Modeling Data from Collaborative Assessments: 
Learning in Digital Interactive Social Networks

Mark Wilson  
University of California, Berkeley, USA

Perman Gochyyev  
University of California, Berkeley, USA

Kathleen Scalise  
University of Oregon, USA

Abstract:  
This paper summarizes assessment of cognitive skills through collaborative tasks, using field-test results from the Assessment and Teaching of 21st Century Skills (ATC21S) project. This project, sponsored by Cisco, Intel and Microsoft, aims to help educators around the world enable students with the skills to succeed in future career and college goals. In this paper, ATC21S collaborative assessments focus on the project’s “ICT Literacy — Learning in digital networks” learning progression. The paper includes a description of the development of the learning progression, as well as examples and the logic behind the instrument construction. Assessments took place in random pairs of two students in a demonstration digital environment. Modeling of results employed unidimensional and multidimensional item response models, with and without random effects for groups. The results indicated that based on this data set, the models that take group into consideration in both the unidimensional and the multidimensional analyses fit better. However, the group-level variances were substantially higher than the individual variances. This indicates that a total individual estimate of group plus individual is likely a more informative estimate than individual alone but also that the performances of the pairs dominated the performances of the individuals. Implications are discussed in the results and conclusions.

Keywords: assessment of collaboration, 21st century skills, group assessment, pairs, dyads, measurement models, IRT, digital literacy, ATC21S, Assessment and Teaching of 21st Century Skills, group effect, multilevel models, Rasch group model.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jedm.12134.

This article is protected by copyright. All rights reserved.
Acknowledgements: We thank Patrick Griffin, Esther Care, Claire Scoular and Nafisa Awwal, from the Assessment Research Center at the University of Melbourne, for developing software used to collect the data, and collecting the data used in this paper. We would like to thank the editors, associate editor, and reviewers for their insightful comments and helpful suggestions.

**Introduction**

Collaboration among people is a long-standing practice in workplaces, schools, and many other environments. Historically, humans often tend to combine skills and knowledge, divide work, and come to common understandings that advance objectives. Numerous recent reports indicate that collaboration is increasingly important in today’s complex work world, therefore of increasing interest for teaching and assessing with students (Binkley et al., 2012; OECD, 2013).

The efforts described in this paper were grounded in the Assessment and Teaching of 21st Century Skills project (ATC21S), launched in 2009 as a response to these transitions in the world economy. The project was founded by three information technology companies (Cisco, Intel, Microsoft—CIM) with a broad group of academics (Griffin, McGaw, & Care, 2012).

The ATC21S project was strongly focused on preparing students for goals and challenges they will face after completing primary and secondary schooling (Griffin & Care, 2015). As the ATC21S project reports, for career-readiness, employers may base their recruitment decisions on school and job qualifications. But decisions to retain people are often made on the basis of team-working skills, collaboration, and ability to work within an organization (Binkley et al., 2012).

This paper summarizes results from one approach to the assessment of cognitive skills through collaborative tasks. Using field-test data from the ATC21S project, unidimensional and multidimensional item response models, both with and without random effects for groups, are explored to generate inferences about student and student group performance in the tasks.

The next section of the paper gives some background on assessment within collaborative environments. This is followed by sections describing the constructs that the ATC21S project
developed in the area of Learning in Digital Interactive Social Networks, and then two sections on methods: (a) one on the collaborative tasks and item formats used, and (b) a second on the analytic techniques used. These are followed results and a discussion of the implications.

**Assessment within collaborative environments**

Teachers, schools, states and regions that would like their pupils to better master their career activities may wish to assess how likely student success is for collaborative efforts, and to do so within groups of students or with computer agents (Rosen & Foltz, 2014). However, there are numerous questions about assessing skills during collaboration (Wilson et al., 2012). Many traditional assessments in other areas focus on measuring individual performance. For collaborative tasks, ATC21S reports a key question is how to assign credit to each group member and how to attribute differences from specific group configurations (Binkley et al., 2012).

The limitations for measurement involving the use of real people rather than standardized avatars include that, if an individual report is desired, it is anticipated that students would need to participate in more groups to randomize over the range of attributes of the partner(s). This would not necessarily be difficult to do in classroom-embedded settings if enough virtual, designed opportunities are made available, since students are often expected to participate in many new and different collaborations in school, even over the course of a day or week. In this case for the ATC21S project, a condition of the assessment specification requirement was that collaboration must take place and be measured between real people. Therefore, this approach was used here.

Many other measurement complexities also enter the picture when assessing within groups (Von Davier & Halpin, 2013). Collaboration has been defined as a “coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” (Roschelle & Teasley, 1995, p. 70). As the Programme for International Student Assessment (PISA;
OECD, 2013) reports, social interaction is a vital but *insufficient condition* for collaboration, because some social interactions do not involve shared goals, accommodation of different perspectives, and/or organized attempts to achieve the goals.

