflected conversation away from themselves or required significant prompting from the facilitator to participate, the score was reduced from '4' down to a '1'.

2.3.4 Scoring Relational Goals (Emotion and Relationship Domains)

The emotion domain was scored by rating the degree to which participants discussed or disclosed sensitive, emotionally vulnerable topics. The neutral score for the emotion domain was set at '1' because, like the content domain, previous studies found that coders were more reliable in their ratings of the 'presence' of emotion than they were in identifying the 'absence' of emotion. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) When participants acknowledged emotion or discussed personal experiences, scores were increased from '1' towards '7'. The highest scores (6 or 7) were assigned when participants explicitly disclosed their *own* emotions, shared sensitive or personal information, or revealed vulnerability.

The relationship domain was scored by rating the degree to which participants affirmed or validated the value of relationships either within the small group (which resulted in the highest relationship scores) or relationships with persons outside the small group setting (e.g. physician-patient relationships). The neutral score for relationships was set at '4' because it was possible for participants to affirm (higher scores) or undermine (lower scores) relationships with their comments. Comments that sought to build group comradery, showed concern for others' experiences, reinforced the group dynamic, or shared personal experiences resulted in increased relationship scores. When participants de-valued relationships, used sarcasm or jabs in clearly non-playful ways, demeaned others or made negative statements about others, the score was decreased towards '1'. If participants refused to discuss a topic or share with the group, relationship scores were also decreased since this demonstrated a lack of good faith effort to trust others in the group.

2.3.5 Scoring Identity Goals (Identity Domain)

Attention to identity goals was scored on one domain (the identity domain). This domain was scored by rating the degree to which participants showed respect for others' views and perspectives. Like the relationship domain, the neutral identity score was set at '4' because identities could be affirmed or undermined. When participants made affirming statements about others' ideas or views, respected the autonomy of others, or acknowledged others' viewpoints (even if they disagreed), identity scores were increased from '4' towards '7'. Similarly, when participants offered alternative, flattering explanations for another's behaviors or comments ('facesaving'), scores were increased. When participants criticized or demeaned others' ideas, disagreed with others in a contentious way, frequently interrupted, acted in a patronizing or condescending fashion, or turned focus towards themselves (ie. being egocentric), identity scores were lowered.

2.3.6 Calculation of the Domain and Multiple Goals Scores

Coders listened to the audio recordings of the class sessions while following along with the written transcripts of the group discussion. At every 5-minute interval of audio, five scores were assigned for each participant for each domain. Thus, for each 50-minute class, 10 scores were assigned per domain. The scores for each domain were averaged together to reach an overall domain score for each participant. This resulted in each participant receiving 5 overall scores for each class session.

Then, we calculated an overall Multiple Goals Score for each student. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) Briefly, this 'breadth' score condenses the five domain scores into three goals scores that represent attention to 'task' (the average of content and engagement domains), 'relational' (the average of relationships and emotion domains), and 'identity' goals (identity domain). Based on the sample

mean, a normative score was assigned for each time segment for each of these three goals: '0' if the conversant scores at or below the sample mean and '1' if the conversant scores above the sample mean. A Multiple Goals Score was then calculated for each time segment by adding the three normative goals, and then averaging those scores across time segments to produce the final Multiple Goals Score (range 0-3).

2.4 Data Analysis

2.4.1 Calculation of the Intra-class Correlation Coefficient and Variance Components

The intra-class correlation coefficient (which measures the agreement between coders) acted as the primary result for which evaluation of successful application of CQA to the medical humanities setting was assessed. The ICC was calculated by taking the difference between the total variability and the variability due to the coder effect and dividing this by the total variability, for each domain. To ensure that silence was not a factor in our analysis, the ICC was calculated in two ways: 1) including all silent time segments in the dataset (when individuals were silent, their scores were set to the neutral value for each domain); and 2) excluding all silent time segments (when individuals were silent, the scores were set to 'missing' values for each domain).

To evaluate potential sources of variability that may have impacted each communication domain, we applied a generalized linear mixed-effects model. A total of five variance components were considered: a group effect, subjects nested within the groups, class topics nested within the groups, time nested within class topics and within groups, and a coder effect.

