Innovative Assessment of Collaboration 2014
Panel 3: Education

Art Graesser

Let me introduce the folks on this panel. We're going to start out with Patrick Griffin, and he's chair of the Education Assessment at Melbourne University and also he played a big role in the assessment and teaching of the 21st century skills. He's also a very active member of PISA 2015. He's contributed substantially to that. Then we're going to go on to a number of people from ETS. There's Paul Borysewicz and Eric Steinhauer, and these folks are very clever at actually making the items for PISA 2015 and they played a role in this as well as like GRE and SAT. It's really very creative item development for collaborative problem solving. Then there's some woman here, Alina von Davier, who we all know, and she's Senior Research Director and leader of the Center for Advanced Psychometrics at Educational Testing Service and she also teaches periodically at Fordham. She just came out with an edited book. If you purchase that she'll be richer and independently wealthy. It's on the table out there in the front. And then finally there's Patrick Kyllonen and I seem to be seeing him about once every other month at various events. At ETS he's Senior Research Director of the Center for Academic Workforce Readiness and Success. And so this is a group that's going to get really down to evaluation of collaborative problem solving at a very deep, detailed and deep level. So why don't we start with Patrick?

Patrick Griffin

Thank you, Art, and to Alina and Patrick and co from ETS for the invitation to come to Washington. It's not a short trip. For a number of years I've been involved with a project initiated by Cisco, Intel and Microsoft to develop assessment materials for collaborative problem solving in a project titled the Assessment and Teaching of 21st Century Skills. It's generally known by the acronym ATC21S. We're pretty much down the track with the development of tasks and scoring and interpretation and we're currently working in just ten countries at the moment conducting some further research. OECD announced today in an email that they would fund a study comparing the human-to-human approach to collaborative problem solving to the human-to-agent approach to it and that will be jointly conducted by my own center at the University of Melbourne and the University of Luxembourg, so I look forward to another opportunity to talk about that to the group as well.

I want to sketch a little bit of the background to the project first because the thing that concerned the corporations that set the project up was that there was a shift in the workforce taking place so drastically that companies like Microsoft, Cisco and Intel could not recruit people ready for employment. Graduates and high school leavers were not being prepared for, at least in the developed countries, for working in a society, a workforce that was digitally dominated. And what they've been able to show through most literature, Autor, Levy and Murnane's work showed the rise of non-routine work and the demise of routine manual and cognitive work in developed economies. That drop in routine and non-routine cognitive and manual shift, an emphasis on that started to put pressure on education systems to start thinking about their curriculum and the changes that were needed to produce graduates ready to work in those sorts of

environments. It meant that the sorts of skills that we needed to prepare students to solve problems, to develop new ways of thinking involving creativity, critical thinking and many of the Cs that Eduardo was talking about early on today, new ways of working are emerging that everybody now tends to work through some sort of digital environment, and new tools that are necessary for working in the work taking place, their part in the workplace, and just to be able to live in a society that is basically becoming much more of a knowledge society or an information-based society and away from that sort of trend.

I want to talk about four dates first of all to talk about this, set this in context and the kinds of things that we need to be thinking about. 2015, five-year-olds enter the formal education system. When they get to year ten, which in the OECD is the first year that students can legally leave school, it'll be 2025 and these people will be entering a workforce looking for fairly routine, perhaps part-time casual work. 2027 will be the first year that that cohort of people actually leave with a high school diploma, and if they choose to enter the workforce at that point they also will be looking at kinds of work that probably are 17-year-olds going into service-type work. But in 2031, which seems like it's so far off in the distance as not to be relevant to us, this cohort that enters the school system next year, that's when they'll be graduating and entering the workforce as professionals. What we don't know is the sort of skills that they're going to need when they enter the workforce,15 or 16 years time, and the kinds of technology-dominated workforce that they enter will be different from the sort of thing that exists when they start school.

Lots of organizations are trying to do something to address this. Partnerships 21 in the United States has been looking at the curriculum throughout the United States to try and address that. The OECD, UNESCO, International Labor Organization have also been addressing these issues of the kinds of skill, and there is a consensus emerging in the kinds of skills that are needed and one of the dominant skills that appears to be emerging is the idea of collaborative problem solving. In the ATC21S project about 250 academics, researchers at the end of AERA in 2009 met and defined what they call the KSAVE Model. That's knowledge, skills, attitudes, values and ethics of the 21st century defining new ways of thinking, new ways of working, new tools for working and the way in which we need skills to live in basically a digital society.

In 2010 the first volume published by Springer actually looks at those particular skills and published that in 2012. In 2010 we decided to combine critical thinking, problem solving, decision making and communication collaboration and we called it collaborative problem solving, and that really became our definition of what collaborative problem solving was. It was a combination of these particular skills. We set up an expert group to being to explore and define it and it was led by an academic from Germany, Frederick Hesse from the University of Tubingen, and he defined, his team defined collaborative problem solving as consisting of social and cognitive skills.

The social skills consisted of participation, perspective taking, social regulation and the cognitive skills task regulation and knowledge building. Participation skills, we look at the action, interaction and task completion or perseverance. Perspective-taking skills were defined by the responsiveness or the ability to integrate the contribution of their partners, audience awareness, to tailor their contributions to the work that other

people were doing. Social regulation skills consisted of metamemory, transactive memory, negotiation skills and responsibility initiative or the acceptance of responsibility. The cognitive skills consisted of the extent to which they can analyze the problem, set goals, manage resources and the way they deal with ambiguity. Many of these things have already been raised in the discussion this morning. The learning and knowledge-building skills of the cognitive aspects consist of the way in which people collect information, the systematic approach to the definition of that, the way in which they see relationships and patterns, explain contingency or formulate rules, generalize and formulate hypotheses. PISA in 2015 as Art has said will be looking to examine the PISA collaborative problem-solving skill using human-to-agent, and it has a history of linking back to Polius(?) problem-solving framework which was explored in the 2003 and up to 2012 in problem solving and there's some need to link to that through PISA. ATC21S didn't have that kind of history.