In PISA’s 2015 assessment, a collaborative problem solving assessment was included and examples are available in the draft framework (OECD, 2013). The framework and task examples show use of avatars, or computer agents, for the collaboration. For PISA, the conception of collaborative, or group problem solving had its origin in the types of problems and collaborative interactions that 15-year-old students face in and out of the classroom, as well as the issue of preparedness for future life (Scalise, Mustafic, & Greiff, pending). PISA’s collaborative problem solving construct built on its 2012 individual problem solving.

Other projects besides PISA also have investigated measures of individual performance during collaboration (Laurillard, 2009) and come to similar conclusions of the complexity involved. Cakir, Zemel and Stahl (2009) show how group participants organize some activities to share discourse and co-construct meaning. An extensive body of additional research, such as from Baker and Lund (1997), documents how collaboration supports reflective activities. Any of these effects may interact for assessment in collaborative settings, so the goal involves a complex measurement challenge, or a hard-to-measure construct (Scalise, 2012).

**Learning in Digital Networks**

For this project, the goal was not to measure collaboration directly but to measure some specific skills within a collaboration context. Measurement models are then applied to generate information about student individual performance and student group performance.

In the ATC21S tasks, collaboration is encompassed within a joint activity where two or more students work together to contribute knowledge, skills, materials and procedures. The intention is that

This article is protected by copyright. All rights reserved.
they move through a series of cognitive states that can involve collection and analysis of information, formulation and testing of hypotheses that the students jointly set out to explore, or other tasks and activities relevant to the applied construct of interest.

For the analysis in this paper, ATC21S collaborative assessments focus on the project’s “ICT Literacy — Learning in digital networks” learning progression. A learning progression involves successively more sophisticated ways of thinking or acting that may be exhibited as learners gain more expert understanding and/or skill in a given area. Note that this paper does not describe longitudinal assessment, or repeated measures of individual students over time. However, learning progressions are intended to describe a construct to measure student learning; hence the use of the word “learning” in the construct name and manuscript title here.

Tasks here involve the digital literacy construct and all data were collected within randomly-assigned dyads of respondents (see Results section for more information). Two-person teams of students worked together to craft responses that focus on the project’s “ICT Literacy — Learning in digital networks” learning progression. The idea of learning in networks (Wilson, Scalise, & Gochyyev, 2015) was seen as made up of four strands (modeled later as dimensions) and summarized here, with examples in Table 1, and in Table 2 in the online appendix:

- Functioning as a consumer in network;
- Functioning as a producer in networks;
- Participating in the development of social capital through networks, which involves using, developing, moderating, leading and brokering the connections within and between individuals and social groups in order to marshal collaborative action;
- Participating in intellectual capital in networks, which involves building collective intelligence and integrating new collaborative insights into personal understandings.
The theoretical/conceptual motivation developed for the relationship between ICT Literacy and collaboration arises from the conception of digital tasks as work to which groups often contribute in digital networks, such as today’s work environments (see introduction). Note that we have picked out collaborative components in two strands of the construct – tapping shared social capital and shared intellectual capital – which expands the definition of ICT literacy into collaborative and networked environments. Thus, participating in the shared development of social capital or intellectual capital within the group becomes a constituent component of ICT Literacy in a digitally networked world. At the same time, for any of the skill estimates on the four strands, it may be possible to yield a picture of individual and group efforts through measurement models with and without random effects for groups.

The four strands are seen as interacting together in the activity of learning in networks. First, functioning as a Consumer in Networks (CiN) involves obtaining, managing and utilizing information and knowledge from shared digital resources and experts in order to benefit private and professional lives. To give an example for this first strand, it involves questions such as:

- Will a user be able to ascertain how to perform tasks (e.g. by exploration of the interface) without explicit instruction?
- How efficiently does an experienced user use a mobile device to find answers to a question?
- What arrangement of information on a display yields more effective visual search?
- How difficult will it be for a user to find information on a website?

Second, functioning as a Producer in Networks (PiN) involves creating, developing, organizing or re-organizing information and knowledge to contribute to shared digital resources.

Third, developing and sustaining Social Capital through Networks (SCN) involves using, developing, moderating, leading and brokering connections within and between individuals and social
groups in order to marshal collaborative action, build communities, maintain an awareness of opportunities and integrate diverse perspectives at community, societal and global levels.

Fourth, developing and sustaining Intellectual Capital through Networks (ICN) involves understanding how tools, media and social networks operate and using appropriate techniques to build collective intelligence and integrate new collaborative insights into personal understanding.

