JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Spring 2017, Vol. 54, No. 1, pp. 54–69

Modeling Collaborative Interaction Patterns in a Simulation-Based Task

Jessica J. Andrews, Deirdre Kerr, Robert J. Mislevy, Alina von Davier, Jiangang Hao, and Lei Liu
Educational Testing Service

Simulations and games offer interactive tasks that can elicit rich data, providing evidence of complex skills that are difficult to measure with more conventional items and tests. However, one notable challenge in using such technologies is making sense of the data generated in order to make claims about individuals or groups. This article presents a novel methodological approach that uses the process data and performance outcomes from a simulation-based collaborative science assessment to explore the propensities of dyads to interact in accordance with certain interaction patterns. Further exploratory analyses examine how the approach can be used to answer important questions in collaboration research regarding gender and cultural differences in collaborative behavior and how interaction patterns relate to performance outcomes.

Educational games and simulations offer interactive tasks that can elicit rich data, providing evidence of complex 21st century skills that are difficult to measure with more conventional items and tests. However, one notable challenge in using such technologies is the difficulty associated with making sense of the data generated in order to make claims about individuals or groups (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012). In this article, we present a novel methodological approach that uses the process data and performance outcomes from a simulation-based collaborative science assessment platform developed by Educational Testing Service, the Tetralogue, to explore the varying ways in which individuals interact in the assessment task. In particular, we present a novel application of the Andersen/Rasch multivariate IRT model to examine the propensities of dyads to interact in accordance with certain interaction patterns. Further exploratory analyses examine how the approach can be used to answer important questions in collaboration research regarding potential gender and cultural differences in collaborative behavior and how interaction patterns relate to performance outcomes.

Background

Collaboration and Assessment

Collaborative activities have been utilized in a variety of contexts such as the school, workplace, and military as a potentially effective way to improve learning and performance outcomes. When individuals engage in collaborative activities they are provided with opportunities to engage in many beneficial processes such as the elaboration of their knowledge, the co-construction of ideas, and the resolution of conflicting beliefs (Andrews & Rapp, 2015).

The assessment of skills associated with collaboration, particularly in problem-solving situations, has recently received increasing attention, especially in computerized environments and large-scale assessment contexts (Griffin, Care, & McGaw, 2012; Organisation for Economic Co-operation and Development, 2013). Competency in collaborative problem-solving (CPS) focuses on two dimensions: a cognitive dimension associated with problem-solving processes and an interpersonal dimension associated with collaboration processes. A great deal of work has investigated these two dimensions as individual constructs (e.g., Brannick, Prince, Prince, & Salas, 1995; Chung, O'Neil, & Herl, 1999), but more recent work has begun investigating the complex interaction between both dimensions (e.g., L. Liu, von Davier, Hao, Kyllonen, & Zapata-Rivera, 2015; Rosen, 2014).

Assessment research has also sought to create innovative methods in determining ways to assess individuals' cognitive and collaborative skills through the use of CPS tasks (Hao, L. Liu, von Davier, & Kyllonen, 2015; von Davier & Halpin, 2013). Much of the current work addressing learning and performance gains in collaborative contexts has focused solely on the outcomes associated with the entire group (e.g., final team performance; Mohammed & Angell, 2003; Watson, Kumar, & Michaelsen, 1993) or the measurement of individuals' skills using traditional assessments that are external to the collaborative activity (e.g., multiple-choice posttests; Deslauriers, Schelew, & Wieman, 2011; Kirschner, Paas, Kirschner, & Janssen, 2011). These methods of assessment provide little information about the content knowledge and skills that individuals demonstrate *during* engagement in a collaborative task, because the complex interactions among group members during task activity are not leveraged to provide information for a more detailed analysis of each individual's cognitive contribution to the group outcome (von Davier & Halpin, 2013). The analysis of the rich process data in a theoretical psychometric framework has been recently considered under the computational psychometrics inference model (von Davier, 2015). For this article, we seek to move in this direction, as we utilize both the actions and discourse among group members during a collaborative assessment to gain information about behaviors indicative of collaborative skills.

Patterns of Interaction

Although collaboration is generally beneficial in facilitating learning and performance outcomes, there are a number of factors that can influence the extent to which collaboration proves fruitful. One factor concerns the nature of interaction among group members. Studies have identified a number of different patterns of behaviors among collaborators in group contexts and have further shown that some patterns are more conducive for learning than others (Li & Zhu, 2013; C.-C. Liu & Tsai, 2008). For example, in the analysis of pair interaction in an English as a second language university course, Storch (2002) identified four patterns of interaction which characterize the behaviors of groups as a whole: collaborative, dominant/dominant, dominant/passive, and expert/novice.

