Listening. Learning. Leading.®

Research Report
ETS RR-13-41

Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations

Alina A. von Davier

Peter F. Halpin

December 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations

Alina A. von Davier
Educational Testing Service, Princeton, New Jersey

Peter F. Halpin
Department of Humanities and Social Sciences, NYU Steinhardt, New York

December 2013

Action Editor: John Sabatini

Reviewers: Madeleine Keehner and Tenaha O'Reilly

Abstract

Collaboration is generally recognized as a core competency of today's knowledge economy and has taken a central role in recent theoretical and technological developments in education research. Yet, the methodology for assessing the learning benefits of collaboration continues to rely on educational tests designed for isolated individuals. Thus, what *counts* as evidence of learning does not correspond to current best practices for teaching, and it does not reflect what students are ultimately expected to be able to do with their knowledge. The goals of this paper are to give an overview of the research conducted in several fields of work related to collaboration, propose *a framework for the assessment of cognitive skills* (*such as science or math*) *through collaborative problem-solving tasks*, and propose several statistical approaches to model the data collected from collaborative interactions. This research contributes to the knowledge needed to support a new generation of assessments based on collaboration.

Key words: cognitive skills, collaboration, problem solving

Table of Contents

Collaboration is generally recognized as a core competency of today's knowledge economy and has taken a central role in recent theoretical and technological developments in education research. Yet, the methodology for assessing the learning benefits of collaboration continues to rely on educational tests designed for isolated individuals. Thus, what *counts* as evidence of learning does not correspond to current best practices for teaching, and it does not reflect what students are ultimately expected to be able to accomplish with their knowledge. Although estimation of individuals' cognitive skills may be most optimal when accomplished through traditional tests, the question remains about individuals' performance in a team, which is often required in school and workforce situations. Is person A more productive and successful when working alone or when working with others? What is A's contribution to the team's outcomes? Can we assess that individual contribution? Should we report a separate score for individual ability estimated in isolation and individual ability estimated from collaborative tasks? These are the kinds of questions we are addressing in this paper.

Specifically, the purpose of this paper is to review the research literature on collaboration in various disciplines, to propose a framework for an educational *assessment of cognitive* skills (such as science or math skills) with collaborative problem-solving (CPS) tasks, and to outline a novel psychometric approach for assessing *individuals' cognitive skills* using tasks that involve CPS.

Hence, this paper is not about directly assessing the construct of collaboration skills as measured in the CPS tasks, as others in the literature have attempted to do (e.g., Griffin, McGaw, & Care, 2012; Organization for Economic Co-operation and Development [OECD], 2013). Nor is it about directly assessing the construct of collaboration skills more generally. Those other approaches posit that noncognitive skills (e.g., teamwork) are construct-relevant aspects of the target skills assessed. Rather, we are interested in extracting evidence of individual cognitive skills (e.g., science or math skills) when students are engaged in CPS tasks, not the extent to which they possess any noncognitive skill sets.

One of the reasons we are not focusing on the collaboration skills is that they are not clearly defined in the literature. In his paper *What Do You Mean by "Collaborative Learning"?* Dillenbourg (1999b) wrote,

> The reader will not be surprised to learn that our group did not agree on any definition of collaborative learning. We did not even try. There is such a wide variety of uses of this term inside each academic field, and a fortiori, across the fields. (p. 1)

As psychometricians, we believe that an assessment should be built around a clearly defined construct, and, therefore, this paper is about measuring clearly defined cognitive skills through CPS tasks that pose significant measurement challenges.

That being said, we posit that the statistical methods we propose in this paper will be applicable to measuring specific collaborative skills that will unequivocally be defined in the future, using observable features from the collaborative interaction such as turn taking in communication, sharing resources and ideas, refraining from interrupting other team members, absence of *social loafing*, absence of abusive language, and so forth, See Liu, von Davier, Hao, Kyllonen, and River-Zapata (2014) for yet another attempt to define, quantify, and measure collaboration skills in CPS tasks.

The concepts, methods, and research ideas presented here are located at the intersection of collaborative learning, educational data mining, and psychometrics. We feel that this kind of multidisciplinary research has the potential to redefine the nature of standardized assessment. One of the main criticisms of standardized tests is that test writing or test taking is not an activity that is inherently useful to individuals. On the other hand, collaboration is regarded as a core competency, essential in today's workplace. Being able to assess individuals in a collaborative context, while also retaining the virtues of reliable, valid, and fair evaluation of their knowledge, represents a major advance in the culture and practice of educational assessment. While this type of assessment may have seemed implausible in the past, existing educational technologies, such as cognitive and peer tutors, multiuser virtual environments for learning, and online classrooms, are rich data sources that could provide the foundations of many future education assessments.

In this paper, we focus on measuring individuals' performances in terms of their *cognitive skills* (e.g., science, math) using CPS tasks. There are three characteristics that make our approach to collaboration innovative. First, most existing research on teamwork has focused on the outcomes of the entire group (e.g., Cohen, Lotan, Scarloss, & Arellano, 1999; O'Neil, 1997; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Any individual assessments are conducted externally to the collaborative tasks (e.g., pretests and posttests). Such approaches do not allow us to say anything about the knowledge that the individual group members

demonstrated during the task, as distinct from one another, because the dynamic interactions during the collaborations are not analyzed. In contrast to these approaches, our focus is on individual-level outcomes in the teamwork context.

Second, as mentioned above, we make a distinction between cognitive and noncognitive skills, with the former denoting the target we are addressing (i.e., individual ability) and the latter denoting those additional skills that might be involved when assessing collaborative skills directly. Frameworks of CPS that include noncognitive skills as central to the construct have been addressed elsewhere (e.g., Griffin et al., 2012; OECD, 2013). Although we will mention the role of noncognitive skills, our target is individual cognitive ability, such as science or math ability. In addition, we will not address the logistics of team assembly in this paper. This distinction and focus on the measurement of cognitive skills is the second innovative aspect of our approach to using CPS tasks.

The third ingredient that features prominently in the present research is education technology. In particular, the activity logs of educational software programs can provide detailed time series describing the actions and interactions of the users. We consider these activity logs to be an important ingredient for modeling and evaluating CPS tasks. Later, we provide a general characterization of what an ideal CPS data set would look like.

The central contribution of the present research is to outline a theoretical framework and preliminary statistical theory that can underlie advances in the assessment of cognitive skills with CPS tasks. The need for a statistical theory of CPS as a data collection design for assessment is apparent. The main priority in teamwork research was recently identified as the analysis of dynamic knowledge—how team members communicate and update each other on goal-related activity during the performance of a task and how this relates to the success of team activities (Cooke, Duchon, Gorman, Keyton, & Miller, 2012; Wildman et al., 2012). Quellmalz, Timms, and Schneider (2009) argued that traditional psychometric methods are not appropriate for accounting for collaborative interactions (CI) because of violations of core assumptions about statistical independence. In a discussion of educational multiuser virtual environments (MUVEs), Dede (2012) wrote that "understanding what students do and do not know as a result of open-ended learning activities requires new types of assessment design and analytic methods." In general, the lack of appropriate statistical methodology is currently a major hurdle in the understanding and evaluation of CI and their contribution to the learning process. The present

paper directly addresses this situation by suggesting such a methodology. In particular, we propose to incorporate traditional psychometric models into a stochastic process framework in which the overall dependency structure is parsimonious enough to be used with the CPS tasks. The core statistical models mentioned here are drawn from work by Halpin and De Boeck on the time series analysis of event data (Halpin, in press; Halpin & De Boeck, in press). Other dynamic statistical models may also be considered, and some of these alternatives are briefly mentioned.

The following sections address: (a) a review of the literature on collaboration and team performance; (b) a review of the data types that are typically obtained from CI and the features that would make an *ideal* data set; (c) a proposed assessment framework for cognitive skills that integrates current research in CPS with existing psychometric theory; (d) considerations on the development of tasks; (e) initial theoretical results on the statistical theory and specific models proposed; and (f) the significance and broader impacts of this research, how it is related to this new generation of assessments, and the research directions that need to be explored. We outline why it is relevant to psychometric theory and educational assessment to construct assessments that include CPS tasks.