Table 1. Functioning as a Consumer in Networks (CiN)

<table>
<thead>
<tr>
<th>CONSUMER IN NETWORKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CiN3</td>
</tr>
<tr>
<td>Discriminating consumer</td>
</tr>
<tr>
<td>Judges credibility of sources/people</td>
</tr>
<tr>
<td>Integrates information in coherent knowledge framework</td>
</tr>
<tr>
<td>Conducts searches suited to personal circumstances</td>
</tr>
<tr>
<td>Filters, evaluates, manages, organises and reorganises information/people</td>
</tr>
<tr>
<td>Seeks expert knowledge (people through networks)</td>
</tr>
<tr>
<td>Selects optimal tools for tasks/topics</td>
</tr>
<tr>
<td>Conscious consumer</td>
</tr>
<tr>
<td>Selects appropriate tools and strategies (strategic competence)</td>
</tr>
<tr>
<td>Constructs targeted searches</td>
</tr>
<tr>
<td>Compiles information systematically</td>
</tr>
<tr>
<td>Knows that credibility is an issue (web pages, people, networks)</td>
</tr>
<tr>
<td>Emerging consumer</td>
</tr>
<tr>
<td>Performs basic tasks</td>
</tr>
<tr>
<td>Has no concept of credibility</td>
</tr>
<tr>
<td>Searches for pieces of information using common search engines (e.g. movie guides)</td>
</tr>
<tr>
<td>Knows that tools exist for networking (e.g. Facebook)</td>
</tr>
</tbody>
</table>

In Table 1, example of levels for one of the strands have been described as hypothesized construct maps, which can be seen as simple learning progressions, showing an ordering of skills or competencies involved in each. At the lowest levels of each are the competencies that one would expect to see exhibited by a novice or beginner. At the top of each table are the competencies that one would expect to see exhibited by an experienced person - someone who would be considered very highly literate in ICT. See Table 2 in the online appendix for another example construct map.

The levels in the ATC21S Learning in Digital Networks framework were developed using the BAS approach (Wilson, 2005) and input from international participating countries in Australia, Finland, Singapore and the U.S. The levels in each strand follow a similar valuing of:

This article is protected by copyright. All rights reserved.
First awareness and basic use of tools,
followed by more complex application directly relevant to teaching and learning,
with evaluative and judgmental skills emerging as experience and knowledge are gained,
moving to leadership and ability to manage and create new approaches.

Note that for the ICT Literacy assessments here, students may not have had a specific curriculum or opportunity to learn the skills in school. In this sense, the new construct with the digital networking perspective is measuring skills that may have been developed in many different settings, both formal and informal, or not at all. The idea is to obtain a measure of how well students are doing, from which schools can consider, for instance, if students should be supported in these skills. Supporting students through explicit training would likely substantially change how effective students may be at deploying these skills on demand.

Methods: Collaborative Tasks and Analytic Techniques

For the Learning in Digital Networks tasks, the Berkeley Evaluation and Assessment Research (BEAR) Center at UC Berkeley developed three scenarios in which to place tasks and questions that could be used as items to indicate where a student might be placed along each of the four strands. The same construct of learning in networks with all four strands was used for all three scenarios, but the contexts differed (Wilson, Scalise, & Gochyyev, 2014).

For the data set used here, two of the three pilot scenario-based tasks developed at UC Berkeley were selected based on the pilot results of the full set of tasks in several countries. The two scenarios were simplified and modified at the Assessment Research Center at the University of Melbourne to perform on multiple systems, be suitable for the digital infrastructure available in the Australian schools, and be suitable for large scale use. They were then administered in a larger field trial. This generated the data set used here. The two tasks that were retained contained a total set of 87...
“items” or observations. The tasks were revised for large scale use in Australia such that all items could be dichotomously machine scored, according to decision rules. The scoring decision rules were developed based on the earlier tryouts of the original scenarios, the item analysis, subject matter expertise, and tryouts of the new formats.

While the final tasks, items, and scoring rules for the project are not released and cannot be described in full detail, sample item formats will be described within the first scenario, called “Arctic Trek.” This was conceived as a "collaboration contest,” or virtual treasure hunt through science and mathematics resources. The Arctic Trek scenario views social networks as an aggregation of different tools, resources and people that build community in areas of interest.

The primary purpose of this paper is to show some approaches to modelling collaborative data, not to explore the development of collaborative tasks, for which many approaches might be taken. However for the purposes of understanding the data set, a brief summary of indicators in the Arctic Trek task are listed in Table 3 of the online appendix, which includes a summary description and overview of scoring. More complete assessment blueprints have been designed for the tasks but are not available to share since (i) the tasks themselves are undergoing study, review and revision, and (ii) more advanced scoring rules and scoring approaches are being developed to more richly and fully use process data and other data types that are available in the tasks.