The collaborative interaction pattern is characterized by each group member working jointly on the task, equally sharing their ideas, and engaging with their collaborators' ideas. The dominant/dominant pattern is characterized by each group member

contributing to the task, but exhibiting unwillingness to engage with each other's contributions in a way that contributes to conflict and difficulty reaching consensus. The dominant/passive pattern is characterized by a dominant member taking control of the task and displaying little effort in trying to encourage or invite contributions from a passive member who maintains a subservient role. The expert/novice pattern describes a pattern of interaction in which one member (expert) contributes more to the task than his or her partner, but also encourages and actively invites contributions from the presumably less knowledgeable partner. Storch (2002) found that pairs with a collaborative orientation (i.e., collaborative and expert/novice patterns) performed better than did pairs with a non-collaborative orientation (i.e., dominant/dominant and dominant/passive patterns; see Watanabe & Swain, 2007, for similar results). These results suggest that patterns of group behavior in which group members work jointly on all parts of the task and engage with each other's ideas can positively influence an individual's learning and performance.

Recent research has also identified an additional interaction pattern, namely cooperative, showing it to be particularly important in computer-mediated communication (CMC) group contexts (Tan, Wigglesworth, & Storch, 2010). The cooperative pattern is characterized by a division of labor such that group members contribute equally to the task, but their contributions come as a result of each member working on their own part of the task. Thus, group members engage little with each other's contributions.

Modeling Interaction Patterns

Interaction patterns, which consist of a number of behaviors associated with collaborative skill, can be a useful conceptualization of collaboration for assessment research. In this article, we present the use of the Andersen/Rasch (A/R) multivariate IRT model as an innovative means by which to model interaction patterns. The A/R model addresses tendencies in observations that can be classified into a set of *m* exhaustive and mutually-exclusive nominal categories. In personality research, for example, a person encountering a stranger might act in ways that can be categorized as "introduce oneself," "avoid contact," or "acknowledge but not engage." Resolving decades of debate, the interactionist model in personality psychology now views such behavior as determined partly by fairly persistent tendencies of individuals, or traits, and partly by fairly stable tendencies toward behavior in recurring types of situations (Endler & Parker, 1992). Self-introductions are more frequent at parties than at business conferences, for example, and less so walking down a busy street in a big city. Researchers can study such behavior in terms of probability distributions, exhibiting variation around pervasive tendencies toward the behavioral categories that are associated with persons and situations (Fleeson, 2001) and expressed in respective parameters. In the A/R model, the parameters would be vector-valued, reflecting tendencies of persons to behavior in each of the categories, and of situations to evoke those behaviors. In the current instantiation of the A/R model, we model tendencies of dyads to behave according to each of the interaction pattern categories, and of items in a simulation-based task to evoke behavior in accordance with each interaction pattern category.

The results from such a modeling approach can be useful in exploring important questions in collaboration research concerning how potential differences in language and communication styles may influence the extent to which collaborative activity is beneficial. Prior research has noted differences in behavior according to gender and cultural background. For example, research has shown that individuals from more collectivist cultural traditions (i.e., Asian, African, and Hispanic Americans) may behave more cooperatively than individuals from more individualist cultural traditions (i.e., European Americans) on group tasks in which participants can choose to compete or cooperate with another party (Cox, Lobel, & McLeod, 1991). Furthermore, males may be more likely to take a leadership role in group interactions and speak more than females (Barrett & Lally, 1999; Prinsen, Volman, & Terwel, 2007), whereas females may exhibit more positive interpersonal behaviors (Wood & Karten, 1986). In this study, we provide exploratory analyses that examine how tendencies toward interaction pattern categories relate to performance outcomes and how tendencies may differ across gender and cultural background.

Research Questions

The literature reviewed suggests that collaborative activity can benefit learning and performance outcomes, but the benefits may be particularly salient for individuals who interact in particular ways with their group members. This motivated two research questions by which to explore the use of our novel modeling approach:

1. What interaction patterns do dyads display in the Tetralogue task?
2. How do the observed interaction patterns relate to performance outcomes?

It is also possible, as suggested in the literature review, that certain ways of interacting are more or less prevalent for individuals of a particular gender or cultural background. As such, some interaction patterns, which may be more or less conducive for learning, may be more likely in dyads with individuals of a particular gender or cultural background. This motivated two additional exploratory research questions:

3. Are the observed interaction patterns differentially exhibited across mixed- and same-gender dyads and mixed- and same-race dyads?
4. Is the relationship between the observed interaction patterns and performance outcomes different for mixed- and same-gender dyads and mixed- and same-race dyads? (Results for this question can be found in Appendix A in the online Supporting Information.)