To account for varied participant-levels of communication that define the group quality standards, a nested effect for subjects within the groups was taken into consideration. Additionally, each group of students participated in small-group discussions on varying class topics. Therefore, the topic of discussion for each class also was dependent on the groups to which the subjects belonged. Effectively, the length of time that the class discussion required, and hence the duration of the class, was measured via a surrogate value, the number of 5-minute time segments for which each student is scored. The number of data points for each student is correspondingly dependent upon the number of five-minute time segments (i.e. the length of the class). The number of time segments is dependent on: a) the class topic (since duration of the class varied); and b) the group (since the class topic differs between the groups). This hierarchical nature of the number of data points provides the rationale for considering such nested effects. For this analysis, time was treated as an ordinal variable, with lower values representing the start of the class and higher values representing the end of the class. Finally, each 5-minute time segment for each student (in every group and for every class topic) was scored independently by three coders.

To describe these different sources of variability that may affect overall domain scoring, SAS 9.4 was used to estimate the specified variance components by fitting a generalized linear model via PROC GLIMMIX, using random intercepts for the five effects mentioned previously above for each of the five outcome domains. A cumulative probit link function was embedded within each model, in which a multinomial distribution was assumed for the data from each domain. The variance components were then calculated as percentages of total variability for each domain response. These variance components provide information on the amount of variability that can be attributed to the outcome due to each component that was considered in the random-intercept-only models.

3. Results

Three small groups (n=25 individuals) participated in the study, of which there were 3 faculty facilitators (2 male; 1 female) and 22 first year medical students (45% male, 55% female). The mean, median, range, and reliability statistics of CQA scores by domain are shown in Table 2. For all domains, the intra-class correlations were >0.93 when including individuals' silent segments in the analysis and >0.87 when removing their silent segments from the analysis. The mean scores and confidence intervals for Group 5 are shown in Figure 2. Group 5 was selected as the exemplar group because the scores in this group were the most variable.

Table 2. Description of CQA domain scores

Domain	Neutral Score	Mean (95% CL)	Std. Dev.	Median (IQR)	Min-Max	Intraclass correlation (including silence)	Intraclass correlation (excluding silence)
Content	1	2.49 (2.42, 2.56)	1.87	1	1-7	0.965	0.899
Emotion	1	1.94 (1.89, 2.00)	1.43	1	1-7	0.926	0.868
Engagement	4	4.91 (4.88, 4.95)	0.90	5	2-7	0.979	0.875
Relationship	4	4.20 (4.18, 4.22)	0.53	4	1-7	0.998	0.998
Identity	4	4.18 (4.16, 4.20)	0.58	4	1-7	0.962	0.950
MGS	-	0.93 (0.89, 0.97)	1.11	0	0-3	-	-

Scores are averages of all time increments, classes, subjects, and coders; ICC reflects Coder Agreement