Table 3 (see online appendix) provides a summary of the types of indicators (items) used from the Arctic Trek task, one of two tasks for the data set in this paper. All items shown here were used in the model analyses in this paper. Each row of the table summarizes a different type of “item concept” employed in the “task surround”, or surrounding scenario of the task. Nine item concepts for Arctic Trek are shown in Table 3 in the online appendix. Percentages are provided for the relative contribution of score points in column 1. A short description of items for the concept are shown column 2. An overview of the scoring rules is provided in column 3.
The relatively high contribution of information foraging in this task (about 31% of score points) is due to the task surround. It involves a science/math “treasure hunt” to solve a series of clues through seeking information, or foraging, in digital resources. Responding to information requests also contributed relatively more indicators (about 18% of score points), again due to the information foraging focus of this particular task.

By contrast, relatively few score points were awarded for voluntary use of the chat tool (about 3%) because this was conceived as a holistic rating over the full task. Future task versions might consider awarding score points over scenes, or portions of the tasks, for the voluntary chatting, or might use scoring engines to further evaluate the chat quality for scoring.

The “explaining” concept in the fifth bullet point also involved relatively fewer score points (about 5%). This was primarily because human raters were dropped from the field trial and scoring engines were not yet available for the task extended response items that involved explanations. Therefore the scoring rules were simply presence or absence of a provided explanation toward contributions to shared intellectual capital. Since this does not evaluate the quality of the explanation, the concept was considered somewhat less useful for the scoring model at this time, and therefore was designed to contribute fewer score points.

To provide a few examples, an opening screen\(^1\) of the Arctic Trek task from the field trial is shown in Figure 1 in the online appendix. Students work together using the chat window on the right, which displays all the chat from either player, and the role sorter tool on the left. Once any role or activity is selected by a team player, it is no longer an available role. Players can return roles to the sorter window and make different selections during their planning process with the partner, while the

\(^{1}\) We wish to thank Claire Scoular and Nafisa Awwal from the ARC at the University of Melbourne for programming this version of these items.
role sorter tool is available in the beginning portion of the task. Scoring is very simple for the data used here and could be made more complex: a player is awarded one point if at least one of the roles has been assigned to the player being scored, and another point if at least one of the activities has been assigned to the player being scored, before players move on to begin their activities. The same is done for the other player when his or her score points are awarded. In some teams, one player will have roles and activities assigned and the other will not, in other cases, other combinations are seen.

A second example is shown in Figure 2 of the online appendix, the screen for Clue 3 in the task. For Clue 3, students need to decide upon and access strategically important links related to the clue in the navigation bar (at bottom of the screen). Here, students access an online table that describes the population of bears in the Arctic Trek region. The table is color coded for population trends. Once students forage through the information for the appropriate evidence, they must adjust the slider bar to represent the colour codes identified in the population trends table and then explain their answers. They are then prompted to provide information from the chat window showing a varying perspective that the partner suggested. Score points are awarded in various strands of the ICT literacy construct for accessing the strategic link that provides the clue information, appropriate setting of the slider bar for the colour representation, provision of an explanation for their colour choice by the player being scored, and also by correctly eliciting and providing a perspective or idea from the partner through the chat tool.

A third screen example is shown in Figure 3 of the online appendix, for which students design a landscape together using a figural drawing tool. Then each player must successfully upload the landscape by choosing and submitting a file. A podcast is a new informational resource. In this screen, students receive one score point for using the digital drawing tools to design a landscape representation, another for successfully saving their landscape as a file, and a third point for uploading and sharing their representation so that others can view it.
Methods: Analytic Techniques

For analytic techniques in cases where group results are the aim, the ATC21S methodological report (Wilson et al., 2012) described techniques such as weighted analyses and/or multilevel models that allow for the dependencies of clustered data in collaborative tasks.

Ideally, some sort of measure will be needed of: (a) each member of the group, and (b) the variability of each individual across groups. Here for this data set, having a group effect in a large scale data context can be considered as introducing an ephemeral variable. Since the group exists just for the one event, this could be modeled as a random effect.

In order to avoid overly confounding group effects with member abilities, potentially one of two approaches could be taken. If classroom-based data are available or other sets of student work from multiple groups, such as from embedded assessments or other sets of classroom activities, potentially a number of reasonably randomized group experiences over the classroom setting might be available for each person. However, few classroom instructional settings or technical systems are yet available to structure formal data collection across numerous planned groups in classrooms, so this is likely a forward-looking solution that is not currently possible to implement in most settings. Alternatively for more current use or for large scale assessments, direct data of non-group results might be used for each member along with group results, to potentially model a student effect and an effect of students within groups.