Method

Participants

Participants were recruited using Amazon Mechanical Turk in order to obtain a sample fairly representative of the U.S. adult population in terms of gender and race. Four hundred and thirty-six unacquainted dyads completed the study, with individuals ranging in age from 18 to 68. Of those participants who reported their gender (918 out of 992), 50% were males and 50% were females. There were 213

same-gender dyads (i.e., 104 male-male and 109 female-female dyads) and 212 mixed-gender dyads (i.e., male-female dyads). Of those participants who reported their race and ethnicity (877 out of 992), 78% of the participants were White, 7% were Black or African American, 5% were Asian, 5% were Hispanic or Latino, and 5% were multiracial. Three participants identified themselves as an American Indian or Alaska Native. There were 243 same-race dyads and 151 mixed-race dyads.

Task and Measures

Prior to completing the Tetralogue task, participants individually completed a science content test to assess their science knowledge, a background questionnaire, and a personality questionnaire. The background questionnaire included items to obtain information such as participants' race, gender, age, highest level of education completed, native language, computer use, and parents' highest level of education.

The three-part task that participants completed was a simulation-based collaborative problem-solving task in which two individuals worked together in a virtual laboratory to solve a science problem about making predictions about volcanic eruptions. Two participants were automatically formed into a dyad based on whether they logged into the system within a similar time range. The current analyses focus on the first part of the task in which a virtual scientist agent presented to the dyads background information about volcanoes and a simulation displaying the sequence of seismic events that lead to volcanic eruptions. Dyads were then asked a series of seven selected response questions in order to assess their knowledge of the previously presented information. Each participant was first prompted to respond to the question individually on the screen. System prompts were then displayed to facilitate collaborative dialogue as participants used a chat box to discuss their answers. Participants were next given the opportunity to revise the individual answer previously chosen and then one participant was randomly selected to submit a team response. When participants submitted the team response, their partners were able to see the selected answer choice (see L. Liu et al., 2015, for more in-depth discussion of the task). Screenshots of the Tetralogue task can be found in Appendix B in the online Supporting Information.

Coding

The chat logs for each dyad and log files detailing participants' actions as they completed the task were used to code for the interaction patterns that were displayed by the dyads. Seven items making up the first part of the task were separately coded for the interaction patterns that were exhibited. That is, interaction patterns were coded at the item level. Modified versions of Storch's (2002) and Tan et al.'s (2010) models for dyad interaction patterns were used to create a rubric for identifying how participants interacted with their collaborator. In line with those models, the cooperative, collaborative, dominant/dominant, dominant/passive, and expert/novice interaction patterns were included in the rubric. One additional interaction pattern, not found in previous literature, was added to account for a recurring pattern of

behavior that could not be classified by these models. This was the "fake collaboration" pattern in which participants' dialogue exhibited collaboration as discussed in the literature cited previously, but their subsequent individual response actions suggested otherwise. Thus, there was a mismatch between participants' chat dialogue (i.e., expressing agreement) and behavior when independently responding to questions (i.e., not choosing the agreed-upon answer). This pattern was labeled *fake collaboration* and will be discussed further in the Results section. Independent raters double-coded 30% of the data. The interrater reliability determining consistency among raters was found to be Kappa = .83. All discrepancies among raters were discussed to reach consensus on assigning a final code. After finding that acceptable interrater reliability was reached on the subset of the data (Kappa > .60; Landis & Koch, 1977), the remaining data were coded by a single rater. Across the first part of the task, only four interaction patterns were observed. These were the cooperative, collaborative, fake collaboration, and dominant/dominant interaction patterns. The remaining two interaction patterns were not exhibited in the data. The observed interaction patterns will be described in more detail in the Results section.

Model

The data were analyzed using an Andersen/Rasch multivariate IRT model[1] (Andersen, 1973, 1995), estimated using a full Bayesian model, using Markov chain Monte Carlo (MCMC) estimation, implemented in the computer program JAGS (Plummer, 2003) via R. Relative to conventional IRT models that deal with a unidimensional latent trait, A/R models allow more than one parameter to be associated with each person, each indicating a propensity toward a certain type of response. A/R models are useful for polytomous items (i.e., items with more than two response categories), in which each response category is associated with one of the response types. Tasks are also associated with a vector of parameters, indicating their propensity to be responded to with the different response types. Thus, A/R models can deal with mixtures of strategies or states within individuals (Andersen, 1995; Huang, 2003).

Huang (2003) and Huang and Mislevy (2010) investigated the use of the A/R model for such an application, using it to model individuals' propensities toward particular types of responses as opposed to expected correctness. In particular, they examined college students' problem solving in physics, viewing students as being in mixed-model states, or simultaneously having a number of conceptions/misconceptions with regard to physical phenomena related to force and motion. With the A/R model, each student was seen as having a certain propensity to answer a question in accordance with one of three conceptions. The specific physics problems were modeled as tending to elicit particular states, in terms of propensities to evoke responses associated with those conceptions. In this article, we use the A/R model in a similar manner, with strategies or states corresponding to interaction patterns. Specifically, at a given point in time, dyads are seen as having propensities to interact in such a way that is in accordance with any of the interaction patterns described previously. In addition, the science simulation items in the Tetralogue task are parameterized in terms of their propensity to elicit certain interaction patterns among dyads.