Literature Review

The goal of this section is to summarize the existing literature related to CPS. We do this to provide a lay of the land and to stake out our general position. We do not attempt an exhaustive overview. Additionally, since this review takes us into many domains in which we are not resident experts, we may characterize these domains in a way that expert readers might find less than ideal. Nonetheless, we hope that by attempting to locate the present research within a broader perspective, we can facilitate the ultimate goal of multidisciplinary research into innovative educational assessments.

While there has been relatively little research on the assessment of cognitive skills through the utilization of CPS tasks, there is a very large body of literature pertaining to the nature and usefulness of teamwork, especially in the workplace. The study of collaboration in educational settings has not become prominent until somewhat more recently, and this has been partially intertwined with the widespread use of educational technology in the classroom (see Hmelo-Silver, Chinn, O'Donnell, & Chan, 2013). Although much of the educational technology literature is not directed at CPS per se, it places CPS within a broader frame of technologically intensive educational applications.

Within our own domain of educational assessment, there has been a strong recent interest in the evaluation of CPS as a social skill (de Jong, 2012; Griffin et al., 2012; OECD, 2013). However, to date, there is no statistical theory for the assessment of individuals' cognitive skills in the collaborative context. Because cognitive skills are not currently assessed using CPS tasks, the evaluation of learning outcomes attributable to CPS continues to rely on traditional methods of assessment, such as, multiple-choice examinations administered in a pretest and posttest design (Cooper, Cox, Nammouz, Case, & Stevens, 2008; Crouch & Mazur, 2001; Deslauriers, Schelew, & Wieman, 2011; Dillenbourg, Järvelä, & Fischer, 2009; Gilles & Adrian, 2003; Kirschner, Pass, Kirschner, & Janssen, 2011; Koedinger, McLaughlin, & Heffernan, 2010; Kolloffel, Eysink, & Jong, 2011; Noroozi, Weinberger, Biemans, Teasley, & Mulder, 2012; Prinsen,Terwel, Zijlstra, & Volman, 2013; Sandi-Urena, Cooper, & Stevens, 2012). This approach to evaluation ignores the wealth of data made available by the collaborative activities themselves, essentially throwing away copious amounts of information that can provide the basis for a much more detailed analysis of what students know and what they can do with their knowledge.

The first part of the literature review addresses conceptual features of collaborative interactions, specifically the interdependence among the collaborators and the type of skills needed to succeed in a group. Then we focus on technology-facilitated learning and interaction. In the last part of the section, we describe several data sets and outline the desired data properties for the assessment of cognitive skills through CPS tasks.

A Taxonomy of Collaborative Interactions Based on Interdependence

A good starting point is the definition of collaboration, and a useful concept here is interdependence (e.g., Thompson, 1967). The concept is often illustrated by contrasting interdependent teams or ensembles (e.g., a football team, a marching band) and independent teams (e.g., a gymnastics team, a chess team). The nature of the distinction is that the actions of members of interdependent teams rely on the actions of their teammates, whereas the members of independent teams operate in relative isolation. The focus of the teamwork literature has been on teams that are interdependent, and many gradients, categories, and sources of interdependence have been identified (Wageman, 2001). More radical approaches have taken group activities as the foundational unit of analysis, rather than the individuals who perform them, leading the notion of an ensemble (Granott, 1998).

We propose to distinguish at least three types of teams that seem to fall on a continuum of interdependence.

Ensemble. An ensemble consists of complementary parts that contribute to a single effect. An ensemble is a team whose activities cannot be performed by its members working in isolation. For example, if, instead of a usual soccer game, we put each of the 22 soccer players on their own field and they each performed exactly how they would in a game together, their individual activities do not make a soccer game. They make 22 different activities, none of which is a soccer game. A musical ensemble is another good example. Stevens et al. (2012) worked on neurosynchronicity of military teams, which are good examples of ensembles.

Group. A group consists of a number of individuals considered together because of similarities. In a group, individuals perform a task together, but not a task that intrinsically requires more than one person to be performed. For example, a gymnastics team requires multiple members because each person has different strengths, not because one person cannot, in principle, perform all of the necessary tasks. The gymnastics team has multiple people because it is trying to optimize its skill level, not because it needs more than one person. Another example is a sales team; a company may have too many clients for a single sales agent, so different agents manage different clients. There seem to be two main reasons to have a group rather than an individual: (a) time pressure and (b) a division of labor/task specialization. Most groups probably involve both of these.

Synchronized individuals. We use the notion of synchronized individuals to describe cases where persons operate in isolation, but on a shared timeline. Some types of teams (or parts of a team) only require people to do different tasks at the same time, for example, to defend the castle in a video game. Other examples of synchronized individuals are cashiers at a retail store or clerks in a law firm. This type of teamwork comprises individuals' work done at the same time or at overlapping times. This means that the individuals need to know the correct starting time, but they do not need to communicate among one another about their activities during a task. To be more precise, the team members' activities are statistically independent: my activities in the team could be done by someone who is equally good at the task but is not aware that the rest of the team exists, and this would not make a difference to team performance. No CPS is required for synchronized individuals. If each person does his or her individual task adequately, then the team succeeds. The only temporal dependency is the starting time.

In summary, ensembles require the activities of more than one person for task performance; groups incorporate more than one individual to optimize task performance with respect to skill specialization or time pressure; and synchronized individuals operate in isolation but are coordinated by a shared clock.

A central feature of the statistical approach we develop below is to measure the degree of team interdependence in terms of the statistical dependence between the activities of the team members. For synchronized individuals, the statistical dependence is zero. For individuals in ensembles, the dependence is (supposed to be) perfect. We can measure the in-between values using various techniques, and below we prefer the use of mutual information. Taking this approach, we are assuming that collaboration involves different people doing different things and that those things *depend on each other*. It is the statistical dependence of the timing and sequencing of the individuals' activities that describes the collaboration. If the timing and sequencing of the activities of two or more people is independent, then those individuals are not considered a team, but merely synchronized individuals.

Individual and Team Characteristics That Affect Performance

In addition to types of teams, many considerations have been made about the nature of team problem solving and the role of individuals therein. Perhaps the most notorious of these is Janis's (1982) notion of *groupthink*. More recent approaches have argued that there is a general group intelligence factor that is not highly correlated with the intelligence of the individual members but is stable over various tasks (Woolley et al., 2010). They have studied group level or *collective* intelligence as a factor separate from individual intelligence. This research involved a variety of team-based activities, but did not focus on recording group interactions. Rather, the focus was on group outcomes and evaluating the cognitive skills of the group as a whole, not of its individual members. There have also been many considerations about group decision-making processes (e.g., Tindale, Kameda, & Hinsz, 2003) and the individual noncognitive skills that lead to successful teamwork (e.g., DeChurch & Mesmer-Magnus, 2010; Griffin et al., 2012; OECD 2013; O'Neil, 1997). Soller and Stevens (2007) used dynamic models (hidden Markov models and Bayesian networks) to describe CPS activities, such as the kinds of problem-solving strategies employed and how participants shared knowledge. This research taps into the minutiae of group interactions but, again, the assessment component was external. In an intelligent tutoring system (ITS), student modeling is often explicit (e.g., Nkambou, Mizoguchi, &

Bourdeau, 2010) but is not intended as a means of reliably comparing students. Instead, it provides a means of targeting the tutor toward the needs of the individual student. This research is significant for understanding the factors that lead to better and worse team performance. However, the present research is focused on a different question: how to measure individual cognitive skills based on team performance.