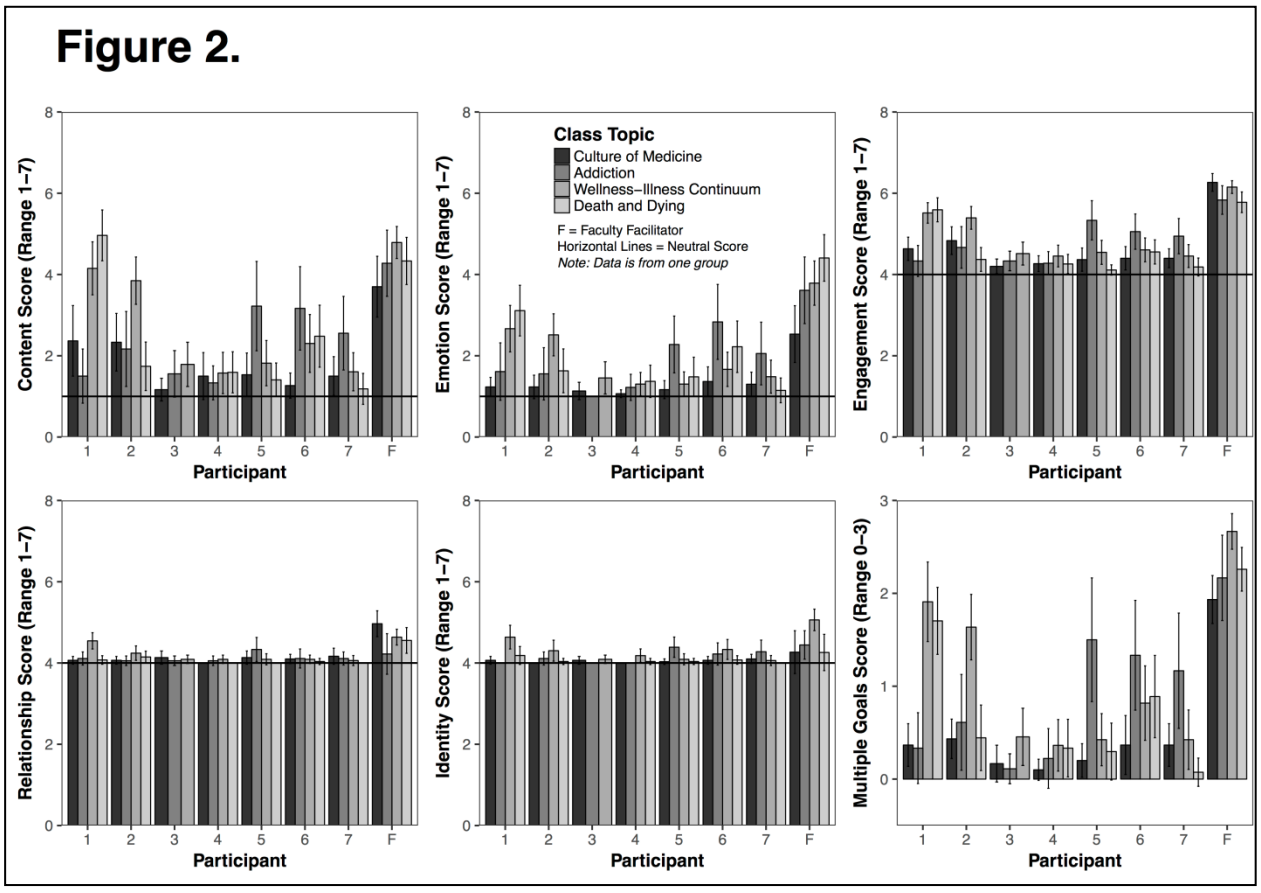

Figure 2. CQA scores for one small group

Bars represent the average scores that were coded in five minute intervals for each domain (1= low quality; 7=high quality). Horizontal lines represent a neutral score (1= neutral for content and emotion domains; 4= neutral for engagement, relationship and identity domains). The Multiple Goals Score represents the overall communication quality across all domains, with a score of 3 representing the highest communication quality.

3.1 Variability in CQA Scores

Figure 3 shows the considered sources of variability and their respective percentage of the total variability for each of the five domains. We found that content, emotion, and engagement domain score variability was primarily driven by the individual subjects (nested within the three communication groups); while relational and identity domain score variability was accounted for by duration of class (i.e. five minutes into class versus 35 minutes into class). The total variability of the Content domain appears to be larger than that for the Emotional, Relational, and Identity domains, all of which had observed minimum scores of 1 and observed maximum scores of 7. It is difficult to qualitatively compare the total variability of the Engagement domain with the others, however, due to the fact that the observed minimum score for the Engagement domain was a 2.