The A/R model takes a form in which there are m interaction pattern categories and dyad and item parameters that correspond to each of them. Let X_{ij} ($i = 1, \ldots, n$, $j = 1, \ldots, k$) be interaction pattern category variables (i is the index for dyads and j is the index for items) where X_{ij} can be any integer between 1 and m. The A/R model gives the probabilities of a behavioral pattern from dyad i to item j in interaction pattern category m as

$$P\left(X_{ij} = p\right) = \frac{\exp\left(\theta_{ip} + \beta_{jp}\right)}{\sum_{p=1}^{m} \exp\left(\theta_{ip} + \beta_{jp}\right)}, \tag{1}$$

where p is an integer between 1 and m, θ_{ip} is the parameter that represents dyad i's propensity to exhibit interaction pattern p, and β_{jp} is the parameter that represents item j's propensity to evoke interaction pattern p. There are m probabilities for each dyad on a given item in the Tetralogue task, thus corresponding to the probability of exhibiting any given interaction pattern for that dyad on that item (Huang & Mislevy, 2010).

The results from the analysis using the A/R model will be interpreted as follows. For the item parameter $\beta(j, p)$, in which j represents the jth item and p corresponds with the interaction pattern, the β with the highest mean value is considered as having the highest propensity for that item to elicit the corresponding interaction pattern. Similarly, for the dyad parameter $\theta(i, p)$, in which i corresponds with the dyad and p corresponds with the interaction pattern, the θ across the interaction patterns that has the highest mean value will indicate the interaction pattern that a dyad has a greater tendency to display (Huang & Mislevy, 2010). For any item and any dyad, however, there is a non zero probability of a response in any response category as given by (1). The A/R model allows the possibility that a given dyad might engage in any of the interaction patterns across the items, but we model whether different dyads have different tendencies toward each interaction pattern. What each dyad does is not constant, but the model presumes that the tendencies are constant over different situations. What each dyad will do in different situations depends partly on the situation's tendencies to evoke certain interaction patterns, but also on the inherent variation in what people do in situations.

In specifying the model, standard normal prior distributions (0,1) were assigned to the item and dyad parameters. In the Bayesian estimation procedure, the means of these normal distributions were themselves modeled with normal prior distributions, creating hyperpriors so that we did not impose too much information about the parameters in the prior distributions. We ran three MCMC chains starting with different initial values for all parameters in each chain. In each case, the initial values were drawn from normal distributions. Convergence was checked by visual examination of history and trace plots juxtaposing the chains, and by using the Gelman-Rubin diagnostic to check stochastic convergence of all parameters in all chains. The Gelman-Rubin diagnostic compares the variances within each chain and the variance between chains to assess whether all chains converge to the same target distribution. Each of these convergence methods demonstrated that convergence had been obtained. The Gelman-Rubin diagnostic was used to determine the number of initial iterations that should be discarded, and accordingly the first 20,000 samples were discarded as a

burn-in leaving (after thinning) 60,000 draws per chain for obtaining the parameter estimates of the posterior distribution. The accuracy of the obtained posterior means was assessed by comparing the Monte Carlo error (MC Error) with the sample standard deviation for each parameter. The MC Error adjusts the sample standard error for autocorrelation. As a rule of thumb, the simulation should run until the MC error for each of the parameters is less than about 5% of the sample standard deviation (Ette & Williams, 2007; Powers & Xie, 2008). The JAGS syntax for the A/R model can be found in Appendix C in the online Supporting Information.

Results

RQ1: What Interaction Patterns Do Dyads Display in the Tetralogue Task?

Four interaction patterns, namely the cooperative, collaborative, fake collaboration, and dominant/dominant interaction patterns, were displayed as dyads completed the first seven items of the Tetralogue task. The cooperative interaction pattern corresponded to a pattern in which both participants in a dyad shared ideas relatively equally, but there was little engagement with the ideas that were shared. Specifically, participants would often work independently to generate their own answer and simply disclose the answer to their partner. If their answers were the same a simple agreement occurred (e.g., "ok" or "yeah") and if the answers were different one member would simply choose to defer to the other. Dyads did not co-construct their answers and there were little to no deliberations, explanations, or evidence provided for contributions, particularly for the easier items. As a result, participants often shared information and that information was accepted by their partner with little information as to why.