Of course, team performance will be influenced by noncognitive individual skills and team composition factors, but our general approach is to treat these as nuisance variables. This may seem to omit from consideration many of the more interesting aspects of collaboration and teamwork, but this omission is, in fact, what makes our approach characteristically psychometric. Here, it is useful to compare our approach to that taken in traditional assessments. In traditional assessments, we do not measure the many noncognitive skills that play a role in test performance (e.g., study habits, test anxiety, time management). Rather, we assume that the systematic contribution of these factors to test performance is captured by conditioning on individual ability. To the extent that this assumption is false, the assessment is subject to revision, not the assumption. This is because the purpose of an assessment is to measure ability and explicitly not to measure extraneous factors—that is just what is meant by *reliable* measurement. Similarly, traditional assessments do not model the role of gender, racial and ethnic background, or social class. Instead, we assume that, after controlling for ability, the assessment performs equivalently in these different groups. Again, if the assumption is wrong, we change the assessment, not the assumption. This is because equivalent test performance in different social groups is what it means for an assessment to be *fair*. Taking this perspective, it is clear that we are not making an unfortunate oversight by eschewing the more interesting aspects of collaboration. Instead, we are taking an approach that allows for the treatment of collaboration as a modality in which individuals' cognitive skills can be validly measured. It seems to us that such a theory of measurement must be in place before we can meaningfully talk about the performance of individuals in the team context. By analogy, if we could not measure ability in a conventional assessment setting, we would not be able to answer questions about, for example, how study habits affect test performance. We expect that answering the psychometric questions will, in the long run, provide for a better understanding of the noncognitive and team composition factors that affect real-world collaborations. However, the present research is focused on the immediate goal of measuring individual cognitive skills in a collaborative context.

Educational Settings

In this subsection, we selectively review research on collaboration in education settings with the purpose of indicating the kinds of collaboration-based approaches intended to improve student learning outcomes that have been proposed (Cohen et al., 1999; Crouch & Mazur, 2001; Dillenbourg, 1999a; Fischer, Kollar, Mandl, & Haake, 2007; Hmelo-Silver et al., 2013; Johnson, Johnson, & Smith, 2007; Stahl, Koschmann, & Suthers, 2006; Webb, Troper, & Fall, 1995). This research is usually classified under the umbrella term *collaborative learning* and, while the exact principles continue to evolve, it essentially means having students work in groups. Some, but not all, of these approaches are technologically intensive, although technological applications are our main focus here.

An initial version of computer-supported collaborative learning was to pair students with an ITS (Graesser, VanLehn, Rose, Jordan, & Harter, 2001; Koedinger & Corbett, 2006; Matsuda et al., 2010; Nkambou et al., 2010). In its most simplistic form, an ITS might provide a determined sequence of hints about stem-and-response style questions; in its most sophisticated form, it involves text-based dialog between one or more students and an *avatar* (anthropomorphic computer agent). Although computer-human interaction might not be regarded as collaboration per se, the advancing technology of computer agents makes the use of avatars a viable way to simulate collaboration, and this can offer researchers more control than is available with real human collaboration (Graesser & McDaniel, 2008; OECD, 2013).[1] Moreover, many modern variations on ITS technology explicitly incorporate multiple users via peer tutoring and other computer-supported collaborations (e.g., Hmelo-Silver et al., 2013; Rummel, Mullins, & Spada, 2012; Walker, Rummel, & Koedinger, 2009; Walker, Rummel, & Koedinger, 2011). Another technologically intensive approach has students *play* within a multiuser virtual environment, engaging with each other, computer agents, and the simulated environment (Dede, 2009; Metcalf, Kamarainen, Tutwiler, Grotzer, & Dede, 2011). This approach essentially turns multiuser video games into a collaborative learning activity and is, therefore, part of the more general trends in using games for learning (Domagk, Schwartz, & Plass, 2010; Villalta et al., 2011). Here we also consider massive online open courses (MOOCs), in which students participate in an overarching, online classroom involving video lectures, chat rooms, forums, automatically graded assignments with instantaneous student feedback, and other course materials.[2] From our perspective, MOOCs provide a framework for situating many types of

interrelated learning activities, of which student collaboration is currently an emerging component (e.g., Bergner & Pritchard, 2013).

In summary, there has been no shortage of innovation about how to provide students with novel and enriching collaborative learning environments. However, as previously noted, much research on the learning outcomes of collaboration has focused on external criteria (e.g., multiple-choice pretests and posttests). Less research has been concerned with (a) the contributions of individuals to the outcomes of a collaborative project or to (b) the processes of interaction that lead to those outcomes.

When task outcomes have been studied, the focus has usually been to describe the success of the entire group (Cohen et al., 1999; O'Neil, 1997; Woolley et al., 2010). It is important to note that this approach does not allow us to say anything about the knowledge that the individual group members demonstrated during the task, as distinct from one another. Some approaches have *reverse engineered* the collaborative processes that lead to different task outcomes (Baker, Corbett, Koedinger, & Wagner, 2004; Graesser, Jackson, & McDaniel, 2007; Soller & Stevens, 2008; Stevens, 2012; Stevens et al., 2012). This research has discovered, for example, sequences of interactions that are commonly associated with (un)successful task performance. Here it is important to emphasize that the students' actions are not always interpreted as cognitive in nature, but may rather concern their emotions or agreeableness; we return to this point below. When students' actions are interpreted in an explicitly cognitive manner, we are then able to say something about the knowledge they demonstrated during the performance of a task. Such an approach can be useful, for example, to identify learners who may be struggling (Soller & Stevens, 2008). However, this approach cannot provide the basis for assessment because it is an ad hoc description of the collaborative processes that are likely to be related to (un)successful task outcomes; it may provide useful feedback to students, but it is not sufficient for making summative decisions about their ability.

Much research focusing on the interactive processes that occur during collaborative learning have been highly qualitative and descriptive in nature (Ding, Li, Piccolo, & Kulm, 2007; Granott, 1998; White, Wallace, & Lai, 2012). We are not aware, however, of any quantitative modeling of teamwork process data. This is perhaps largely because such data has only become accessible with relatively recent advances in educational technology. As we mentioned above, applications of CPS do not, in principle, need to be technologically intensive. However, we have

highlighted such applications because they yield quantitative data. These data are typically recorded as a time-stamped sequence of events registered by a computer program during its use via an activity log. From a statistical perspective, these activity logs provide detailed time series describing the actions and interactions of the users. In the following section, we outline the general flavor of these activity logs and how we envision their use in developing assessments based on CPS tasks.

Data, Interdependence, and Dynamics

As mentioned above, the data from CPS consist of time-stamped sequences of events registered in a log file. From a statistical perspective, these activity logs or *logfiles* are detailed time series describing the actions and interactions of the users. We refer to this as process data. In addition, we also have outcome data, such as the correct/incorrect assessment of an action or task at the individual or team level. In this subsection, we discuss these types of data and their role in assessment. We also review the type of data structures encountered in the literature.

Process data. The process data offer an insight into the interactional dynamics of the team members, which is important both for defining collaborative tasks and for evaluating the results of the collaboration.

In a CPS assessment, the interactions will change over time and will involve time-lagged interrelationships. Thus, if there are two people on a team, the actions of one of them will depend both on the actions of the other and on his or her own past actions. The statistical models used should accurately describe the dynamics of these interactions. These dynamics, which are defined by the interdependence between the individuals on the team, could also offer information that could be used to build a hypothesis about the strategy of the team. For example, by analyzing the covariance of the observed variables (the events), one might hypothesize that an unknown variable, such as the team's type of strategy, explains why the team chose a particular response and avoided the alternative.

Hence, the data of interest are the activity logs. An activity log or logfile is basically a time-stamped record of what a program does during its operation. In the context of learning technology, these time-stamped events can represent the minutiae of students' activities, right down to the level of where the pointer is located on the screen. This kind of information has so far been the grist of educational data mining (EDM; Baker & Yacef, 2009; Romero & Ventura, 2005), a domain which has been intimately connected to recent developments in educational

technology and especially ITS. However, the interests of EDM often depart from those of conventional psychometrics. In particular, while many new data sources are available, these have yet to be harnessed for the purpose of educational assessment.

Individual outcome data. The outcome data are collected through the evaluative scoring throughout the process (collaboration). For example, an individual's actions during the collaboration can be scored as *correct* or *incorrect* by a human rater or an automatic scoring engine. Pretests or posttests, if available, also result in individual outcome data. If either of these tests is available, then the test scores that contain information about the test-taker ability can be corroborated with the information contained in the actions scored throughout the CPS task.

Team outcome data. The team level outcome data are straightforward to collect. These data indicate whether a team solved the task successfully or whether parts of the problem were solved correctly.

Data examples. To indicate the general character of activity logs, we summarize three example data sets in Table 1. These datasets were obtained via personal communications with the indicated authors.