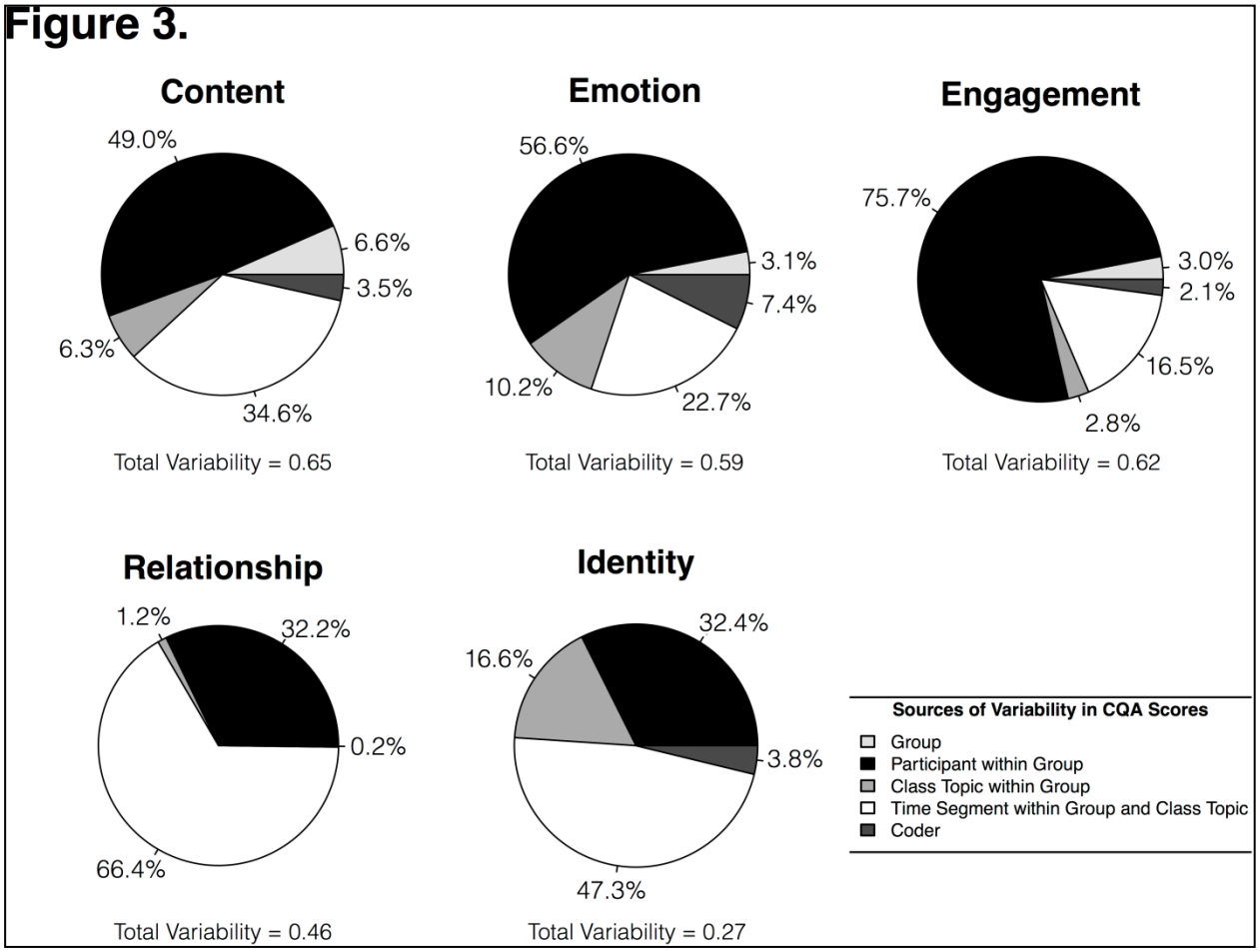

Figure 3. Sources of variability in CQA scores

Percentages represent the degree to which the source accounts for the variability within the scores (by domain).

4. Discussion

This study was the first to apply Communication Quality Analysis (CQA) to a medical education small group setting. Small group learning has become a prevalent and important element of medical education that is thought to promote development of important communication skills. However, there are, to our knowledge, no theory-based, rigorous evaluation methods available to educators wishing to assess the communication quality that results in these small groups. As such, this study was intended to adapt CQA for use in medical humanities small groups, and to determine whether its use resulted in reliable and valid results for the assessment of small group communication quality. Our results demonstrate that the use of CQA can be successfully adapted to the medical education setting (covering a variety of class topics) while also maintaining high inter-coder rater reliability as assessed by the intra-class correlation coefficients. We also observed that the scoring system displays considerable variability between

participants, regardless of the class topic, which is a key characteristic of a discriminating assessment tool. In doing so, this study has taken the first step towards validating a useful and informative methodology that could be applied to medical education research, and eventually used as a small group discussion evaluation tool. This discussion will focus first on the methodological and statistical findings of this study, and will then consider how this measure may be useful for medical educators in general.

4.1 Methodological and Statistical Considerations

4.1.1 Adaptability and Reliability of CQA Scores

Our finding that CQA was easily adapted to a context beyond end-of-life communication (the topic on which the method was originally developed and validated) (Scott, A. M., & Caughlin, J. P., 2014; Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017; Scott, A. M., & Caughlin, J. P., 2012) while also maintaining reliability is important because it provides evidence that CQA could be used more broadly as a communication quality assessment tool across fields and content areas. We observed very strong inter-rater reliability with high intra-class correlations among the three coders (>0.93). To ensure that this high level of reliability was not associated with the ease of coding silence (i.e., segments that are quite easy for coders to agree upon), we analyzed the data both including and then excluding segments where students made no verbal contributions. When excluding silence, we found no substantial difference in intra-class correlations, suggesting that coders were reliably coding active discussion and that the high reliability was not attributed to coders agreeing that students were silent. We also found CQA scores were reliable even when coding diverse small group discussion topics within this dataset (e.g., the culture of medicine, spirituality, disability, etc).