The collaborative interaction pattern was indicated by both participants contributing relatively equally during the problem-solving process in such a way that they were jointly constructing a response. Furthermore, the dyads fully engaged with each other's ideas as participants critically evaluated their partner's contributions and provided explanations and evidence for their own contributions. This interaction pattern is distinguished from the cooperative interaction pattern by joint construction and negotiation of ideas to come to a solution rather than a more simple division of labor in which participants individually construct their own response and merely share it with their partner as with the cooperative pattern.

The fake collaboration interaction pattern included instances in which both participants in the dyad contributed information and seemingly worked together to reach a consensus. However, revised individual response choices demonstrated that participants maintained their own (and different) responses. In particular, participants would appear to agree with their partner's answer choice (e.g., by saying "that sounds good" or "I agree"), but then choose a different response when given the opportunity to input their private, individual response. Thus, this pattern is characterized by a mismatch between participants' chat dialogue and behavior in revising their individual responses such that participants appeared to collaborate publicly and agreed with their partner, but privately maintained their own answer choice.

The dominant/dominant interaction pattern was indicated by both participants contributing to the task, but often outwardly seeking to maintain their own response.

As a result, they exhibited an unwillingness to engage with each other's contributions, making it difficult to reach consensus and causing disagreements to occur. Often participants did not treat the Tetralogue as a joint task; they did not even attempt to reach consensus in their individual or team responses. For example, participants displaying this interaction pattern often shared their responses with their partner, chose not to deliberate about which answer to use, and utilized their own answer choices for the final answer. Sample dyad conversations displaying each of these interaction patterns can be found in Appendix D in the online Supporting Information.

Table 1 displays the number of instances each of these interaction patterns was exhibited across the first seven items of the task. The cooperative interaction pattern was displayed more than all other interaction patterns. The dominant/dominant interaction pattern was the second most likely pattern to be displayed, followed by the collaborative and fake collaboration interaction patterns. This pattern of results was most extreme for item 7, in which virtually all dyads displayed the cooperative interaction pattern and none displayed the fake collaboration interaction pattern. This lack of variation in interaction patterns was likely due to item 7 being much easier than the other items; most dyads (99%) answered the question correctly. Item 4 showed a slightly different pattern of results, with the collaborative, fake collaboration, and dominant/dominant interaction patterns being displayed more often than in other items.

Table 2 summarizes the dyad parameter estimates for the first six dyads from the A/R model in order to show how the dyad parameter estimates are interpreted (a sample of the item parameter estimates can be found in Appendix E in the online Supporting Information). In Table 2, $\theta[i, p]$ corresponds with the parameter estimates for dyad i and interaction pattern p (1 = cooperative; 2 = collaborative; 3 = fake collaboration; 4 = dominant/dominant). Higher mean values correspond to the interaction pattern that a dyad had a greater propensity to exhibit. This table demonstrates some of the variation in propensities for dyads to display particular interaction patterns. For

Table 1
Frequency of Observed Interaction Patterns for Each Item

	Interaction Patterns			
Item	Cooperative	Collaborative	Fake Collab	Dominant
1	360	46	36	54
2	407	14	29	46
3	395	32	31	38
4	295	68	64	69
5	464	12	9	11
6	431	23	17	25
7	487	2	0	7
Total	2,839	197	186	250

example, the interactions of dyad 6 on the seven items were coded as 1,1,2,3,1,1,1. The posterior means for (θ_{61}, θ_{62}, θ_{63}, θ_{64}) expressing this dyad's propensities for the four interaction patterns based on the coding of their interactions across the seven items was 1.11, 0.19, 0.16, −1.46 as shown in Table 2, indicating higher propensity toward the cooperative pattern relative to the other interaction patterns. The posterior standard deviations for these parameters is fairly large—each around 1—which indicates that these parameter estimates are not very precise in distinguishing whether this dyad had, for example, a higher propensity toward the collaborative (2) or fake collaboration (3) interaction pattern. This is likely due to the small number of items per dyad, which suggests the need for future work in exploring these patterns with more items.

Most of the dyads in Table 2 had a greater propensity to display the cooperative interaction pattern than other patterns. Although dyad 2 had a large propensity to show the cooperative interaction pattern and no other patterns, all other dyads listed in the table had a tendency to display other patterns as they engaged in the task. For example, dyads 1 and 4 had a tendency to also display the dominant/dominant interaction pattern, and dyads 3, 5, and 6 had propensities to display the collaborative and/or fake collaboration interaction patterns.