Table 1

Properties of Example Data Sets

Data source	Activity log	Avatar interaction	Human interaction	Evaluation	Pre-/post-testing
SimStudent	Yes	Yes	No	Possible	Yes
EcoMUVE	Yes	Yes	Possible	No	Yes
Dr. Bob	Yes	Yes	Yes	Yes	Yes

SimStudent is a simulated student that *learns* using principles of artificial intelligence (Matsuda et al., 2010, 2011). The data we examined involved a real middle school student teaching SimStudent how to do algebra via an avatar interface. The number of activities that can be treated as interactions between the two is quite large (e.g., requests for help, input from the real student on whether the SimStudent acted *correctly*). However, the activity logs currently recorded by SimStudent allow very few of these interactions to be treated in an evaluative way (i.e., as actually correct, not merely correct in the view of the student). This decision is quite natural when one considers that SimStudent was not designed for assessment but as a learning interface.

EcoMUVE (Dede, 2009; Metcalf et al., 2011) is a multiuser virtual environment for ecosystem education that is directed at high school and middle school students. Students can log into different worlds and navigate around these worlds, visiting different locations and taking chemical measurements of the environment.

The interactions students have with the EcoMUVE avatars are usually quite short and deterministic (e.g., clicking on an animal provides a description of the animal). Multiple students can be logged in at the same time and can communicate with one another via chat, but in the current implementation there is no record of to whom each student is speaking. In its present formulation, EcoMUVE is not evaluative. Students are not required to perform specific tasks or missions, and there is no *end* to each world. As with SimStudent, the current implementation focuses on learning, not assessment.

Dr. Bob is the name of a computer avatar from another system who mediates chat room interactions between two engineering students (Adamson & Rosé, 2012; Howley et al., 2012). Each student is charged with the task of designing a power plant; one student's power plant is eco-friendly and the other's is designed to maximize energy output. The conversation is guided by Dr. Bob, who, among other things, uses a natural-language parsing algorithm to evaluate whether students' responses to certain knowledge questions are correct. The activity logs of this program are for the raw chat data only, with Dr. Bob's internal functioning controlled by a black box and the final power plant designs completed through external software.

These examples indicate the general flavor of the kinds of data currently available via ITS learning technology. In our experience, ITS researchers have not been interested (to date) in using this technology for assessment, which is evidenced by the fact that their data records do not include the possibility for evaluating the correctness of students' activities. However, the potential for evaluation is clearly there and can be achieved by relatively straightforward modifications to the task designs and the recorded information. Pretest and posttest data for each student in the content area would be useful to have both for measurement and validation purposes.

In summary, we feel that most of the ingredients of the research proposed below have appeared in various guises and in various literature. Our next goal is to formulate these ingredients into a CPS assessment framework.

A Conceptual Framework for an Assessment With Collaborative Interactions

CPS requires that individuals work together to complete a complex task. On the other hand, traditional psychometric methods have been most successful with isolated individuals responding to a single test question at a time. While there have been many advances in the theory and practice of CPS as a tool for learning, there has been much less research on the evaluation of individuals' cognitive contributions to the outcomes of specific CPS tasks and the interactive processes that lead to those outcomes.

In addition to cognitive skills, one also needs to address the role of noncognitive skills in CPS. Here we face topics like leadership, communication, and the nature of collaboration itself. An important contrast between our approach and that proposed by Griffin et al. (2012) is that we do not propose to measure noncognitive skills. As discussed above, psychometricians traditionally have not been concerned with incorporating the noncognitive skills involved in successful test writing into the measurement model—for example, test-taking strategies, study habits, or student motivation are studied separately. This is because any assessment modality involves a plethora of auxiliary skills that are not necessarily a direct manifestation of the knowledge domain that one wishes to measure. This issue is closely related to the role of confounding variables in assessment, such as language ability, culture/country of origin, or gender. These also have not been traditionally included in psychometric models. Rather, the effects of these variables are disentangled from the target construct through data collection design and post-administration analysis (e.g., factor analysis, differential item functioning, equating invariance), and confounded tasks are removed from circulation. There is a large amount of literature on how to deal with confounding in individual-level testing (e.g., see von Davier & Oliveri, 2013).

Figure 1 summarizes how CPS fits into our overall conceptual assessment framework. Starting at the top of the figure, the first tier is the cognitive skill to be measured, which is the hypothesized target of any educational assessment. The second tier denotes the types of activities that count as evidence of the cognitive skill; these are the focus of the next section. The third tier makes explicit the modality of assessment, and we have taken the position that CPS is on par with, or an alternative to, individual-level procedures. The bottom tier lists some specific assessment strategies, and, naturally, the strategies for collaborative assessments are not yet as well developed as those of the individual case. The final component of the figure represents

noncognitive skills and potential confounding variables that can affect any and all levels of the assessment hierarchy. As described above, these are not explicitly modeled but are dealt with through auxiliary procedures.

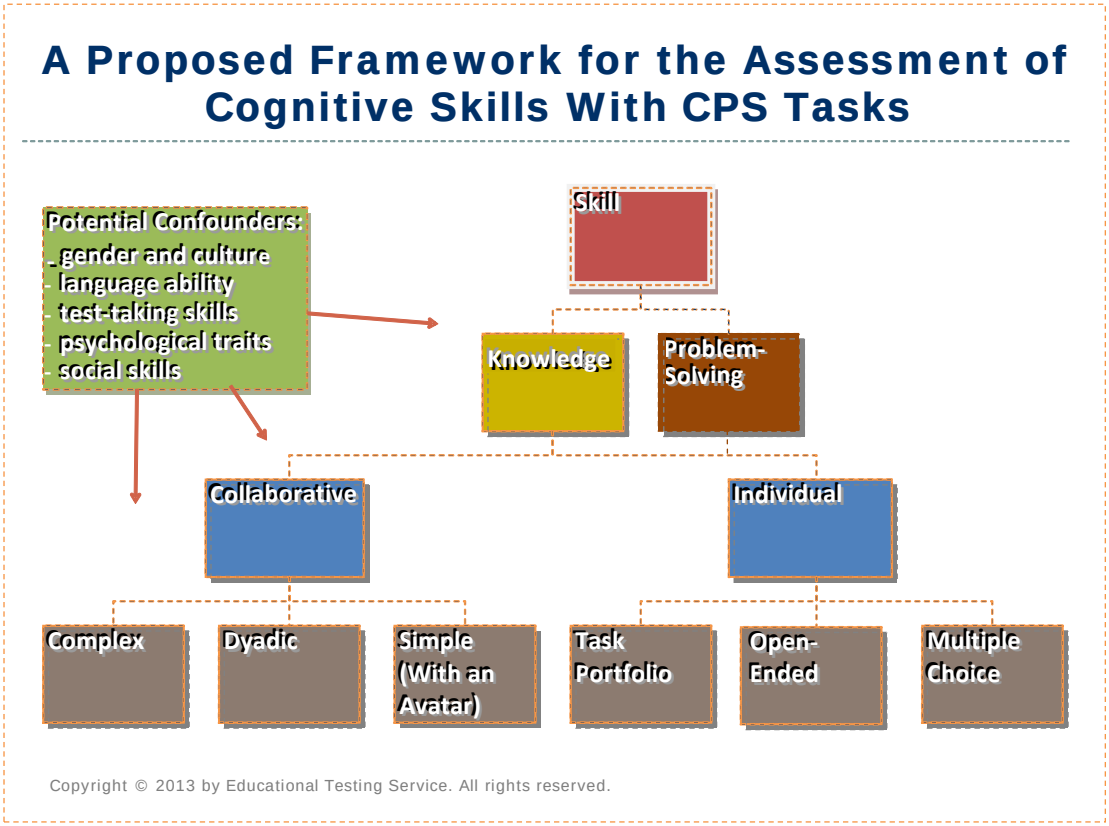

***Figure 1*. Conceptual assessment framework using collaborative problem-solving tasks (see text for details). The layers are the skill, assessment target, assessment modality, evidence activities.**

Previous research has suggested both that people behave differently when they interact in teams than when they work alone, and that the team members' individual domain scores might not correlate highly with the team's outcome (e.g., Woolley et al., 2010). The latter is a very interesting hypothesis and one that could be investigated closely using the framework for the assessment mapped out in Figure 1. By assessing the differences between the individual problem solving and team CPS, one could generalize several scores instead of the usual single total test score. In this situation, we have a score obtained in isolation, a score obtained in collaboration, and the team score as described in the previous section.