I want to show you at least one example of a task that we've developed and used, two students working independently on computers, can't see one another's screen. Student A in this example is just looking at a scientific task of the beam balance, can pass weights to student B. Student B's task is to explore where to put the weights on the beam in order to balance the task. In real life we could think about this being something like a seesaw. We've got; I saw this as I walked through the park once where kids had simply put rocks on the seesaw to balance it. What they've done here is that they've participated in a group activity. This was something that they can't do on their own. They need two people to do this task. Otherwise it becomes a kind of comedy routine as they run backwards and forwards. So we can think about what's their level of

activity, their interaction with one another and the kind of persistence and perseverance of the social activities. They also need to respond to one another and they need to be thinking about what the other person is both listening to and contributing to the task. They also need to be thinking about their own skills or the metamemory and the transactive memory that was discussed earlier today in terms of the role of the other person and the contribution that the other person makes. They also need to be able to negotiate and to explore, and to explain, and resolve differences and to accept responsibility. In the task itself they need to be able to analyze the task. They need to be able to set goals of what they're trying to do, be flexible and deal with ambiguity and the management of the resources, in this case the rocks.

The thing that sets collaborative problem solving and collaboration apart from all other things is the fact that every partner in this has the command of a unique set of resources which have to be contributed systematically to the solution, and unless everybody contributes their resources, this problem can't be solved. The knowledge-building task in this, they need to be able to examine the elements of the task. They see patterns of behavior. From patterns people can formulate rules. Rules become generalizations and challenges to generalizations become hypothesis testing that the students get engaged in. What we like to tell the teachers is to encourage the students to ask questions that begin with the words, "What if?" or "What about?" So that's where we go with the task. We're looking for particular behaviors for each of those things and we're able to plot out the; because they command a specific resource we can track the contribution of each student by what happens to that resource.

We also log the activity file of everything that's happening in the background so that we get a data file, a log stream activity file which we then begin to analyze according to that particular conceptual framework that I've just been through. And as we move through that, we start to develop a data file by coding particular behaviors, then using – is that a signal that I'm finished? Once we've got the data file scored – what's happening here? Call on your imagination there. Once we've got the data file scored we're able to estimate the difficulty of each of the items or behavioral indicators as items and the difficulty parameters. And what this graph would show if I were able to fit it to the screen, it would show that in the four or five or six countries the difficulty parameters are invariant across those countries, so that we can argue that we are measuring the same thing across the countries, but we don't actually know exactly what it is that we're measuring until we start identifying and interpreting the construct.

We can map that out into five dimensions, each of the dimensions being the first three – participation, perspective taking and knowledge of social regulation, the last two being task regulation, task analysis and knowledge building. What you can't see is on the right-hand side of that task is the indicators, the number of the indicators which enables us then to interpret and to develop a described scale of increasing proficiency.

We set a response probability of naught point five when we are doing this analysis so that the description that the student is matched onto in terms of their performance is not an achieved level of performance in collaborative problem solving, but it's a performance of which the odds of success are 50-50 and we've gone ahead and interpreted that as a Vygotskyan sign of proximal development, so what we're telling teachers is this is what the student is ready to learn. And we were able to break

that down into two dimensions, which normally fit onto the screen, and five dimension, so we're able to give the teacher a description of what the student is ready to learn in terms of participation skills, in terms of perspective taking skills and in terms of social regulation skills or group composition skills. We're also able to show the teacher what the student is ready to learn in terms of task analysis and in problem solving itself. What we can do then is give the teacher, who doesn't necessarily have to depend upon technology, a very simple method of monitoring the student which doesn't work. The method works. This doesn't. We show the teacher to just use a highlighter pen to color in the sorts of things on the scale. In a pretest it would be yellow, and in the posttest some months later we ask them to use a different color and it then becomes green. And so we get a whole lot of yellow at the bottom and a lot of green at the top, and we're able to see the point where about 50-50 percent of the actions are shaded in and that's the point where the teacher should intervene for the student. And it looks good. I promise you that, but you'll have to look at it later. Thank you.

Eric Steinhauer

Hi. I'm Eric Steinhauer from Educational Testing Service. Unlike some other people in this room I was given a construct, which was a very happy circumstance for us. We were given the challenge at ETS working together with some international collaborators to design a test of collaborative problem solving as defined by the PISA collaborative problem-solving expert group. So I'm going to talk a little bit about the construct for PISA collaborative problem solving, the challenges we faced and some of

the design decisions that we had to make. In this case we were working with agents, so it's a different situation than the situation in which we have human-to-human interaction.

Currently the status of PISA collaborative problem solving is that the field trials have been completed. The field trial was completed in 54 countries for collaborative problem solving and in 69 different languages, so the challenges of design came out of having to assess with different languages, different education systems and different cultures. The definition of problem solving according to, that was defined by the collaborative problem-solving expert group, collaborative problem-solving competency is in this case the capacity of an individual – we were testing individual collaborative problem-solving skills – to effectively engage in a process where by two or more agents, and in fact we presented, there was always one, two or three computer agents that the student was working with, attempt to solve a problem, and by sharing understanding and effort required to come to a solution, and by pooling their knowledge, skills and efforts to reach that solution. The problems that were defined were intended to require all of the agents to collaborate in order to solve the problem.

So in terms of the construct there were three critical problem solving competencies – establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization – and those were crossed with the problem-solving processes that were defined in the PISA 2012 problem-solving construct with the result that you see what Art showed earlier, a matrix of CPS skills. And this set of CPS skills were the skills that we were targeting in designing the assessment tasks. So you might be called on as a student, you might be called on to engage in something that would require them to discover the perspectives

or abilities of team members. The student might be called on to describe tasks to be completed for the completion of the problem and identify situations in which their teammates weren't holding up their end of the bargain or weren't engaging in their role.

So the approach that we took, and this is not yet a Holodeck, as in some of the other approaches the student was communicating with team members via chat and there was a shared task space that the teammates were working with. The agents and the student could produce chat. The agents and the student could take actions in the task space.

So some of our challenges, assessment across languages. It would be great if we could do natural language processing of type text maybe in one language, not in 69 different language versions. So the design involved selection of chat. It was authored in a game scripting software. Also because of the need to go across the different languages, the need to translate, we wanted to make sure that we took off the table things having to do with the tone of the collaboration because that might differ across the different translations. So the basic structure behind the agents was a branching script that branched where the students made choices of actions or choices of chat and what we scored were their choices. But in some cases those were choices of paths through a branching tree.