We estimate the abilities of students in this complex context of observations of students working within groups using a generalized multilevel model (Adams, Wilson & Wu, 1997; Wilson, 1999; Kang and Wilson, 2003). As a start, a simple Rasch model is the following:

$$\eta_{pi} = \theta_p - \beta_i$$  \hspace{1cm} (1)

This article is protected by copyright. All rights reserved.
with $\eta_{pi}$ as the logit for item $i$ and student $p$; with $\theta_p$ as a random student ability, $\theta \sim N(0, \sigma^2)$, and $\beta_i$ as a fixed item difficulty. A multilevel version of the Rasch model is the following:

$$\eta_{pi(g)} = \theta_{p(g)} - \beta_i + \varsigma_g$$  \hspace{1cm} (2)

with $\eta_{pi(g)}$ as the logit for item $i$ and student $p$ belonging to group $g$; with $\theta_{p(g)}$ and $\varsigma_g$ distributed independently as a random student ability and random group effect respectively, assuming $\theta_{p(g)} \sim N(0, \sigma^2)$ and $\varsigma_g \sim N(0, \xi^2)$; and with $\beta_i$ as a fixed item difficulty. We will call this the Multilevel Rasch model and present it in detail below. Basically, for the Multilevel Rasch model, the probability of a certain item response depends on the item difficulty, the individual’s ability, and the group ability.

Note that Equation 2 reduces to the Equation 1 when $\varsigma_g$ (along with the $g$ subscript) is dropped from the model. Note that each student can be involved in only one group.

We set the estimation of the model within a multilevel version of the multidimensional framework called the Multidimensional Random Coefficients Multinomial Model (MRCML; Adams, Wilson & Wang, 1997). In this setting, the different dimensions are conceived of as the different construct maps in the ICT project. In order to find the relative effects of the different models, we will estimate (a) a unidimensional model without group effects, (b) a unidimensional model with group effects, (c) a multidimensional model without group effects, and (d) a multidimensional model with group effects. These will be examined for relative fit, and also other interesting patterns of results.

The MPlus Software (Muthén, & Muthén, 1998-2014) was used in the estimation of all of these models. The full presentation of the MRCML model, also presented in Adams, et al., 1997, is available in the endnote of this article (available online). We present the multilevel variant of the MRCML model next. Note all models and analyses presented in this paper are within the Rasch modeling framework (we present evidences for the model fit), and can be easily extended to the less-restrictive models (i.e., 2-PL). We focus on the Rasch model mainly due to its interpretational simplicity (see for instance Wilson & Gochyyev, 2013).
Multilevel Rasch model

For simplicity of the presentation, the multilevel unidimensional Rasch model will be presented. Extension to the multilevel multidimensional Rasch model is straightforward.

Following Raudenbush et al. (2003) and Kamata & Cheong (2007), the Rasch model can also be expressed using a two-stage formulation, as follows:

The item level (level 1) of the model can be expressed as:

$$\tau_{ip} = \pi_p + \sum_{i=1}^{I-1} \pi_{pi} W_{pi},$$

\[ (3) \]

where $W_{pi}$ is the $i^{th}$ item dummy variable for the subject $p$ with value equal to unity when $i = q$ and 0 otherwise; $\pi_p$ is the intercept term, and $\pi_{pi}$ is the coefficient associated with $W_{pi}$. Note that there is no dummy variable for the $I^{th}$ item in order to achieve full rank of the design matrix. The formulation in Equation 3 in the endnote allows elements of $W$ and $\tau$ to vary across subjects (by incorporating the subject subscript). By restricting these to be invariant at level-2, we implement the Rasch assumption regarding the item slopes (i.e., equal discriminations). Thus, the subject-level (level-2) is expressed as

$$\pi_p = \beta_0 + u_{0p}, \quad \pi_{pi} = \beta_i, \quad i = 1, \ldots, I-1,$$

\[ (4) \]

in which $u_{0p}$ is the latent variable at the respondent level (“ability”).

Note that with this restriction,

$$\tau_{ip} = \beta_0 + \beta_i + u_{0p},$$

\[ (5) \]

which is similar to the expression in Equation 1 if we express:

$$u_{0p} = \theta_p; \quad \delta_i = -(\beta_0 + \beta_i).$$

\[ (6) \]

In addition, by adding a third level subscript (e.g., for team or classroom) to the expression in Equations 3, and 4, we can express the three-level Rasch model using a three-stage formulation, with level-1 expressed as:

This article is protected by copyright. All rights reserved.
\[ \tau_{pis} = \pi_{ps} + \sum_{i=1}^{l-1} \pi_{pis} W_{pis}, \]  

and level-2 expressed as:

\[ \pi_{ps} = \beta_{0s} + u_{0s}; \pi_{pis} = \beta_{is}. \]  

and level-3 can be expressed as:

\[ \beta_{0s} = \gamma_{00} + v_{0s}; \beta_{is} = \gamma_{i0}, \]  

with \( v_{0s} \) as a level-3 (e.g., school or classroom) latent variable, assuming \( v_{0s} \sim N(0, \psi) \).