In exploring the interaction patterns with the highest mean value for all dyads who completed the study (i.e., the one interaction pattern a dyad had the greatest propensity to exhibit), 423 dyads (85.3%) had a tendency to display the cooperative interaction pattern, 8 dyads (1.6%) had a tendency toward the collaborative interaction pattern, 3 dyads (0.6%) had a tendency toward the fake collaboration interaction

Table 2
The First Six Dyads' Summary Statistics for the Dyad Parameter

Dyad Parameter	Mean	*SD*	MC Error[b]	Dyad Parameter	Mean	*SD*	MC Error[b]
θ[1,1][a]	1.32	0.97	0.04	θ[4,1]	0.83	0.95	0.03
θ[1,2]	−0.91	1.11	0.03	θ[4,2]	−0.92	1.09	0.03
θ[1,3]	−0.94	1.11	0.03	θ[4,3]	−0.22	1.05	0.03
θ[1,4]	0.53	1.60	0.05	θ[4,4]	0.31	1.58	0.05
θ[2,1]	2.14	1.07	0.03	θ[5,1]	1.59	1.03	0.03
θ[2,2]	−0.44	1.15	0.03	θ[5,2]	−0.48	1.13	0.03
θ[2,3]	−0.47	1.16	0.03	θ[5,3]	0.24	1.10	0.03
θ[2,4]	−1.23	1.91	0.05	θ[5,4]	−1.35	1.87	0.05
θ[3,1]	1.59	1.03	0.03	θ[6,1]	1.11	1.00	0.03
θ[3,2]	0.27	1.09	0.03	θ[6,2]	0.19	1.07	0.03
θ[3,3]	−0.51	1.14	0.03	θ[6,3]	0.16	1.08	0.03
θ[3,4]	−1.35	1.87	0.05	θ[6,4]	−1.46	1.84	0.05

Note. The number of MC draws (i.e., the chain length) to obtain these statistics is 60,000.
[a]θ[*i*, *p*] corresponds with the parameter estimates for dyad *i* and interaction pattern *p*(1 = cooperative; 2 = collaborative; 3 = fake collaboration; 4 = dominant/dominant). In each MCMC cycle, the four parameters of a dyad are constrained to sum to zero.
[b]MC (Monte Carlo) Error is the autocorrelation-adjusted standard error of the estimated posterior mean.

pattern, and 62 dyads (12.5%) had a tendency toward the dominant/dominant interaction pattern when compared to all other patterns.

RQ2: How Do the Observed Interaction Patterns Relate to Performance Outcomes?

In investigating RQ2, dyad parameter mean estimates that corresponded to each of the four interaction patterns were correlated with dyads' team performance outcomes for the seven items of the simulation task. Propensity for displaying the cooperative interaction pattern, $r(494) = .28$, $p < .001$, and collaborative interaction pattern, $r(494) = .11$, $p = .02$, was positively correlated with performance outcomes. Conversely, propensity for displaying the dominant/dominant interaction pattern was negatively correlated with performance outcomes, $r(494) = -.21$, $p < .001$. There was no significant relationship between propensity for displaying the fake collaboration interaction pattern and performance outcomes, $r(494) = -.021$, $p = .64$.

RQ3: Are the Observed Interaction Patterns Differentially Exhibited Across Mixed- and Same-Gender Dyads and Mixed- and Same-Race Dyads?

For RQ3, the dyad parameter mean estimates corresponding to each interaction pattern were utilized to determine if these values were different across mixed- and same-gender and mixed- and same-race dyads.

Gender. The mean dyad propensities for each interaction pattern across mixed- and same-gender dyads are displayed in Table 3. A one-way multivariate analysis of variance (MANOVA) was conducted to examine differences for propensities toward each pattern across mixed- and same-gender dyads. Results revealed no significant differences between these groups in their propensities to display each of the observed interaction patterns, $F(4,420) = 1.27$, $p = .28$, $\eta_p^2 = .01$. Additional anal-

Table 3
Mean Dyad Propensities Across Interaction Patterns for Mixed- and Same-Gender Dyads

Interaction Pattern	Mean	*SD*	*N*
Cooperative			
Mixed-gender	1.45	0.60	212
Same-gender	1.42	0.64	213
Collaborative			
Mixed-gender	−0.31	0.56	212
Same-gender	−0.40	0.56	213
Fake collaboration			
Mixed-gender	−0.39	0.45	212
Same-gender	−0.44	0.47	213
Dominant/dominant			
Mixed-gender	−0.75	0.98	212
Same-gender	−0.59	1.13	213

Table 4
Mean Dyad Propensities Across Interaction Patterns for Mixed- and Same-Race Dyads

Interaction Pattern	Mean	*SD*	*N*
Cooperative			
Mixed-race	1.47	0.64	151
Same-race	1.40	0.61	243
Collaborative			
Mixed-race	−0.30	0.56	151
Same-race	−0.41	0.56	243
Fake collaboration			
Mixed-race	−0.39	0.43	151
Same-race	−0.42	0.49	243
Dominant/dominant			
Mixed-race	−0.78	1.00	151
Same-race	−0.58	1.11	243

yses were conducted to determine if separately including female-female and male-male dyads for same-gender dyads led to any differences. Results showed no significant differences in propensity toward each interaction pattern for female-female, male-male, and male-female (i.e., mixed-gender) dyads, $F(8,838) = 1.27$, $p = .25$, $\eta_p^2 = .01$.