Another challenge assessment across cultures and educational systems, one of the things that we selected to do was make sure that all of the teams consisted of fellow students, peers. We wanted to make sure that across different cultures we didn't introduce the possibility that somebody's reaction to perhaps an authority figure would be to confirm something they say even if it's not accurate. The other thing that; another

piece was to make sure that the problems that were solved were practical and nonacademic. The PISA has measures of science and mathematics and reading. This is a measure of their collaborative skills. And we had to choose things, activities that could be taken across different countries. It turned out to be rather challenging to find things that would be accepted across all different countries.

So a problem might be working together with a team to plan an event. Another challenge was making sure that we assessed effectively through extended problem-solving scenarios. So one of the things that would happen, we had to build in were rescues to ensure that a student who took a misstep at the beginning of the problem, that we continued to get some evidence of their problem-solving ability. Another advantage of this approach is that when an agent made a mistake in their choice of, in what they said or what they did, another agent could come in and note to the, observe to the human student that an error had been made. So there was a way to introduce the notion that the computer agents were fallible.

So we built in rescues and some resets, and we also built it in so that there was scripting and CPS orientation of the agents to ensure that we were covering the range of collaborative problem-solving skills. So the agents might neglect the goals or the rules of the project. The agents might forget the role that they play. And one of the things that we know, one of the things we were concerned about was maintaining student engagement, and we found that the field trial results showed that the students stayed with the collaborative problem-solving tasks. They went through the; in PISA they're standardly given four different sections and usually there's a significant falloff in terms of the amount of time they spend on the fourth section they received. In the CPS

measure in the field trail there wasn't that falloff, so we had very good engagement. It appears that engaging with the chat and having the agents respond to them kept them in the story. I'm going to hand it over to Paul Borysewicz.

Paul Borysewicz

Thank you. I know our time is almost up so I won't take too much time. What I'm showing you here needs to have a few caveats, and I'm not showing it to you now because I'm not. It's an example of an early, a set that we considered, a problem scenario that we considered using, but in the end we chose not to. So I'd like to leave you very quickly with two caveats – thank you – about this and I think three takeaways that are perhaps useful for further discussion.

The two caveats are that this instance in which the student is asked to imagine that they're in some sort of laboratory situation in which they're collaborating with other members of a research team, the two caveats are, one, this did in its initial conception have a hierarchical structure. It was imagined that; the student was asked to imagine that they were sort of in charge of the project and this was an approach that was later abandoned for reasons that Eric pointed out. Different cultures may see collaboration working differently according to when there are situations of differing status. The other caveat is that this situation, this lab situation was in the end further removed from typical 15-year-old experience than we wanted to go. We realize there's always an element of suspension of disbelieve in students engaging in these kind of scenarios and it was felt that this was maybe going a little too far.

But the three takeaways I'd like you to have from this, and here's the beginning of; this is a couple of introductory screens which we give to the students to get them into the story. And here's an instance, same screen that Eric showed you before, three things to think about. One, although this is in some respects a highly traditional form of assessment – it's selection of chat choices, but it also has a task space in which the students can take action – it did allow for multiple and indeed interactive modes of gathering evidence about students' ability to collaborate in these situations.

We were also able to model, rather surprisingly successfully I think, a variety of agent behaviors and orientations. In the case that I show you here we have an agent who has failed to understand the exact nature of the task, but is nevertheless clear about communicating his or her misunderstanding to the student. It was also possible at times to model a behavior where the agent not only misunderstood the task but didn't even know what they didn't know, and that, pleasingly to us, made for a more difficult collaborative problem-solving challenge for the student.

And finally the last takeaway what I'd like you to think about is that an unfolding narrative context, even in the simple branching tree of selected chat, it was possible to create an unfolding narrative context. It was indeed possible for people to begin to form a picture of these agents that they were dealing with which might influence their subsequent interactions with them in a way that I think mirrors the kind of evolving collaboration that we're all interested in seeing. So thank you very much.

Alina von Davier

My presentation is about the research program that we put together at Educational Testing Service around the idea of constructing collaborative assessment. So this is the outline of my brief presentation, hopefully brief enough, so I will talk about why we need the comprehensive research program, though by now after these two-and-a-half panels you probably already saw why it's necessary to look at this complex and hard-to-define construct from different perspectives. Then I will talk about the assessment design and data challenges and I will focus a bit more on the statistical models that account for dependencies. And I will do that because first of all that's where my background is, but second I think this is a good segue into the work that will be presented tomorrow which will go into more details into modeling this type of data.

I also want to point out that in my presentation I will explicitly separate measuring collaboration skills about which we've heard quite extensively today, but I'll also talk about measuring cognitive skills through collaborative problem solving tasks. That means to look at the collaboration as another mode of measurement, so it's a quite different perspective on using collaboration. And I will also refer for individuals and teams, so I will have at least two constructs and that will be for the individuals and teams. I also want to acknowledge the contributors to all this work at ETS and not only at ETS. Definitely I want to acknowledge my colleague L… who is here in the room, who worked with us as the content expert on many of our projects, and all my colleagues in the Center for Advanced Psychometrics, many of them being presenters here, the interns from the summer, Diego Luna Bazaldua from Teacher College, Columbia, and …. Wong from the University of Michigan, colleagues from other departments, …., and colleagues who are not from ETS, Art Graesser who, well, you know him, and he

worked with us indirectly by being behind the technology on the trialogue, and Peter Halpin who will present tomorrow, with whom actually my work on this area started.

So this is pretty much what it means at ETS to talk about collaboration, and we have research around all of these issues. So definitely we have still people talking about the construct, what is what we define to be a collaborative interaction, what are the best features, what does it mean to have a successful collaboration? Then we have a lot of research around the technology and delivery and scoring, and I will include here the work done by my colleague Saad Kahn, as well as …. Chan, and Gary Frank(?) and a few others. Then of course we still have work on assessment and psychometric design and that, I will say a but more about that because it actually matters how you think about your data collection. So and we try to bring to the table our ETS rigorosity(?) with the traditional assessment into this field that's a bit less structured such as assessing collaborative skills.