Thus, the response of person \( p \) who is in school \( s \) to item \( i \) is given by:

\[ \tau_{qsi} = \gamma_{00} + \gamma_{i0} + u_{0s} + v_{0s} \]  

with

\[ u_{0s} + v_{0s} = \theta_p; \delta_i = -\gamma_{00} + \gamma_{i0}. \]  

**Model Fit Comparisons**

When a model is a special case of the less restricted model (such as Rasch model being nested in multilevel Rasch model, or unidimensional Rasch model nested in multidimensional Rasch model), the difference in deviances is generally assumed to have a chi-square distribution with the difference in the number of parameters as degrees of freedom (Wilks, 1938). Thus, we can statistically test if the less restricted model fits the data significantly better than the simpler (more restricted) model. Note, however, that the group-level variance cannot be nonnegative and thus the null hypothesis is on the boundary of the parameter space, and thus the LR statistic does not have a simple chi square distribution. A conservative test, then can be obtained by simply dividing the naive p-value from the LR test by 2 (Rabe-Hesketh & Skrondal, 2005), used below.

**Example and Results**

Data were collected in a demonstration digital environment in Australia, for 1,562 students and 781 groups. There were 87 dichotomous items in all. The sample consisted of 47% females and

This article is protected by copyright. All rights reserved.
ages ranging between 13-17 at the time of assessment. The schools were all volunteer schools in the state of Victoria, in Australia. The students were also volunteers from within those schools. Before beginning the assessment, students were told that they will engage in collaborative activities online and will be provided with partners. They are told that the activities are for a math/science task, or for an English language literacy task, but in both cases activities are intended to provide information on how students work with the technology tools, processes, and partners provided. Students are encouraged to tap into assistance from their partners through the chat tools provided. Students are told they cannot collaborate face-to-face. They also are required to restrict all collaboration and conversation to the online tools provided within the assessment environment, which does share some screen information but now full screen sharing. Teachers in the room during the assessment monitor that instructions are followed. Once beginning the task, students can however communicate their identity and any other information they wish to share to each other, through tools such as chat windows.

Groups were assigned in a random fashion within the classroom before the task was administered—for this data set, each group consists of two students. The specific method of assignment that generates the degree of randomness is somewhat new to large-scale assessment but is a more common approach in e-learning products used in the classroom. It is controlled by the delivery platform, through the use of timestamps. For a given classroom, an assessment delivery window (for instance a particular day at a given hour) is selected. Students in that classroom are assigned to take a certain number of items or tasks in advance, in this case, the items in both of the retained scenarios for the field trial. When they log in, each student in the class receives a timestamp, with the timestamps recorded to fractions of a second, and the individual timestamps of login are used to place the students into groups. The first two timestamps are grouped together, then the next two are paired, the next two and so forth. Students are not allowed to collaborate or plan their login in advance or off-line.
In this case, the use of timestamp randomization was used at the start of each task, and was the only form of assignment available in the platform. The technique is practical for schools. The assignment technique has been piloted in prior work with other collaborative problem solving tasks (Care, Griffin, Scoular, Awwal, & Zoanetti, 2015) and was found to generate a reasonable degree of intermixing and randomization of groups, although other approaches might be used if a more systematic stratification is desired. Looking first at the fit results for unidimensional models, as shown in Table 4, the models incorporating the group effects fits considerably better in a statistical significance sense using a LR test ($\chi^2 = 583.89$ on one degree of freedom), and also in terms of Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). Considering practical significance too, the person variance for the Rasch model is 1.230 (SE 0.063), while the variances for the Multilevel Rasch model are 0.146 (SE 0.019) and 1.207 (SE 0.084), for individual and group effects, respectively. These results show how strongly the performances of the pairs figure compared to the results for the individuals—about 90% of total variance is explained by groups.