Race. The mean dyad propensities corresponding to each interaction pattern for mixed- and same-race dyads are displayed in Table 4. A MANOVA was used to compare dyad propensities toward each observed pattern for mixed- and same-race dyads. Results showed no significant difference in propensities across these two groups, $F(4,389) = 1.14$, $p = .34$, $\eta_p^2 = .01$.

Discussion

In this study, we sought to explore the use of a novel modeling approach to examine how dyads interact with each other while completing a simulation-based collaborative science assessment. We further sought to explore how the approach could be used to determine if the observed interaction patterns were differentially related to performance outcomes and whether particular interaction patterns were more inclined to be exhibited by dyads of a similar or different composition with respect to gender and race. Four distinct interaction patterns were observed, specifically the cooperative, collaborative, fake collaboration, and dominant/dominant interaction patterns. Using a novel analytical approach, we were able to determine each dyad's propensity toward displaying each of these interaction patterns as they completed the task. There was some variation in propensities toward different interaction patterns in that dyads often displayed more than one interaction pattern during the task. However, overwhelmingly the dyads' greatest propensity was in exhibiting the cooperative interaction pattern relative to all other interaction patterns.

The observed interaction patterns were differentially related to performance outcomes. In particular, greater propensity toward the cooperative and collaborative interaction patterns were related to higher scores on the Tetralogue items. Unsurprisingly, greater propensity toward the dominant/dominant interaction pattern was related to lower scores on the Tetralogue items. Thus, engaging in collaborative behaviors, either simple or complex, were helpful in facilitating performance among the dyads. However, when collaborators did not work jointly or attempt to reach agreement, their performance suffered.

There were no significant differences in how individuals in the varying dyadic compositions interacted with each other. With regard to fairness, these results provide some preliminary evidence that this task may elicit similar response patterns from dyads of varying gender and race. While these results are important given the growing interest in the assessment of collaborative skills, future work will need to examine whether similar patterns of results emerge with a more diverse sample. Also, future work can explore additional individual characteristics and background variables (e.g., socioeconomic status, age, personality) for differences in the display of interaction patterns.

The results of the current study have important implications for the assessment of skills associated with collaboration, particularly with respect to concerns about fairness and methods for determining how to evaluate collaborative behavior. Modeling dyadic interaction patterns using the Andersen/Rasch model is a novel approach to the analysis of such data in which the output can essentially generate a profile displaying propensities to exhibit each of a number of interaction patterns. Importantly, each interaction pattern is characterized by elements of effective and poor collaborative behavior. While these results offer promise in this burgeoning area of research, much more work is needed in this area to ensure the valid and reliable assessment of collaborative skills.

Acknowledgments

This work was completed when Alina von Davier was with Educational Testing Service. The opinions presented in this article do not represent the opinions of Educational Testing Service or ACT, Inc.

Note

[1]Rasch gave a general form of a model with properties he called "specific objectivity" in Rasch (1977). The Andersen/Rasch multivariate model used here is a member of that family and has multiple person and task dimensions and multiple categories of outcomes (see Andersen, 1977, for an in-depth discussion of this model). The most familiar of these is Rasch's (1960) model for dichotomous items. The dichotomous Rasch model can be expressed as a special case of the A/R model, with two categories. This form is not very interesting when there are only two possible responses, however, since if the response for one category is p, the probability for the other response must be $1 - p$, and the parameters for the two response categories are simply mirror images of each other.

References

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Danish Institute for Mental Health.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81.

Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. H. Fischer & I. W. Molenar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 271–291). New York, NY: Springer.

Andrews, J. J., & Rapp, D. N. (2015). Benefits, costs, and challenges of collaboration for learning and memory. *Translational Issues in Psychological Science*, *1*, 182–191.

Barrett, E., & Lally, V. (1999). Gender differences in an on-line learning environment. *Journal of Computer Assisted Learning*, *15*(1), 48–60.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*, 641–651.

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, *15*, 463–493.

Cox, T. H., Lobel, S. A., & McLeod, P. L. (1991). Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of Management Journal*, *34*, 827–847.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, *332*(6031), 862–864.

Endler, N. S., & Parker, J. D. (1992). Interactionism revisited: Reflections on the continuing crisis in the personality area. *European Journal of Personality*, *6*(3), 177–198.

Ette, E. I., & Williams, P. J. (2007). *Pharmacometrics: The science of quantitative pharmacology*. Hoboken, NJ: John Wiley.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011–1027.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*, 111–143.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 1–15). Heidelberg, Germany: Springer.