We consider nowadays including multimodal analytics in our research and Saad already mentioned some of that. Statistical modeling and data mining, we'll be talking more about that tomorrow, but a little bit more in my presentation today. And also more on data science, data management and log files. That's not a trivial thing to do nowadays when we have that many types of data and we want to merge different types of datasets.

So this has some examples of the instruments and data that are available to us and on which we do all this research. Sometimes a research project might not have the word collaboration in it, but it's one of the building blocks, for example working on log files and data mining. So one of the projects is the one called a tetralogue, and you

have here a screenshot of that and this is built on a trialogue on which Art helped us build initially. We also have; we also see there on the screen an example of the log file on which we do various forms of research in terms of what is the best way to capture the data in a meaningful way so that it is available for the second analysis that are necessary. These are other types of data and instruments that we've been using. We work in partnership with the GlassLab and we work on the SimCityEDU. This is a game for schools and we worked, we had various projects on that going from analyzing the process data to building the log file for it. Then we work on NAEP, NAEP's special task …. technology on which we analyze process data from different perspectives. We have a project on simulating data for complex systems using agent-based modeling and we have the basketball data on which …. …. our journey in learning about collaboration.

So what is important and what's different, which can be the same thing in this case? Definitely analyzing data from an assessment that includes collaborative problem solving involves aspects that are not to be found in a traditional assessment. One of them is the fact that both items and people exhibit different types of dependencies that are not encountered in a traditional assessment. Another assumption is that people behave differently when they interact in teams from when they work alone, and that the individual domain skill might not correlate highly with the team's outcome, which on some examples would be counterintuitive, but this is what research seem to indicate. Also we need to consider multiple scores. We have an individual domain score measured in isolation, as we traditionally do that. We have an individual domain score measured in collaboration and we have an overall team score. And this is just I would

say the first level of scoring, while depending on our results we might actually have even a more complex scoring rubric.

So what is what we are trying to do? I mentioned before we try to bring to this complex measurement situation the same rigorosity that we applied in the traditional assessment. So the way we approached this problem at ETS was by building interdisciplinary teams, and in a way this conference reflects that mindset as well, evidence-centered design to design the simulation, the games collaborations task, use large and representative samples, have robust measurement models, conduct predictability studies, investigate the fairness claims and analyze the process …. modeling. And as you can see in that little picture there, there is a process of collaboration. Even for such a simple message we still have several steps that show how those little donkeys(?) got to a successful result. So that means that without those interim steps we might actually not really understand what is going on there.

So I propose a test design that includes collaborative tasks in a way that takes the advantage of what we know about testing. So on the right side, so this is about measuring a cognitive skill, say science, and you want to measure the knowledge as well as problem-solving skills, and usually and traditionally we do that by having the individual tested in isolation through either multiple choice, open-ended or even games or other tasks, forms of performance task. So the addition here is that we add a collaborative component to it in which we would have a simple task, a human with an avatar. It could be dyad, so two humans, or it could be a more complex task that could include more humans or more humans and agents.

So what is that I'm talking about when I say about interactions? And this will go a little bit, I will link it back to panel one this morning where people were talking about collaborative propensity or the fluid skill. We call that in psychometric a latent variable. So if we have a hidden variable there that will be A1 A for person and B for the other person in a dyad, and we have a sequence of states for that person during the collaboration, the A1 going to AN(?). But actually what we observe that the behavior, so we observe the X1, X2, X3 and so on. Now what you see on that plot is that we have various types of dependencies there. You have the time dependence, you have the dependence between the two people and you also have the responses from each individual going forward. Another way to look at dependence which several authors before us considered were to think in terms of the interactions as once sort of dependence, the autoregressive events during the collaboration as well as …. event that would take place, could occur during the collaborative events.

So I am going to talk next about psychometric considerations when we deal with the assessment of collaborative skills. So here definitely we have to consider the possibility of multidimensionality and definitely dependence. So that multidimensionality of skills and team members can be presented as multidimensional skills, so we have a MIRT, multidimensional item response model, that may work in some circumstances, multiple time(?) series(?) for multiple people in a team. But the first one, the multidimensional IRT actually would not work very well because a dependence, a local independence assumption that's necessary for an item response model to hold definitely in this case doesn't. And it doesn't, and you see there the multiplication(?), it's over people and over item, and in the case of collaboration it doesn't hold either for over

items nor for over people and I will say more about that. So it's not over people because we have group interdependence or we even may have multiple teams nested in bigger multisystem teams. But it's definitely not over items because both the time dependence and a dependence in a complex task may occur.

So if you look at the portfolio of statistical models that are available to us for those different types of dependencies, we see that actually we have a good start. We could start analyzing this data by using static models for that account, for test takers' dependencies to meet team and group dependence, and we have multilevel modeling, social network analysis and so on, nonparametric models such as your article segmentation and so on. We have latent(?) class and …. networks. None of these elements will work for dynamic systems, so these will not work for a fully-collaborative analysis, but they may work in analyzing how teams work together, how teams for example vary one from another.

We can also consider other model like base …. and MIRT test-like(?) models to account for the dependencies in the items that are due to the task dependence. And I was told I don't really have any time, but I want to point out that we have continuous processes and discrete processes that could account for the real process data. We have dynamic factor analysis, dynamic linear model, differential equation model, and you will hear Sy-Min Chow talk about that tomorrow, inter-variability models, machine-learning models, and we also have more common for many of us at least the …. model, mark(?) of decision process, dynamic Bayes Nets and so on. The point processes, so perhaps speaker Halpin will say about that a few words tomorrow.

So here is an example of how we could use this model and this is an example of simulated bivariate(?) …. processes. …. processes are …. processes that work on the temporal structure of the data, and they actually have been quite extensively used in the past year or so as I found online for analyzing dependencies in economics, but also social dependencies. And this is an example on how you can model the interaction over a period of time, and you can see what happened at the given period of time, how the two interact with each other. This is an example of, and I found that also in the literature, where people use what is called a mixture …. model to analyze what happens when you actually have people belonging to more than one team. And it's not trivial to estimate the …. process. Well, imagine what it means to estimate a mixture …. process model.