Table 4. Model fit statistics for the estimated models

<table>
<thead>
<tr>
<th>Model</th>
<th>deviance</th>
<th># of parameters</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional Rasch$^a$</td>
<td>50306.91</td>
<td>87</td>
<td>50480.91</td>
<td>50946.68</td>
</tr>
<tr>
<td>Unidimensional Multilevel Rasch$^a$</td>
<td>49723.02</td>
<td>88</td>
<td>49899.02</td>
<td>50370.15</td>
</tr>
<tr>
<td>Multidimensional Rasch$^b$</td>
<td>49824.48</td>
<td>96</td>
<td>50016.48</td>
<td>50530.44</td>
</tr>
<tr>
<td>Multidimensional Multilevel Rasch$^c$</td>
<td>49028.98</td>
<td>106</td>
<td>49240.98</td>
<td>49808.48</td>
</tr>
</tbody>
</table>

$^a$ MML estimation with 50 nodes  
$^b$ Monte Carlo MML estimation with 5000 nodes
Monte Carlo MML estimation with 100000 nodes

Switching to the multidimensional models, we can see that a similar story is repeated here, with the results equally supporting accepting the model incorporating group in the sense of statistical significance using a LR test ($\chi^2 = 795.50$ on 10 degrees of freedom), as well as AIC and BIC, for the example data set here.

With non-group models, we first investigated item fit using the weighted mean square (infit) statistics often used in the Rasch framework. A convention of $3/4$ (0.75) and $4/3$ (1.33) is used for an acceptable lower and upper bound (Adams & Khoo, 1996).

In both models (non-group unidimensional and non-group multidimensional models), there were two items that were outside of the boundaries of the acceptable fit (for item wsl905, infit statistic of 1.63 with unidimensional and 1.53 with multidimensional; for item atl0503, infit statistic of 1.36 and 1.47 from unidimensional and multidimensional models respectively).

When we account for clustering using both unidimensional and multidimensional models, we relied on the univariate item fit statistic that is based on standardized residual for each item (Muthen & Muthen, 2011). These residual values are assumed to have a chi square distribution with 1 df. This is used to investigate whether the model (i.e, the multilevel MRCML model) fits each item marginally. Using chi-square critical values of 3.84 for the 0.05 level and 6.64 for the 0.01 level, we found that only two items have standardized residual significant at 0.05 (one of which is the same item: wsl905).

When comparing these univariate fit statistics from the multilevel model with the ones obtained from the simpler models (the ones that ignore group clustering), we confirmed that standardized residuals do not get worse when a multilevel model is used.
We can examine two different indices of practical significance: the relative variances (as above), and also the correlation coefficients. The variances for the estimated multidimensional models are shown in Table 5. The impact of including the group effects can be clearly seen—the variance of each of the dimensions in the non-group model is split among the person and group levels in the multidimensional Rasch Group model, with preponderance going to the group level. This varies somewhat, with the largest individual proportion for Intellectual Capital.

Table 5. Person and group variances for estimated multidimensional models (SE in parentheses).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Non-group</th>
<th>Multilevel-level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Person-level</td>
</tr>
<tr>
<td>Consumer</td>
<td>1.096 (0.070)</td>
<td>0.100 (0.027)</td>
</tr>
<tr>
<td>Producer</td>
<td>2.051 (0.155)</td>
<td>0.156 (0.043)</td>
</tr>
<tr>
<td>Social</td>
<td>0.923 (0.171)</td>
<td>0.120 (0.089)</td>
</tr>
<tr>
<td>Intellectual</td>
<td>2.154 (0.147)</td>
<td>0.555 (0.075)</td>
</tr>
</tbody>
</table>

The correlations estimated in the non-group model are shown in Table 6, and those in the model incorporating group in Table 7. These are all direct estimates from the model solutions.

Table 6. Correlations among the dimensions for different levels from the non-group model

<table>
<thead>
<tr>
<th>Level</th>
<th>Dimension</th>
<th>Consumer</th>
<th>Producer</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Producer</td>
<td>.798</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td></td>
<td>.907</td>
<td>.744</td>
<td></td>
</tr>
<tr>
<td>Intellectual</td>
<td></td>
<td>.737</td>
<td>.755</td>
<td>.879</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved.
The results indicate that based on this data set, the models incorporating group in both unidimensional and multidimensional analysis fit better than the non-group models. Thus, these multidimensional models would provide a more sound basis for interpretation and use in policy-based and research-based analyses. In addition, the group-level variance was in all cases substantially higher than the individual variance. Some may be surprised at this result—but we believe that this surprise is actually due to fact that we seldom have data like these, and that it will become less surprising as more cases of collaboration are examined. What it does do is illustrate the power of the collaborative link between the two students—in this case, the collaborative environment seems to have had a greater effect than the individual differences between the students. This is parallel to the findings in work environments that effective teamwork and collaboration skills are key to workplace success, see paper introduction.