We also learned that when adapting the CQA codebook to other contexts, the normative expectations of the group dynamic and expected goal attention should be considered when modifying the CQA codebook. Specifically, when coding the relationship domain, we found that, unlike previous studies (Scott, A. M., 2011) scores in the relationship domain were predominantly neutral (mean 4.19). This neutral score makes sense when considering that the codebook defined the highest scores in the relationship domain as those that affirm or validate the value of relationships either *within* the small group (which results in the highest relationship scores) with only smaller increases in score for discussions related to relationships with persons outside the small group setting (e.g. physician-patient relationships). Specifically, in the medical education small group setting, comments about the patient-provider relationship are more normative than comments about student- student relationships. Thus, for future iterations of the CQA codebook, we recommend considering the normative expectations of the group conversation when setting the highest and lowest score limits for each of the domains.

4.1.2 Variability in CQA Scores

We observed that the scoring system of CQA demonstrated considerable variability across individual students and topics. This indicates that CQA is sensitive enough to detect nuanced differences between the participants being studied. We assessed variability qualitatively by examining magnitude of each participant's scores and the respective confidence intervals (Figure 5) for each student and also for each class topic. Figure 5 shows that there was meaningful variability (non-overlapping confidence intervals) between participants, but also that some individual participants had variable CQA scores depending on the class topic of the day (within-subject variability appears to be dependent on the class topic). Furthermore, when we examined the source of variability of the dataset as a whole, we found that for the content, engagement and emotion domains, the individual student was responsible for 49-76% of the variability in scores and the class topic for only 3-10%. This finding suggests that the topic of the day contributes less to the quality of student communication with respect to all domains than does individual characteristics. However generally speaking, the topic of the day had a larger influence on some domains than others, suggesting that perhaps the topic is less salient to communication quality than the student composition of the small groups.

Less variability was observed in the identity and relationship domains than in the other three domains, which is consistent with findings from some previous studies utilizing CQA in groups, (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) yet not others. (Scott, A. M., 2011) Similarly, while some variability was present in the engagement domain, this was less pronounced in our dataset compared to prior studies. (Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H., 2017) Such differences imply that different goals become more salient in different contexts and groups. For example, education contexts, where student grades are in large part determined by class participation (ie. engagement), our results suggest that students are more likely to make a good faith effort at contributing to the dialogue than perhaps some family groups discussing end-of-life issues where there is not a normative expectation that all individuals contribute equally to the dialogue. Similarly, in education context, relationships center primarily around professionalism and

mutual respect, rather than explicit affirmations of relationships (i.e. expressions of love). Thus, the variability (and definition) of each domain should be considered when adapting CQA to various contexts.

We also found that the primary source of variability for the content, emotion, and engagement domains was most dependent on the individual participant (versus topic, time segment, etc), whereas the primary source of variability for the relational and identity domain, was the time segment (early versus late in the class). This suggests that the content, emotion, and engagement domains are a function of individual factors, whereas relationship and identity domains are more attributable to group norms and operate at a group level. This offers further support that CQA adapts to the setting and context in which it is applied.

Several methodologic limitations of this study warrant consideration. For this first study using CQA in small group sessions, only three small groups were included in this analysis (albeit the analysis included 2,658 data points). Further study of additional small groups is warranted to establish generalizability and reproducibility of using CQA in this context. Second, in order to learn whether CQA could be adapted broadly across class topics (rather than to compare group scores on a single topic), we opted to analyze different class topics for each of the three groups which limited our ability to compare scores across class topics. Third, the method does not account for important non-verbal communication behaviors (such as head nodding, active listening, or silent engagement). Finally, this study did not assess various aspects of construct validity using CQA. (Royal, K. D., 2017)

4.2 Implications for Medical Educators

Although small group learning is a major element of modern medical education, very little attention has been paid to on how best to evaluate the communication quality within groups in an objective and quantifiable manner. CQA offers a promising approach to doing so by enabling study of small group communication quality in an more objective, rigorous, and theory-based way. CQA provides an important measure that could be used to evaluate the impact of novel curricular activities, varied facilitation techniques, and/or various types of groups (e.g. homogenous vs. interdisciplinary groups) on communication quality and other educational outcomes. In its current state, it is likely too labor intensive for general educators to use CQA as a routine individual student assessment, but further development and refinement may enable CQA to be applied more broadly. Studies are underway to simplify the transcript-based CQA methodology to one that can be utilized in real-time without the reliance upon transcripts, making it a more generalizable method for medical educators. Ongoing studies are also assessing discriminant validity by comparing CQA scores with traditional communication quality assessments.