Then we have indices available to us from statistics and psychometrics that actually capitalize on covariance and dependence and these are mutual information function, the entropy that will be used by Ron Stevens tomorrow. Co…. …. divergence measure. That's also going to be used in Peter's presentation tomorrow. Conditional information function, it's another index. Keep in mind that all of these models, as fancy as they seem to be, they are not easy to be implemented. We have often identifiabilty issues, so the parameters cannot be distinguished uniquely from each other. We have situations, and this can be related, ill(?) conditioning of the livelihood(?) function which leads to the difficulty of separating, estimating the parameters as we need them actually for various reasons, or the algorithm itself can become really cumbersome.

Now why that book there, the book on adaptive testing is at the entrance is because there are actually links between adaptive testing and collaboration, and the

CPS in itself can be seen as an adaptive test. The interactional team members influence the next outcome, there is learning that takes place potentially during a collaboration and there are adaptive features in the collaboration itself. Moreover the algorithm by which we can, assembly teams can be absolutely similar with we use to assembly test. So this is quite interesting and we have some research that started to look in that, that the way you adapt, the way you build tests can be used to build themes if you have appropriate parameters for that purpose. And why we use multimodal effect detection in collaboration? Primarily, so for two reasons, for measurement and for validity. So these are at least the areas in which we look. Thank you.

Patrick Kyllonen

So this is the summary of my talk. I'll start out with that. Employers want workers with collaborative skills. I'm going to show you why I say that. Consequently schools are starting to teach collaboration. But we don't really know how to measure collaboration very well, as you might have been able to tell by the talks so far, other than with ratings. And so we can't really tell, the school systems can't really tell whether they're being successful in teaching students the collaborative skills that employers want.

There are a number of challenges in collaborative assessment such as how to assign credit, what the test process measure, what the best process measures actually are, how student background variables and task variables affect team performance and how best to measure individual and collective outcomes. But this is the takeaway. Regardless of all that, regardless of whether we know how to do it, there is an immediate demand for good collaborative assessments to monitor student growth, for

developmental uses and for admissions, especially higher education and employment selection.

So first of all why do I say that employers want workers with collaborative skills? I'll show you some data. Department of Labor maintains a database called the O*NET, Occupational Network, and we did, Jeremy Burrus and several of us at ETS, we did an analysis of the hundred-plus ratings that are collected in O*NET, and we found that, we did principal components out of those ratings and we found that the teamwork component, the teamwork factor was rated the third most important of 15 components we extracted behind a problem-solving factor and a reasoning-ability factor, ahead of an achievement and innovation factor and ahead of a science and ITC(?) literacy factor. There are a number of employer surveys. I'm going to review those in a second, and there also is performance evaluation tools out there in the workforce that will make the same point.

One employer survey, recent survey, was based on this organization, National Association of Colleges and Employers, NACE, and as you can say when asked about the various candidate skills that employers want the number one rated skill on a five-point scale, average rating, was the ability to work in a team structure. Ability to solve problems also were very important. Number one, ability to work in a team structure. Another recent survey, Millennial Branding Student Employment Gap Study, a couple hundred employers, two of the top-rated skills in terms of what skills are employers looking for when they are hiring and what skills are the ones that are hardest to find, but most important are communications skills and teamwork skills. Another, this is a six-years-old study now, the Conference Board and others, it was a number of employers,

400 and some employers were asked what skills were most important, and for college graduates two of the most important were oral communications and teamwork and collaboration skills. For high school graduates those were also among the most important three factors.

One other slide to make the point. This is Lominger Competencies. These are widely used in employment for employee evaluations. And so employees, a very common use is that employees are evaluated on some subset of these competencies for promotion and for other kind of development. And you can see that roughly a third to a half of these competencies arguably could be called collaborative kinds of, collaborative and communicative competencies, negotiation and dealing with other people, relationships with the boss, delegating, developing others, directing others. This is why we talk about these skills as critical for the workforce.

So consequently schools are beginning to focus attention on teaching collaboration and let me just list some examples. There was a National Research Council report that Margaret Hilton, who's here in the audience, and Jim Pellegrino edited, and what they did among other things is looked at the Common Core State Standards. The Common Core State Standards, for those of you who don't know, are potentially a kind of a revolutionary change in American public education where states are adopting standards that are common across states, and so one of the other things about adopting common standards is the promotion of deeper learning. And in this report, the NRC report, there is documentation of the fact that there's a lot of collaborative, promotional collaborative activity in these new standards, these new common standards.

We already talked in this panel about OECD and the PISA 2015 emphasis on collaborative problem solving. Just last week the US Department of Education, NCES put together something called the NAEP Innovation Symposium where the subject featured was collaborative problem solving, and a quote from the motivation from that was that this is necessary to succeed in today's society, and so there's interest in the US Department of Education for doing maybe something perhaps along the lines of what PISA 2015 is doing. We talked throughout this panel also about it being promoted by educational nonprofits, ATC21S and P21. Also I should mention a project in Brazil going on right now as well as other jurisdictions where they're looking at the development of non-cognitive skills growth from grade 1 to grade 12, and this is in recognition of the importance of non-cognitive skills growth in education. Schools teach more than math and reading and yet we don't know much more about what schools do besides what they do in math and reading areas and so Brazil is experimenting with new assessments for non-cognitive skills. And then finally within higher education we now have pilot projects. In graduate school we have the ETS Personal Potential Index, which is a non-cognitive skill assessment which measures collaborative skills, and in business school, for example Yale School of Management, there's a pilot study that looks at collaborative skill.