The correlations also tell a story. Comparing the results in Tables 6 and 7, we note that, for each relationship, the non-group correlation is in-between the student level and the group-level correlation (in one case its the same as the group-level). This is not too surprising, as one might think of the non-group correlation resulting from a “blending” of the group-level and the student-level information. However, another pattern was also revealed—for the three relationships involving the

<table>
<thead>
<tr>
<th>Level</th>
<th>Dimension</th>
<th>Consumer</th>
<th>Producer</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Producer</td>
<td>.832</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social</td>
<td>.895</td>
<td>.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intellectual</td>
<td>.581</td>
<td>.921</td>
<td>.883</td>
</tr>
<tr>
<td>Group</td>
<td>Producer</td>
<td>.781</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social</td>
<td>.907</td>
<td>.665</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intellectual</td>
<td>.789</td>
<td>.679</td>
<td>.960</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved.
Producer construct the person-level correlation is higher, and for the other three relationships, the group-level is higher.

Note that item parameters are assumed to be same across groups. The same is true for other estimated parameters such as variances and covariances. An alternative to this would be the so-called multiple group IRT – also known as the fixed-effects IRT model. Instead, the approach we have taken is to disentangle the residual into within and between group effects assuming these groups effects have distributions and hence, are random-effects. Therefore, the variances and covariances in the model are varying across levels only (i.e., student level and group level).

**Conclusion and Implications for Future Work on Collaboration**

As noted in the Results section, this paper explores results for collaborative assessment for a data set collected in a demonstration digital environment including 1562 students composed in pairs of 781 groups. Assessment activities included 87 dichotomous scored observations. For both unidimensional and multidimensional models, the model incorporating group fit significantly better, as well as yielding interesting results. It also illustrates the challenge of using results from collaborations if the main aim is to gain measures of individual students. On the one hand, if one uses the individual student residual estimate ($\mu_0s$ in Equation 1), which is the estimate most free from the collaborative interaction, then one faces the possibility that there will be not much variation left (as is the case in this data set). On the other hand, if one uses the total estimate (as in Equation 11), then, although one regains variation, it is also clear that much of that variation may be due to the collaboration rather than the individual (as is the case in this data set). This leads one to speculate that multiple group experiences might need to be combined for a given student, in order to disentangle the group and individual effects, as discussed in the paper introduction. This is theoretically consistent with what is known of group collaboration, where a student will often perform somewhat differently.
or accept different roles, responsibilities and processes from one group situation to the next. Of course, it will also require a more complex model, one that incorporates crossed random effects.

Of course, one could also consider the consequences of estimating student abilities in the different settings—such as, alone, in groups (perhaps different types of groups), and when the source of the data is both individual and group-based. This is an important research frontier for the application of psychometrics to data arising from group-level interactions.

More work is needed in understanding the behavior of correlations between dimensions within each of the levels (i.e., student-level or group-level). For this dataset, we couldn’t find a clear and consistent interpretation of the pattern regarding these correlations when one moves from multidimensional model to its multilevel version. We do observe, not surprisingly, that correlations between dimensions in the non-group multidimensional model are bounded by the relevant correlations at the person-level and group-level from the more complex model.

Future work might profitably consider both the constancy and change in student performance across groups during collaboration. Other important considerations include the possibility of including some individual observations for some items, which could be represented by having a group labeled “0” for those items, with a null group effect, (i.e., $\xi_0$ constrained to be 0), so that we would have a standard Rasch model for these items as in Equation 1, where, $p$ indexes students and $i$ indexes items. In addition, students could be involved in different subgroups within the class, and, indeed, one group might well be the whole class. Of course, the dichotomous Rasch model illustrated here can readily be extended to other models, including polytomous models such as the partial credit model (Masters, 1982), and to more complex item structures such as testlets (Rosenbaum, 1988; Wilson & Adams, 1995). Also, explanatory item response modeling could be explored following De Boeck and Wilson (2004). Please see the second endnote (available online) for a note related to the Rasch model.
In summary, this paper employs an empirical example to explore the challenges and opportunities of 21st century collaborative assessments. New measurement modeling approaches may be needed to examine individual and group effects, such as modeling a group aspect of the collaboration within psychometric models. Results presented here are promising, and indicate that assessment of collaborative performances may open new vistas in educational assessment.

References


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.
Author 1 is Professor, Graduate School of Education, University of California, Berkeley, Berkeley, CA, 94720; MarkW@berkeley.edu. His primary research interests are in psychometrics and educational assessment.

Author 2 is Ph.D., Graduate School of Education, University of California, Berkeley, Berkeley, CA, 94720; perman@berkeley.edu. His primary research interests are in psychometrics and statistics.

Author 3 is Associate Professor, Department of Educational Methodology, Policy and Leadership, University of Oregon, Eugene, OR, 97403; kscalise@uoregon.edu. Her primary research interests are in psychometrics, adaptive assessments, and innovative e-learning products.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Wilson, M; Gochyyev, P; Scalise, K

Title:
Modeling Data From Collaborative Assessments: Learning in Digital Interactive Social Networks

Date:
2017-03-01

Citation:

Persistent Link:
http://hdl.handle.net/11343/292561