Considerations for Collaborative Problem-Solving Task Construction

This subsection briefly outlines some psychometric aspects of CPS task development. The overall framework for task development should be based on the evidence-centered assessment design (Mislevy, Steinberg, & Almond, 1999). The basic principles of this framework require defining in advance the evidence needed to support the claim that mastery of a particular skill has been demonstrated. This information is used to guide the construction of tasks that provide the requisite evidence. Therefore, task construction should draw on the existing literature for structuring collaborative activities (Cohen et al., 1999; Dillenbourg, 1999a; Rummel et al., 2012; Stahl, 2009; Walker et al., 2011; Webb et al., 1995; White et al., 2012).

Because of the many options available for the design of CPS tasks, research should be devoted to developing one or more useful formats. At this point, we have identified the following four basic factors that need to be considered. First is the degree of structure imposed on the task. In general, it is necessary to have a fixed bank of elements that comprise the overall task. It must then be decided whether students must complete the task on their own, or whether they are guided or prompted to follow a sequence of moves, or something in between. Second, rules for turn taking may be considered, for example, by a random or fixed order. Third, the kinds of actions that students can perform on each turn must be predetermined. Can students undo their actions from previous turns? Can students skip a turn? Do all students need to indicate their agreement on a specific action at each step? The fourth consideration is task termination, for instance, when further responses become uninformative about students' ability levels, when arriving at a correct answer/solution becomes impossible, or when all students agree to move ahead.

After a set of tasks has been developed, an initial round of data must be collected so that the tasks can be calibrated or *normed*. Roughly, this means that we need to know how difficult the tasks are for the population in which they will be administered. More specifically, it means that we need to estimate the parameters of the models proposed in the following section. Based on the parameter estimates, we can then select a subset of CPS tasks that permit reliable estimation of the range of cognitive ability, such as science or mathematics ability, in the target population. In addition to calibrating the CPS tasks, we also recommend the calibration of an external measure of the cognitive ability as illustrated in Figure 1; a set of traditional multiple-

choice format items can be used as the external measure. We also recommend administering a brief background survey for the purposes of identifying potential confounding variables.

Statistical Models for Collaborative Problem Solving

There are a number of modeling strategies available for use with process data. These do not come from educational assessment, however, so they must be adapted from other fields. Some modeling strategies include dynamic factor analysis; multilevel modeling; dynamic linear models; differential equation models; nonparametric exploratory models, such as social network analysis; intra-variability models; hidden Markov models; Bayesian belief networks (BBNs); Bayesian knowledge tracing (BKT); machine learning methods; latent class analysis; neural networks; and point processes.

From the perspective of psychometric theory, BBNs are by far the most useful means that have been used so far for modeling student knowledge that is demonstrated during a complex task (Russell & Norvig, 2003). BBNs model the probability that a student has mastered a specific knowledge component conditional on the sequence of responses given to previous elements of a task. BBNs have long been applied in ITS to represent student knowledge and, thereby, guide the activities of the tutoring system (Corbett & Anderson, 1995; Desmarais & Baker, 2012; VanLehn, 2008). They have also played a central role in the design of complex assessments (Mislevy et al., 2002; Shute, Hansen, & Almond, 2007; VanLehn & Martin, 1998; Vomlel, 2004); therefore, BBNs are an obvious methodological bridge between CPS and traditional psychometric theory. However, the practical implementation of BBNs often requires highly simplifying assumptions, and, as with traditional models, they have not been adapted to represent the knowledge of multiple individuals simultaneously.

One of the main contributions of the present research is to generalize existing models of individual cognitive skills to incorporate the collective actions of a group of collaborators. In the following section, we develop a point process framework for measuring individual cognitive skills through CPS. We indicate how this approach falls in line with existing psychometric research, and in future research we will more explicitly consider how this approach can be integrated with existing models.

A Statistical Representation of Collaboration

It is intuitive to model collaboration in terms of statistical dependence among the activities of two or more individuals. This approach also fits nicely with traditional theories of teamwork that distinguish between independent and interdependent teams (Thompson, 1967), as discussed earlier. While there have been many conceptualizations of interdependent teams (Wageman, 2001, provides a review), our intention is to use this idea as a building block for a statistical theory of collaboration. In this section, we present some initial work along those lines.

Begin by letting $X_j = \{X_{jt}: t \in T\}$ be a sequence of random variables that describes the activities of individual $j \in \{1, \ldots, J\}$ over time. Most simply, X_{jt} could be a dichotomous variable indicating whether individual j has performed an activity at time index t. Let $p(X_j)$ denote the marginal probability of the time series of individual j, and let $p(X_1, \ldots, X_J)$ denote the joint probability of the time series of all individuals. We propose to measure the degree of interdependence demonstrated by the activities of the J individuals using the Kullback-Leibler divergence of the marginal distributions from the joint distribution:

$$I_j = E_{X_1,\ldots,X_J}\left(\ln \frac{p(X_1, \ \ldots, X_J)}{p(X_1)\cdots p(X_J)}\right). \tag{1}$$

When $j = 2$, I represents the mutual information of two stochastic processes. While other measures of multivariate dependence can also serve our intended purpose, Kullback-Leibler divergence is a theoretically powerful quantity with many well-known results (Cover & Thomas, 1991), and it is also useful as a data analytic device (Brillinger, 2004). The interpretation of I in the context of collaboration is intuitive. $I = 0$ describes an independent team. When $I > 0$, some interdependence is exhibited among the activities of the J individuals, with larger values indicating more interdependence (i.e., a greater divergence from the model of independence). Importantly, we cannot draw conclusions about the nature of the interdependence based on a value of $I > 0$. For example, we do not know which individuals are responsible for the interdependence, or whether the interdependence is useful to the team's goals. However, we can use this basic definition to formulate implications for both the processes and the outcomes of a CPS task that would be expected under the hypothesis of productive collaboration. We make this idea concrete with the following two illustrations.

Example 1: Process data. Define the time series $U_j = \{X_{js}, s \in S\}$ by applying the lag $s = t - u$ to X_j. Here it is assumed that $u \geq 0$ is constant and $S \subset T$ such that for all $s\_ > s$, $s\_ \in S$ if $s \in S$. Then define $I_j(u)$ as the interdependence index obtained by applying lag u to the activities of individual j:

$$I_j(u) = E_{X_1,\ldots,U_j,\ldots,X_J}\left(\ln \frac{p(X_1, \ldots, U_j, \ldots, X_J)}{p(X_1)\cdots p(U_j)\cdots p(X_J)}\right). \tag{2}$$

$I_j(u)$ allows us to examine the change in interdependence that results from delaying the activities of individual *j* by a period of time *u*. Treating this as a function of *u* and assuming that the memory of the process does not extend past some lag *u**, then if $I_j(u^*) = I$, we can conclude that *I* does not depend on X_j . In other words, the activities of the team do not depend on those of individual *j*. For a team whose activities were dependent on those of individual *j*, we would, in general, expect to see $I_j(u)$ decrease in *u*, perhaps after some initial latency period. Further hypotheses can be formed about $I_j(u)$ and other lagged interdependence functions, and when utilized as a data analytic device, such functions also provide a means of learning about real-world collaborations.

In summary, a definition of interdependence should be useful in describing the contributions of each individual's activities to the ongoing process of collaboration. We still, however, have not said anything about the quality of an individual's actions (e.g., whether he or she helps or hinders the group). This requires a statistical model, and we discuss some candidates below.

Example 2: Outcome data. Define an outcome as a function $g = g(X_1, \ldots, X_J)$ of the complete time series. Most simply, if the activities recorded by X_{jt} are correct responses to the components of a CPS task, we could define the group's total score on the task as:

$$g = \sum_j \sum_t X_{jt}.$$

Then the expected value of g for an independent team is

$$E_{X_1,\ldots,X_j}(g) = \sum_j E_{X_j}(\sum_t X_{jt}). \quad (3)$$

Thus, for an independent team, the expected outcome is simply the sum of its parts. A productive collaboration can then be defined as one for which the expected outcome is greater than the sum of its parts, and an unproductive collaboration would have an expected performance worse than that of the independence model. A similar approach can be applied to other collaborative outcomes. Instead of sum scores, it will be generally advantageous to have a psychometric model for the entire response pattern X_{jt}, for instance, an item response theory (IRT) model or a BBN.