Now a question is how do we measure that in these kinds of projects? And the answer is, the answer is coming up. We measure it with these rating skills. This is the basis for probably 95 percent of all measurement of collaborative skills in the world. So I work well with others is a typical rating scale item. Now we can improve on this. We have some good data. It's challenging and it's considered to be very radical and

innovative, but teacher ratings are one way to improve on it. Anchoring vignettes, we introduced that in PISA 2012 and that's a way of getting people to rate hypothetical vignettes which you can then use to rescale student self-ratings and that gives you superior measurement to the simple self-ratings. Forced choice and ranking methods, also we have good data to show that those provide better data than simple self-ratings. Situational judgment tests, some of you are familiar with situational judgment tests. That's another method. Example, as part of a class project you serve as a volunteer for a nonprofit. In a discussion about how to find new volunteers you bring up what you think is a great new idea, but others tell you that the idea is off base and not workable. How would you handle the situation? So there is an example of a situation that you can talk about, drop your idea, point out several good reasons why your idea is actually right, drop it for now but tell it to the boss later. These are all ways of getting at collaborative problem solving that are starting to come onboard in terms of prominence, and these are fine as they go, but there is an express need to have performance-based measures and not just ratings and rankings, and that's where we get into the ATC21S measure that Patrick talked about in PISA 2015, a measure that the folks over here have been talking about today.

And so we need a broader range, kind of assessment, and some of the issues that we have to deal with is – we've talked about this; Vincent mentioned it in the last session – when is collaboration useful? When are we better off working alone? Does collaboration improve individual performance, and if so why does it? Is it because students learn better strategies or they're just more engaged? Not to downplay the importance of student engagement and the collaborative problem-solving experience in

increasing engagement and motivation, so very important. Collaboration fails because coordination is difficult. People get their feelings hurt, and they disrespect each other, and they ignore each other and do all kinds of things that are not productive for healthy human relationships. And then there's the credit assignment issue.

So some of the factors we think we want to investigate over the next year are the types of interaction, the various kind of task content, curricular, math, reading and science, cross-curricular problem solving and critical thinking, social, interpersonal and physical. We've developed a situational judgment test for athletics that gets at some good team collaborative constructs. The nature of the task well and poor defined, group composition, various kinds of evaluations, individual team and process evaluations and a variety of process measures which we're working together with Alina to try to see how we can measure and instantiate. And so we have a kind of a framework, a beginning framework that talks about how we go about looking at that and we hope to evaluate this framework over the next year.

So here's some questions really are what is the quantifiable benefit of collaboration as a function of the type of individuals, the nature of the task, group composition, individual versus team versus process outcomes and using various process measures? And is there such a thing as a good team, and how task specific is that and what are the characteristics of good teams? And are there good team members and is that task specific or does that vary with types of tasks? And what are the characteristics of good team members? What is the relative importance of personal attributes versus task prior relationships?

So in summary again this is the takeaway. There is a strong and immediate demand – I hope I documented that for you – for assessments that can identify individuals and measure teams on collaborative skill in K-12 systems around the world, and higher education and the workforce. And these assessments will be used for selection and admissions development, student, employee and team development and growth monitoring at the school, district and jurisdiction levels. Thank you very much.

AG: = Art Graesser
PG: = Patrick Griffin
PB: = Paul Borysewicz
ES: = Eric Steinhauer
AD: = Alina Davier
PK: = Patrick Kyllonen
R: = Other Speakers

AG: Just like the other panels we have a number of questions. You have them up there on the PowerPoint. One is how can we disentangle the collaborative problem-solving skills from other skills? And secondly how can we assess cognitive skills through collaborative problem-solving tasks? Number three is how can we standardize the task when humans work together? And four is how can we disentangle the individual contribution from the team contribution? So why don't we start kind of similar to Alina's approach where people can comment on one or other of these questions and why don't we start with Patrick?

PG: Let me tackle the first one, how can we disentangle the collaborative problem-solving skills from other skills? I guess the first thing we need to have a fairly clear idea in our mind what we mean by collaborative problem-solving skills, and for that we need to have some conceptual framework, either the PISA 2015 framework where

problem solving is cross-matched with collaboration, and the intersection of those two skill sets begin to define a framework for what would be collaborative problem solving. The other framework that exists and is now published, I should say that Springer released the electronic version or the e-version of a volume on ATC21S on Thursday of last week and it's available on electronic form where the approaches and methodologies over the ATC21S project has now been fully published. The framework there looks at both the social and cognitive skills associated with collaborative problem solving. So the first thing to do is to know what you're looking for and to begin to identify what might be regarded as evidence of those particular skills, evidence that we can either identify and from which we can make inferences about the possession of collaborative problem-solving skills in this case, the social and/or cognitive skills.

I guess that gets on to the second question, how do we assess the cognitive skills through collaborative problem-solving tasks and I would have the same answer. Know what it is that you're looking for and the evidence, and know what the skills are and what they might appear to be. It always begins from a measurement point of view by defining the construct first. And I know this came up in earlier sessions about whether the construct exists. My reaction to that is of course it exists because a construct is nothing more than something we construct to help us understand our observations. It's not something that exists in and of itself. We build a construct to help us understand our observations. If we know what our data is and we've got some idea of what consists more or less of some behavior and we can identify indicative evidence of that, we can begin to identify the construct of the cognitive skills, as long as we can define them, and then look for evidence of them.

AD: I agree with everything that Patrick said. I would only want to add that research design is quite important …. these questions. So in addition to specific collaborative tasks, one may want to consider having different types of assessment that will target different parts of the construct. For example, if you are interested in a science assessment, electronic skills for example, you might also benefit in your measurement if you can collect data about electronics where people would be tested in isolation, including multimodal data could also help if you are able to build as Saad Kahn explained today from the images that you collect or the speech that you have to build into the affect that's part of, that could be part of the construct. So I think that I mentioned in one of my slides that research design, assessment design is quite important and I think it can help in addressing at least one, two and four definitely, the questions here.

AG: Well, Alina, since you have the floor or the table, whatever metaphor you want to use, do you want to address any of the other ones, any of other questions?