5. Conclusion

We found that CQA is easily adapted to a medical education small group learning context, and that the resultant scores have high inter-rater reliability. Further, we observed meaningful variability in CQA scores that would allow the measure to be a useful tool for assessment. CQA is a promising, well-evidenced communication evaluation methodology that, with further refinement, has the potential to be used in wide variety of health settings.

Declaration of Interest and Confidentiality Statement: All authors declare that they have no conflict of interest. The authors confirm all patient/personal identifiers have been removed or disguised so the patient/person(s) described are not identifiable and cannot be identified through the details of the quotations.

References

Caughlin, J. P. (2010, September). A multiple goals theory of personal relationships: Conceptual integration and program overview. *Journal of social and personal relationships, 27*, 824. https://doi.org/10.1177/0265407510373262

Christopher, D. F., Harte, K., & George, C. F. (2002, March). The implementation of Tomorrow's Doctors. *Med Educ, 36*, 282-288. https://doi.org/10.1046/j.1365-2923.2002.01152.x

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006, September). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA, 296*, 1094-102. https://doi.org/10.1001/jama.296.9.1094

de Jong, Z., van Nies, J. A., Peters, S. W., Vink, S., Dekker, F. W, & Scherpbier, A. (2010). Interactive seminars or small group tutorials in preclinical medical education: results of a randomized controlled trial. *BMC medical education, 10*, 79. https://doi.org/10.1186/1472-6920-10-79

Edmunds, S., & Brown, G. (2010). Effective small group learning: AMEE Guide No. 48. *Medical teacher, 32*, 715-726. https://doi.org/10.3109/0142159X.2010.505454

Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A. (2013, August). Toward a common taxonomy of competency domains for the health professions and competencies for physicians. *Acad Med, 88*, 1088-94. https://doi.org/10.1097/ACM.0b013e31829a3b2b

Ferris, H. A. (2015). The Use of Small Group Tutorials as an Educational Strategy in Medical Education. *International Journal of Higher Education, 4*, 225. https://doi.org/10.5430/ijhe.v4n2p225

Kinkade, S. (2005, March). A snapshot of the status of problem-based learning in U. S. medical schools, 2003-04. *Acad Med, 80*, 300-301. https://doi.org/10.1097/00001888-200503000-00021

Koh, G. C., Khoo, H. E., Wong, M. L., & Koh, D. (2008, January). The effects of problem-based learning during medical school on physician competency: a systematic review. *CMAJ, 178*, 34-41. https://doi.org/10.1503/cmaj.070565

Roter, D. L., Hall, J. A., Kern, D. E., Barker, L. R., Cole, K. A., & Roca, R. P. (1995, September). Improving physicians' interviewing skills and reducing patients' emotional distress. A randomized clinical trial. *Arch Intern Med, 155*,. 1877-84. https://doi.org/10.1001/archinte.1995.00430170071009

Royal, K. D. (2017). Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract, 8*, 567-570. https://doi.org/10.2147/AMEP.S139492

Schmidt, H. G., Vermeulen, L., & van der Molen, H. T. (2006, June). Longterm effects of problem-based learning: a comparison of competencies acquired by graduates of a problem-based and a conventional medical school. *Med Educ, 40*, 562-567. https://doi.org/10.1111/j.1365-2929.2006.02483.x

Scott, A. (2014). Communication about end-of-life health decisions. *Communication Yearbook, 38*, 242-277. https://doi.org/10.1080/23808985.2014.11679164

Scott, A. M. (2011). Family conversations about end-of-life health decisions. AAI3452207.

Scott, A. M., & Caughlin, J. P. (2012, November). Managing multiple goals in family discourse about end-of-life health decisions. *Research on Aging, 34*, 670-691. https://doi.org/10.1177/0164027512446942

Scott, A. M., & Caughlin, J. P. (2014). Enacted goal attention in family conversations about end-of-life health decisions. *Communication Monographs, 81*, 261-284. https://doi.org/10.1080/03637751.2014.925568

Van Scoy, L. J., Scott, A. M., Reading, J. M., Chuang, C. H., Chinchilli, V. M., … Levi, B. H. (2017, May). From Theory to Practice: Measuring end-of-life communication quality using multiple goals theory. *Patient Educ Couns, 100*, 909-918. https://doi.org/10.1016/j.pec.2016.12.010