In conjunction with these models for ability, it is also important to consider response times. For example, at a fixed level of group ability, it is natural to prefer teams that perform faster than expected by their individual members operating independently. Individual-level models incorporating both ability and response time have been developed recently in the psychometric literature (Maris & van der Maas, 2012; van der Linden, 2007), and their generalization to the collaborative context presents many exciting opportunities. However, in the assessments with CPS tasks, where test takers have never worked together previously and where the tasks are not long enough to allow for familiarity with each other's style, the potential advantage of speed in solving the task may not be apparent. In summary, by treating the independence model as a reference point for group outcomes, it becomes straightforward to incorporate existing models from psychometric theory. These models define a baseline outcome against which to judge successful and unsuccessful collaboration. The overall result of this approach is to provide a framework that generalizes what we already know about psychometric assessment to the collaborative context.

Statistical Models

At this point, we have sketched some characteristics of a statistical theory of collaboration and made a distinction between the processes and outcomes of a collaboration. In modeling the processes of collaboration, we are concerned about describing the statistical dependence exhibited by the activities of groups of individuals and the roles of specific individuals therein. In modeling the outcomes of collaboration, we are concerned with judging the performance of a group relative to what we would expect from the individual group members

had they not collaborated. While these two topics can be dealt with productively in relative isolation, a major goal of this research is to provide statistical models that can incorporate both of these aspects. As we will outline in the next section, the theory of point processes is particularly well suited for this purpose.

A point process is a model for isolated events occurring in continuous time, with familiar examples including the Poisson and renewal processes. The more general theory of point processes (Daley & Vere-Jones, 2003) provides a framework for multivariate event series with long memory and random, time-varying covariates. The central component of this framework is the conditional intensity function (CIF), from which many known results concerning maximum likelihood estimation and goodness of fit are directly available. We develop this framework in enough detail to recommend an approach for this type of research.

We have previously presented a formalization of interdependence in terms of a dichotomous-valued, discrete-time stochastic process. In applications to CPS, this means we are concerned with how the probability of an individual's activities changes over continuous time as a function of previous activities. The passage to point processes can be made assuming the time indices $t \in T$ are equally spaced with distance Δ over a fixed interval of time $[a, b)$ and letting $\#T \rightarrow \infty$. Then $X_{jt} = 1$ denotes the occurrence of an event in the infinitesimal interval $[t, t+\Delta)$.

The CIF of a point process, $\lambda(t)$, is a mechanism for describing the expected rate of events occurring in the interval $[t, t + \Delta)$, conditional on all events that have occurred previous to time t:

$$\lambda(t) = \lim_{\Delta \to 0} \frac{E(M\{[t, t+\Delta)\} \mid H_t)}{\Delta}, \tag{4}$$

where $M\{[a, b)\}$ is a random counting measure representing the number of events (i.e., isolated points) falling in the interval $[a, b)$, $E(M\{[a, b)\})$ is the expected value, and H_t is the σ-algebra generated by the time points t_k, $k \in N$, occurring before time $t \in \mathbf{R}+$. It is convenient to assume that the probability of multiple events occurring simultaneously is negligible, in which case M is said to be orderly. Then for fixed t and sufficiently small values of Δ, $\lambda(t)\,\Delta$ is an approximation to the Bernoulli probability of an event occurring in the interval $[t, t + \Delta)$, conditional on all of the events happening before time t.

Point processes extend immediately to the multivariate case. $M\{[a, b)\}$ is then vector-valued, and each univariate margin gives the number of a different type of event occurring in the time period $[a, b)$. In the CPS context, the univariate margins correspond to the individuals $j \in \{1, \ldots, J\}$. Although the different individuals can also be encoded using covariates, we prefer a multidimensional approach because it explicitly relates to our definition of interdependence in Equation 1. Linear systems analysis provides a general framework for modeling $\lambda(t)$ [6], and, in particular, we assume that $\lambda(t)$ is a J-dimensional causal filter:

$$\lambda(t) = \mu + \int_0^t \phi(t-s)dM(s), \tag{5}$$

where $\mu > 0$ is a J-dimensional baseline, which can be a function of time but is treated here as a constant, and $\varphi(u)$ is a $J \times J$ matrix of impulse response functions that govern how each margin of $\lambda(t)$ depends on the past of the multivariate process. The diagonal elements of $\varphi(u)$, denoted $\varphi_{jj}(u)$, describe how the j-th margin depends on its own past. The off-diagonal elements, $\varphi_{jk}(u)$, $j \neq k$, describe how the j-th margin depends on the past of the k-th margin. When $\varphi(u)$ is diagonal, this corresponds to the independence model described previously. In particular, the general form of Equation 1 for a point process with CIF given by Equation 5 is:

$$I = E_{M_1,\ldots,M_J}\left(\int_0^T \ln\frac{\prod_j \lambda_j(t)}{\prod_j \gamma_j(t)}dt\right), \tag{6}$$

where

$$\gamma_j(t) = \mu_j + \int_0^t \phi_{jj}(t-s)dM(s), \tag{7}$$

$\lambda j(t)$ denotes the univariate margins of Equation 5, and the process is observed over the interval $[0, T]$.

Up to this point, we have focused on only the process aspect of collaboration. In order to introduce a component that relates to the outcomes, we need to be able to describe each event in relation to the overall CPS task. For this purpose, we introduce the marked point process. A mark Y is a random variable that describes additional information about an event, such as whether the

action was *correct* or *incorrect*. In psychometrics, the mark is called *score*. Letting $Y = y$ denote the realizations of the mark, the conditional intensity function of a marked point process is defined by

$$\lambda(t, y) = \lambda(t)\, p(y \mid t, H_t) \tag{8}$$

$\lambda(t)$ is defined as above and is called the *ground intensity*. $p\ (y \mid t, H_t)$ is a probability model for the marks, conditional both on the history of the process and the occurrence of an event at time t. It is important to note that H_t now includes the marks for time points previous to t.

In an application to CPS, the general idea is to build a measurement model for $p\ (y \mid t, Ht)$. A latent variable model, for instance, an IRT model, takes the form

$$p(y \mid t, H_t) = \int_{\Theta} p(y \mid \theta, t, H_t)\, p(\theta \mid t, H_t) dF(\theta), \tag{9}$$

where θ denotes the latent variable and F its distribution function. Alternatively or subsequently, the marking variable can be used to define a posterior probability for the latent variable:

$$p(\theta \mid y, t, H_t) \propto p(y \mid \theta, t, H_t)\, p(\theta \mid t, H_t). \tag{10}$$

Letting θ denote the mastery of a skill, we then enter the domain of BBNs. The main novelty introduced by this approach, in comparison with usual psychometric models, is that we condition not only on the ability of an individual, θ, but also on the event history of the entire process, H_t, which includes the actions of the other individuals in the CPS task. In intuitive terms, we reckon on the demonstrated ability of the entire group when estimating the latent ability of each individual.

The log-likelihood of a J-dimensional marked-point process is (see Daley & Vere-Jones, 2003):

$$l = \sum_j \sum_k \left(\ln \lambda_j(t_{jk}) + \ln p(y_{jk} \mid t_{jk}, H_{tjk}) - \Lambda_j(T) \right), \tag{11}$$

where

$$\Lambda_j(T) = \int_0^T \lambda_j(t)dt \cdot$$

Recent literature (Halpin, in press; Halpin & De Boeck, in press; Rasmussen, 2013; Veen & Schoenberg, 2008) has addressed the computational aspects of estimation for Hawkes processes (Hawkes, 1971; Hawkes & Oakes, 1974) using Equation 11. In particular, Halpin has written extensive code in R and C for the maximum likelihood estimation of Hawkes processes and found the model to be well conditioned and to yield good fit for a number of human interaction data sets, including two e-mail databases and a large database taken from the Twitter website.