AD: Well, so I think what I said before is true for one, two and four. So after you have a construct in mind and you know what you want to measure, then I believe that the design, a more complex design that includes different instruments can help. That was the point I was trying to make About the standardization of the task, I mean that's definitely something that's very important to me given that my previous work in assessment was on test equating. And in test equating we are driven by fairness and we want to make sure that an assessment is of equal difficulty across administration. So now we bring people to work together. What does that mean? Will we make the task more difficult by assigning people to specific teams or will we make it easier? And what

is what we can do to adjust for those differences in difficulties or perceived difficulties? And one idea again I believe it's in design. The design can help us and that's again where multi-states(?) testing or computer-adaptive testing algorithm can help. For example, on the work that we do at ETS on the …. that Jiangang Hao is going to describe tomorrow, we plan to standardize this by using the multiple-choice on science as the first assessment and then assign people to teams based on their ability as measured by the multiple-choice test. So it's like a multi-stage test in which the difficulty in the second stage is not driven by the difficulty of the task, but is driven by the difference in ability of the partners. Well, it turned out that so far we haven't been able to achieve that design, but this will be done in the next step. So I think that the design is quite powerful and we should keep that in mind for all of our next studies. So it's related to my talk that the psychometric toolbox is quite useful for addressing this new type of questions. We just need to know what to look for or which tool to pull out.

AG: Paul or Eric, do you want to speak to any of these issues among the questions?

ES: Sure. Well, so we have the advantage of having the agent, so we didn't have to standardize the test when humans worked together. The advantage of having agents is that you can present the test taker, each test taker with essentially the same team to work with and with a set of collaborative problem-solving challenges, a set of essentially hurdles to measure their ability to carry out a collaborative problem-solving skill. When you get into the situation – and this is where the challenge lies – when you get into the situation where you have humans working together, I think Patrick you have ways of standardizing the task for the team as a whole. And then the challenge that we

are having, and it hits with question four, is that if you are able to evaluate the success of the team, how can we disentangle the individual contribution? So standardizing for the team is viable and Alina's ideas of figuring out how to team people up is another way of trying to standardize the team. But standardizing the team is viable, but when you; the team task is viable, but the challenge then becomes disentangling the individual contribution.

AG: Paul, did you want to? Okay, Patrick.

PK: Yeah, I wanted to approach this from a fairly simplistic idea and it's this, that when you look at PISA the correlation between the verbal, the reading test, the math test and the science test and the problem-solving test, again given 60 or so countries, 70 whatever languages, is surprisingly high. The correlation is about .85 on average between these four(?) …. problem-solving, problem solving, reading, science and mathematics. In other words there's not much differentiation. If you know someone's reading score, they're going to do a very good job of predicting their problem-solving score, and their math score and so on. So to me the issue of disentangling CPS from other skills is, is there a CPS test that can be developed that doesn't correlate .85 with the math, reading and science test that already exists? If there is then that would be very exciting. My guess is we already; there are some suggestions in the literature that there is some differentiation and that's the W…. study, the Carnegie Mellon and MIT media study where they looked at group output on a variety of tasks, a battery of cognitive tasks, and found that individual abilities were not that predictive. The correlation was about 20 or so with group outcomes. So that's a nice illustration, if it's replicated, that there might be something to person-to-person collaborative problem

solving that's not the exact same as math, reading and science ability. And this is why I think people are very excited about PISA 2015 because we know, we haven't seen the data yet, I haven't seen the data yet, but we know what the league tables will be like. Shanghai-China will be half a standard deviation above everybody else, Korea, Singapore, Finland and so on will be in the top five, and we see this in cycle after cycle after cycle. My thought was that perhaps with collaborative problem solving, if it's not the exact same as what we're already measuring, with problem solving there might be some change to that league table. There might be some countries that are not on the top now that might actually be on the top and some other countries that are on the top that might sink down 20 or 30 places, and that would be very interesting and that would convince me that collaborative problem solving is different from some of these other skills.

ES: So I know a little bit about the field trial data. It's still being analyzed. One of the things that I, at least I was most concerned with given the way that collaborative problem solving is being measured was that we would have a very, very high correlation with reading. It would just be another way of assessing reading. I can't give you the numbers, but the correlation with reading is not, wasn't problematic at least in the field trial data, so it looks to be, you know, basically we're testing something different from just reading. That's something.

PG: Patrick's point is a good one. What might be a bit of a fly in the ointment with that one is that the countries currently practicing like crazy on collaborative problem solving include Korea, Singapore, Finland, Taiwan, so there could well be a sustained world order in collaborative problem solving that's not related to the skill itself but the

practice effect of it. I wanted to have a crack at a simplistic answer to number four on this as well. Both Alina and I were talking about dyads and there were other speakers talking about dyads in the collaboration sphere. We actually need quite urgently to develop tasks and three people, minimum of three people in a team so that we can get a multilevel model operating and we can get an estimate of the team as well as the individual, but we can't do that unless we get a minimum of three in the teams. Unfortunately that dawned on me about two years down the track after we'd sat down and only had two. But if we can encourage people now to build collaborative tasks with a minimum of three in a team, then we probably do have the psychometrics to have a first go at this with multidimensional, multilevel response modeling that we could approach this in and we could get a measure of the team and the individual at the same time.

AD: I really like that suggestion, so Paul Horwitz and John Chamberlain, who are here in the audience, and I and Al Coon(?), we have an assess project on teaching collaboration using electronics and the teams will have three people.

PG: Oh, great.

AG: So, Patrick, you do need to have a name for that. I know we have trialogues with two agents and a human, and we have tetralogues, two humans and two agents, so what are you going to call yours?

PG: Well, don't extend the idea of dyads to triads. That will have different connotations.

AG: This will be maybe a good time to turn to folks out there to raise questions, concerns, and so please approach the microphone if you have questions, or comments, or expressive evaluations or whatever. Who's going to start? There we go.

R: [Eduardo Salas] So this morning I gave you ten observations after 30 years of working on this. I'd like to add a few more after hearing you. Observation number 11, we know a lot about what affected(?) things do. I think that you need to be informed about that science of teamwork that's out there. There's a lot of work that's been done over the last 50 years that can not perfectly, but can tell you what effective teams do and what are their characteristics. I suggest that since this was funded by ARI that you read Jim Dyer's report 1984, a seminal piece which …. review from 1955 to 1980, what is it that we know about team performance and team performance measurement. Since then we've done work to update that. But there is a science of teamwork that can inform what this panel has been discussing.

Observation 12, team training works. There are programs out there, especially in the medical community, there's a bunch of meta(?) analyses that are coming out that show that effective teams in complex environments like hospitals, those who are trained on three or four skills save lives period. I mean so there are programs out there that work.