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In *Proceedings of the 11th international conference on computer-supported collaborative learning*. Gothenburg, Sweden: International Society for the Learning Sciences.

Huang, C.-W. (2003). *Psychometric analyses based on evidence-centered design and cognitive science of learning to explore students' problem-solving in physics* (Unpublished doctoral dissertation). University of Maryland, College Park, College Park, MD.

Huang, C.-W., & Mislevy, R. J. (2010). An application of the polytomous Rasch model to mixed strategies. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 211–228). New York, NY: Routledge.

Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning and Instruction*, *21*, 587–599.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Li, M., & Zhu, W. (2013). Patterns of computer-mediated interaction in small writing groups using wikis. *Computer Assisted Language Learning*, *26*(1), 61–82.

Liu, C.-C., & Tsai, C.-C. (2008). An analysis of peer interaction patterns as discoursed by on-line small group problem-solving activity. *Computers and Education*, *50*, 627–639.

Liu, L., von Davier, A. A., Hao, J., Kyllonen, P., & Zapata-Rivera, J.-D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA: IGI Global.

Mohammed, S., & Angell, L. C. (2003). Personality heterogeneity in teams which differences make a difference for team performance? *Small Group Research*, *34*, 651–677.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015 collaborative problem solving framework*. Paris, France: OECD Publishing.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria: Technische Universit at Wien.

Powers, D. A., & Xie, Y. (2008). *Statistical methods for categorical data analysis* (2nd ed.). Bingley, UK: Emerald Group.

Prinsen, F. R., Volman, M. L. L., & Terwel, J. (2007). Gender-related differences in computer-mediated communication and computer-supported collaborative learning. *Journal of Computer Assisted Learning*, *23*, 393–409.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58–94.

Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. *Technology, Knowledge and Learning*, *19*(1–2), 147–164.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119–158.

Tan, L. L., Wigglesworth, G., & Storch, N. (2010). Pair interactions and mode of communication: Comparing face-to-face and computer mediated communication. *Australian Review of Applied Linguistics*, *33*(3), 27.1–27.24.

von Davier, A. A. (2015, July). *Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement*. Invited talk presented at the Workshop on Machine Learning for Education, International Conference of Machine Learning, Lille, France. Retrieved May 24, 2016, from http://dsp.rice.edu/ML4Ed_ICML2015

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report No. 13-41). Princeton, NJ: Educational Testing Service.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*(2), 121–142.

Watson, W. E., Kumar, K., & Michaelsen, L. K. (1993). Cultural diversity's impact on interaction process and performance: Comparing homogeneous and diverse task groups. *Academy of Management Journal*, *36*, 590–602.

Wood, W., & Karten, S. J. (1986). Sex differences in interaction style as a product of perceived sex differences in competence. *Journal of Personality and Social Psychology*, *50*, 341–347.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Appendix A includes the results from Research Question 4 regarding the differences in relationships between the observed interaction patterns and performance outcomes for mixed-and same-gender and mixed-and same-race dyads.
Appendix B displays screenshots of the simulation-based collaborative science assessment, the Tetralogue, used in the study.
Appendix C displays the JAGS syntax for the Andersen/Rasch model used in the study.
Appendix D provides sample dyad conversations that display each of the observed interaction patterns.
Appendix E displays a sample of the item parameter estimates from the Andersen/Rasch model.

Authors

JESSICA J. ANDREWS is Associate Research Scientist, Educational Testing Service, 660 Rosedale Rd, MS 12-T, Princeton, NJ 08541; jandrews@ets.org. Her primary research interests include collaborative problem solving, educational applications of cognitive psychology, and technology-enhanced learning and assessment.

DEIRDRE KERR is Associate Research Scientist, Educational Testing Service, 90 New Montgomery St, San Francisco, CA 94105; de_kerr@hotmail.com. Her primary research interests include educational data mining and the analysis of educational video games and simulations. She is now at Sony Interactive Entertainment.

ROBERT J. MISLEVY is Frederic M. Lord Chair in Measurement and Statistics, Educational Testing Service, 660 Rosedale Road, MS 12-T, Princeton, NJ 08541; rmislevy@ets.org. His primary research interests are psychometrics and assessment design.

ALINA VON DAVIER is Vice President of Research and Development, ACT, Inc., 500 ACT Drive, Iowa City, IA, 52243-0168; alina.vondavier@act.org. Her main research interest is in developing learning and assessment systems.

JIANGANG HAO is Research Scientist, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541; jhao@ets.org. His primary research interests include collaborative problem solving, game-based assessment, and data analytics.

LEI LIU is Research Scientist, Educational Testing Service, 18E, 660 Rosedale Road, Princeton, NJ 08540; lliu001@ets.org. Her primary research interests include science education, collaborative problem solving, and technology-enhanced assessment.