The future research requires two things be developed beyond the existing literature on point processes. First is the incorporation of psychometric models for the marks, which would involve original mathematical formulations of modest difficulty. Second, it is desirable to find simplified parameterizations of the response functions that provide good fit and stable estimates for a relatively small number of events. In the special case that only the sequence of events is of interest and not the event times per se, then an easily implemented reduction of this framework would be to use the uniform distribution for the response kernels. In general, this second step is mostly a matter of customizing existing models to CPS data.

Conclusion

The research directions presented in this paper are located at the intersection of collaborative learning, educational data mining, and psychometrics. This paper has proposed a novel assessment framework and a novel statistical approach to assessment that we hope can provide the initial impetus for many future research developments in CPS. The more general idea behind this research is to develop psychometric theories that exploit data sources made available through new educational technologies, and to develop general principles and specific models for the next generation of data-intensive educational assessments.

Any new types of assessment will require considerations regarding how to satisfy traditional assessment requirements, such as reliability, validity, and comparability. To provide specific directions for future research into CPS assessments, we list several research questions in particular:

- What does it mean to have a reliable test that contains complex tasks? How can we define and elicit the right evidence from the process data?
- How long should a task be so that the process data are rich enough to allow for the intensive and dependent longitudinal models?
- How many tasks are needed for a reliable assessment?
- What is the best way to evaluate the validity of such an assessment? What type of data should be collected for a predictive type of study?
- How can one construct comparable complex problems that differ from one administration to the next? In other words, how can we rethink the notion of test equating?

Table 2 summarizes the type of research that has been done to address some of the issues discussed in this paper and identifies gaps in the knowledge necessary to build CPS assessments. It is easy to see that a comprehensive research agenda that will span several years is needed to fully operationalize a new type of assessment of cognitive skills through CPS tasks.

The development of an assessment should also require data collection for calibration and validation. The main purpose of this data collection is to investigate the quality of the proposed psychometric models and, in particular, to validate the estimates of cognitive ability obtained by CPS against those obtained through the concurrent use of multiple-choice items. A validation study would also allow for a large number of potential follow-up projects. Possible areas for follow-up include studying: the measurement invariance over potential confounders such as gender, culture/country, and language; the relative advantages of CPS versus individual-level assessments; methods of optimizing CPS task formats and data collections; improved models for CPS tasks scoring; the potential for generalization to other knowledge domains; and the nature of collaboration empirically. These are questions that will be the subject of research for many years to come.

Table 2

Summary of Findings and Research Recommendations

Assessment structure	CPS research	Research questions	Future research
Task development	• Rich research portfolio needed for collaborative tasks • No research exists on building tasks to assess cognitive skills through CPS	• What is the construct? • What evidence needs to be collected to demonstrate success?	• ECD, student and task models • Item creation
Team assembly	Rich research portfolio	How do you assemble successful teams?	Use linear programming to optimize the team
Logfiles	Rich research portfolio	How do you structure the file optimally?	Use machine learning tools and data mining to extract data
Scoring	Some research exists (O'Neil, 1997; Soller & Stevens, 2007; Woolley et al., 2010)	How do you score the data captured from a CPS task?	• Correct/reaction time scoring of logs • Bayes nets for getting a posterior estimate of individual ability given a group score (limited number of groups) prior skill knowledge • Value added for estimating a person's average contribution to a group, only group score (multiple groups) but no prior skill required
Validity	Some research (O'Neil, 1997)	• What are the effects of team assignment? • What is good evidence for the external predictive validity of a test with CPS?	• Validity of the ECD (programming an avatar to pass the Turing test; putting real test takers in a group scenario) • Investigate different team assignments • Transfer of CPS skills from a specific task to real-life situations
Reliability	Not much	What type and how many CPS tasks are needed for a reliable score?	• How many interactions are needed depends on the SE of the parameters we want to estimate

Assessment structure	CPS research	Research questions	Future research
Fairness	Some research on gender (Mazur, 2002)	How do you define fairness?	• Study team assignment • Compare scores by method with multidimensional scores and with augmented scores • Subgroups (as defined by various background variables) impact
Comparability of test forms	None		• Comparability of tasks across forms • Pre-equating/calibration of tasks prior to the assessment • Adaptive CPS tasks • Estimation of CPS task difficulty
Measurement models	Some research (O'Neil, 1997; Soller & Stevens, 2007)	• How do you build psychometric models for the process and outcome data? • Should the individual scores from the CPS and from the multiple-choice test be augmented?	• Psychometric models for interactions • Models that match the ECD framework • Successive separate models vs. a concurrent complex model • Data mining vs. theoretical models • Model fit and validity of predictions
Logistics/ administration	Some research on avatar interactions	• What is the impact of the test modality? • How do you ensure test security?	• Effects of testing mode (ITS vs. games vs. simulations) • Effects of team assignment (real people vs. avatars)

Note. CPS = collaborative problem solving; ECD = evidence-centered design; ITS = intelligent tutoring system, SE = standard error.

References

Adamson, D., & Rosé, C. P. (2012). Coordinating multi-dimensional support in conversational agents. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Lecture notes in computer science: Vol. 7315. 11th international conference, ITS 2012* (pp. 346–351). Heidelberg, Germany: Springer-Verlag.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students game the system. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of ACM CHI 2004: Computer-human interaction* (pp. 383–390). New York, NY: ACM Press.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009. *Journal of Educational Data Mining, 1*, 3–17.

Bergner, Y., & Pritchard, D. E. (2013). *Homework collaboration via discussion boards in a massive open online course*. Paper presented at an invited symposium at the international meeting of the Psychometric Society, Arnhem, Netherlands.

Brillinger, D. R. (2004). Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics, 18*,163–183.

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. *Theory Into Practice, 38*(2), 80–86.

Cooke, N. J., Duchon, A., Gorman, J. C., Keyton, J. J., & Miller, A. (2012). Preface to the special section on methods for the analysis of communication. *Human Factors, 54*(4), 485–488.

Cooper, M. M., Cox, C. T., Nammouz, M., Case, E., & Stevens, R. (2008). An assessment of the effect of collaborative groups on students' problem-solving strategies and abilities. *Journal of Chemical Education, 85*(6), 866. doi: 10.1021/ed085p866

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics Teachers, 69*(9), 970–977.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes: Elementary theory and methods* (Vol. 1, 2nd ed.). New York, NY: Springer.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork. A meta-analysis. *Journal of Applied Psychology, 95,* 32–53.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66–69.

Dede, C. (2012, May). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights.* Paper presented at the invitational research symposium on technology enhanced assessments, Washington, DC. Retrieved from http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

de Jong, J. H. A. L. (2012). *Framework for PISA 2015: What 15-year-olds should be able to do*. Paper presented at the 4th annual conference of the Educational Research Center.

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

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22,* 9–38.

Dillenbourg, P. (Ed.). (1999a). *Collaborative learning: Cognitive and computation approaches* (2nd ed.). Bingley, UK: Emerald Group Publishing Limited.

Dillenbourg, P. (1999b). What do you mean by "collaborative learning"? In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (Vol. 1, pp. 1–19). Bingley, UK: Emerald Group Publishing Limited.

Dillenbourg, P. J., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer supported collaborative learning. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-enhanced learning* (pp. 3–19). The Netherlands: Springer.

Ding, M., Li, X., Piccolo, D., & Kulm, G. (2007). Teacher interventions in cooperative-learning mathematics classes. *The Journal of Educational Research, 100*(3), 162–175.

Domagk, S., Schwartz, R., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior, 26,* 1024–1033. doi:10.1016/j.chb.2010.03.003

Fischer, F., Kollar, I., Mandl, H., & Haake, J. (Eds.). (2007). *Scripting computer-supported collaborative learning: Cognitive, computational, and educational perspectives*. New York, NY: Springer.

Gilles, R. M., & Adrian, F. (2003). *Cooperative learning: The social and intellectual outcomes of learning in groups*. London, UK: Farmer Press.

Graesser, A. C., Jackson, G. T., & McDaniel, B. (2007). Autotutor holds conversations with learners that are responsive to their cognitive and emotional states. *Educational Technology*, *47*, 19–22.

Graesser, A. C., & McDaniel, B. (2008). Conversational agents can provide formative assessment, constructive learning, and adaptive instruction. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 85–112). New York, NY: Routledge.

Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39–51.

Granott, N. (1998). Unit of analysis in transit: From the individual's knowledge to the ensemble process. *Mind, Culture, and Activity: An International Journal, 1*(1), 42–66.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer.

Halpin, P. F. (in press). An EM algorithm for Hawkes process. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th annual Psychometric Society meeting*. New York, NY: Springer.

Halpin, P. F., & De Boeck, P. (in press). Modeling dyadic interaction with Hawkes process. *Psychometrika*.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90. doi: 10.1093/biomet/58.1.83

Hawkes, A. G., & Oakes, D. (1974). A cluster representation of a self-exciting process. *Journal of Applied Probability*, *11*, 493–503.

Hmelo-Silver, C. E., Chinn, C. A., O'Donnell, A. M., & Chan, C. (Eds.). (2013). *The International handbook of collaborative learning*. New York, NY: Routledge.

Howley, I., Adamson, D., Dyke, G., Mayfield, E., Beuth, J., & Rosé, C. P. (2012). Group composition and intelligent dialogue tutors for impacting students' academic self-efficacy. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Lecture*

notes in computer science: Vol. 7315. 11th international conference, ITS 2012 (pp. 551–556). Heidelberg, Germany: Springer-Verlag.

Janis, I. L. (1982). *Groupthink* (2nd ed.). Boston, MA: Houghton-Mifflin.

Johnson, D., Johnson, R., & Smith, K. (2007). The state of cooperative learning in postsecondary and professional settings. *Educational Psychology Review, 19*(1), 15–29.

Kirschner, F., Pass, 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.

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). Cambridge, UK: Cambridge University Press.

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research, 43*, 489–510.

Kolloffel, B., Eysink, T. H. S., & Jong, T. D. (2011). Comparing the effects of representational tools in collaborative and individual inquiry. *International Journal of Computer Supported Collaborative Learning, 6*(2), 223–251.

Liu, L., von Davier, A. A., Hao, J., Kyllonen, P., & River-Zapata, D. (2014). *A tough nut to crack: Measuring collaborative problem solving*. Manuscript in preparation.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*, 615–633. doi: 10.1007/S11336-012-9288-Y

Matsuda, N., Keiser, V., Raizada, R., Tu, A., Stylianides, G., Cohen, W. W., & Koedinger, K. R. (2010). Learning by teaching SimStudent: Technical accomplishments and an initial use with students. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Lecture notes in computer science: Vol. 6094. Intelligent tutoring systems: 10th international conference, ITS 2010, Part I* (pp. 317–326). Heidelberg, Germany: Springer-Verlag.

Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G., Cohen, W. W., & Koedinger, K. R. (2011). Learning by teaching SimStudent—An initial classroom baseline study comparing with cognitive tutor. In G. Biswas & S. Bull (Eds.), *Lecture*

notes in computer science, Vol. 6738. 15th international conference, AIED 2011 (pp. 213–221). Heidelberg, Germany: Springer-Verlag.

Mazur, A. (2002). *Theorizing feminist policy*. Oxford, UK: Oxford University Press.

Metcalf, S. J., Kamarainen, A., Tutwiler, M. S., Grotzer, T. A., & Dede, C. J. (2011). Ecosystem science learning via multi-user virtual environments. *International Journal of Gaming and Computer-Mediated Simulations, 3*, 86–98.

Mislevy, R. J., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., Yan, D., & Senturk, D. (2002). *Modeling conditional probabilities in complex educational assessments (CSE Technical Report 580)*. Los Angeles, CA: University of California, Los Angeles, the National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (1999). *Evidence-centered assessment design*. Princeton, NJ: Educational Testing Service.

Nkambou, R., Mizoguchi, R., & Bourdeau, J. (Eds.). (2010). *Advances in intelligent tutoring systems.* Heidelberg, Germany: Springer-Verlag.

Noroozi, O., Weinberger, A., Biemans, H. J. A., Teasley, S. D., & Mulder, M. (2012). Fostering multidisciplinary learning through computer-supported collaboration script: The role of a transactive memory script. In A. Ravenscroft, S. Lindstaedt, C. D. Kloos, & D. Hernandez-Leo (Eds.), *21st century learning for 21st century skills* (pp. 413–418). London, UK: Springer.

O'Neil, H. F. (Ed.). (1997). *Workforce readiness: Competencies and assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Prinsen, F. R., Terwel, J., Zijlstra, B. J. H., & Volman, M. M. L. (2013). The effects of guided elaboration in a CSCL programme on the learning outcomes of primary school students from Dutch and immigrant families. *Educational Research and Evaluation, 19*(1), 39–57.

Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009, October). *Assessment of student learning in science simulations and games.* Paper presented at the National Research Council workshop on gaming and simulations, Washington, DC.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes Processes. *Methodology and Computing in Applied Probability, 15*(3), 623–642. *doi: 10.1007/s11009-011-9272-5*

Romero, C., & Ventura, S. (2005). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Rummel, N., Mullins, D., & Spada, H. (2012). Scripted collaborative learning with the cognitive tutor algebra. *International Journal of Computer-Supported Collaborative Learning, 7*, 307–339.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Sandi-Urena, S., Cooper, M., & Stevens, R. (2012). Effect of cooperative problem-based lab instruction on metacognition and problem-solving skills. *Journal of Chemical Education, 89*(6), 700–706. doi: 10.1021/ed1011844

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility* (Research Report No. RR-07-27). Princeton, NJ: Educational Testing Service.

Soller, A., & Stevens, R. (2007). Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment. Alexandria: Institute for Defense Analyses.

Soller, A., & Stevens, R. (2008). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 109–111). Charlotte, NC: Information Age Publishing.

Stahl, G. (Ed.). (2009). *Studying virtual math teams.* New York, NY: Springer.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge, UK: Cambridge University Press.

Stevens, R. (2012). Charting neurodynamic eddies in the temporal flows of teamwork. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 208–212. doi: 10.1177/1071181312561020

Stevens, R., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., . . . Buckles, R. (2012). Modeling the neurodynamic complexity of submarine navigation teams. *Computational and Mathematical Organizational Theory, 19,* 346–369. doi: 10.1007/s10588-012-9135-9

Thompson, E. J. (1967). *Organizations in action*. New York, NY: McGraw-Hill.

Tindale, R. S., Kameda, T., & Hinsz, V. B. (2003). Group decision making. In M. A. Hoog & J. Cooper (Eds.), *The Sage handbook of social psychology* (pp. 381–405). Thousand Oaks, CA: Sage.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. doi: 10.1007/s11336-006-1478-z

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). New York, NY: Lawrence Erlbaum Associates.

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education*, *8*, 179–221.

Veen, A., & Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, *103*, 614–624.

Villalta, M., Gajardo, I., Nussbaum, M., Andreu, J. J., Echeverria, A., & Plass, J. L. (2011). Design guidelines for classroom multiplayer presential games (CMPG). *Computers in Education, 57*, 2039–2053. doi:10.1016/j.compedu.2011.05.003

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, *12*(1 supp), 83–100.

von Davier, A. A., & Oliveri, M. (2013). *Psychometrics in support of a valid assessment of linguistic subgroups: Implications for the test and sampling designs.* Unpublished manuscript.

Wageman, R. (2001). The meaning of interdependence. In M. E. Turner (Ed.), *Groups at work: Theory and research* (pp. 197–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Walker, E., Rummel, N., & Koedinger, K. R. (2009). Integrating collaboration and intelligent tutoring data in evaluation of a reciprocal peer tutoring environment. *Research and Practice in Technology Enhanced Learning, 4*(3), 221–251. doi: 10.1142/S179320680900074X

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of*

Computer-Supported Collaborative Learning, 6, 279–306. doi: 10.1007/s11412-011-9111-2

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87*(3), 406–423.

White, T., Wallace, M., & Lai, K. (2012). Graphing in groups: Learning about lines in a collaborative classroom network environment. *Mathematical Thinking and Learning, 14*(2), 149–172.

Wildman, J. L., Thayer, A. L., Pavlas, D., Salas, E., Stewart, J. E., & Howse, W. (2012). Team knowledge research: Emerging trends and critical needs. *Human Factors, 54*(1), 84–111.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688. doi: 10.1126/science.1193147

Notes

[1] See http://www.learnlab.org

[2] See https://www.coursera.org; https://www.edx.org