Observation 13, task work is necessary, but insufficient for effective teamwork. So like Steve Fiore pointed out this morning, you need to look at those two tracks as [is?] the individual …., that is things I do in my own job and the collaborative things, and the collaborative things add variance, predictive [predicted?] variance, so you don't need to look at those two things. Finally I was going to say something about task

interdependency and I think that's a concept we cannot forget here. Whatever the problems are, I mean in the literature we make a distinction between – this is kind of academic, but I think we need to discuss – within a group and a team, and the big distinction is task interdependency. So when we talk about CPS, what is the level of interdependency among the team members? That needs to be clearly defined because if there's high interdependency then a lot of this literature I'm telling you about generalizes. If it's low interdependency then we have a problem.

AG: One thing I might say is in the 2015 framework interdependency was a requirement in all the tasks, yeah.

R: [Eduardo Salas] Then 80 percent of what's out there from the science of teamwork applies to what we're talking about here.

AG: Thanks.

PK: And more generally I think those are very good points, Eduardo, and more generally training works. Right? So we know that training is effective in getting people to be better performers regardless of the job, as well as team training and other kinds of training work. Nevertheless people vary in the degree to which they benefit from training, and so we see large individual differences in the training growth in performance as a function of training time and so forth, huge ratios from best to worst. And so one of the issues I think is unresolved is the degree to which training works. Team training works, but the degree to which individuals differ in how much they benefit from team training. And so in other words team training works doesn't mean that we get everybody up to the same level as a performer on a team in a given amount of time. And so this is where I haven't seen much literature on this kind of phenomenon. What are the

characteristics of individuals that are most amenable to team training and what characteristics of individuals best predict who are going to be the best team performers?

AG: Do any of the other members want to also respond to Eduardo's comments?

AD: Just a small comment. I think they are very good comments. I want to mention that we also considered this degree of interdependence in our report, Peter Halpin and mine. We also analyzed or at least discussed and classified the teams depending on how much they have to interact with each other and we decided that; I just said that a certain level of interdependence is needed, but if it is too high then we cannot use it for assessment purposes because people, you cannot just have people assigned to teams if they've been working together for a while.

PG: I wonder if Steve Fiore might have a comment about that, the distinction between just teamwork and team training, collaboration, cooperation, coordination and those terms that he raised this morning and whether or not they apply differently to different contexts and different kinds of teamwork or team tasks. Or has he disappeared now?

R: [Steve Fiore] Well, I think the points I made this morning do still stand, but with regard to your question about team training there have specifically been methods that are developed to train coordination that has been shown to facilitate adaptation. So in the team training literature, team adaptivity and team coordination training have been shown to be effective. But those are pretty much in action teams, and I think what we're talking about here with regard to problem solving is a different kind of coordination. We're talking about knowledge coordination and not necessarily behavioral

coordination. And there has been some work by Joan Rench(?) that is doing training for knowledge development and knowledge integration that is somewhat akin to this idea of knowledge coordination.

AG: Next question. Thanks.

R: I had a question about the PISA part of the assessment. The skills that you guys set out in that matrix, they didn't seem to have a cognitively-demanding aspect to it in the sense of the task.

AG: They didn't seem to have what? A little bit louder.

R: Sorry, a cognitive aspect to it in the sense that the task wasn't supposed to be a very hard math task or a very hard chemistry task. It was supposed to be relatively low level in terms of those domain skills, if we can call them that. And I'm wondering does that affect the kinds of collaborations you might expect in those tasks and will those be I guess ecologically valid or is that an open point I guess?

ES: It's a good question. I think one of the risks associated with trying to develop tasks that place the problem-solving aspect of the collaborative problem solving at the relatively lower end of what would be the 2012 PISA problem-solving spectrum is that the concern you might have that these collaborative problem-solving skills we are assessing are not going to be generalizable. So that's, I mean for any test of this sort you're going to have that challenge. The rationale behind going that way was that by adding the aspect of the task where you had to be attending to the reactions of, to the knowledge of, to the contributions of your team members that the overall cognitive load was going to be significant, and that if we made the problem-solving task itself more challenging we would get, that would swamp the measurement of the collaborative

problem-solving skills. So it's a good point and it's something that will have to be looked at when we get the data. We did find that there was a range of difficulty of the collaborative problem-solving skills that was not entirely dependent on the difficulty of the problem-solving task itself.

R: Thank you.

AG: Yeah, I'd add that I'd put it more like low and intermediate in problem solving that they had. Other questions maybe. Anybody else or is it five o'clock?

R: I could re-ask my gender question, but I won't. I'm a big believer in collaborative problem solving. I worked with a simulation at the University of Maryland about 10 or 15 years ago that had students simulating diplomats from different countries, and computing in computer system, and I did a lot of work on the various constraints and co-construction that those students did. But when I or we tried to take that into some schools, and I'm addressing now the issue of the employers want this, the question is what happens when you try to get this into a developmental framework with adolescents? The reaction of many parents to this is, "I want my kid to achieve on his own, to be able to own his product. Don't tell me that he should be helping somebody else or collaborating with somebody else. I don't care what the employers want. My kid needs to shine for himself." And I think that we need to; I think it's a fine direction to take here that you're proposing, but we need to be aware that there are people, I think especially in this country – I've done a lot of international research; it's somewhat less true in some other countries – but the whole idea of collaboration means you're taking something away from the individual and their opportunity to shine has to

be considered as you move forward this effort. And it's not a question as much as it is just an observation. I would like to think it was different, but I don't think it is.

PG: There are lots of connotations in that. One is trying to change the teacher as much as trying to change the parents. But also when you start to put collaborative tasks in the classroom as assessment tasks and you've got students working together to solve a problem in an assessment context, that's cheating, and that's a real mentality that you've got to change in the classroom and the capacity to produce individual performance measures in a collaborative task actually can go some way towards breaking that down. But if we only produce a team grade we then have the problem of what we call lurkers, those who sit back and watch the rest of the team do all the work while they get the credit at the end. So being able to produce individual scores and individual grades within the team is an important way of breaking that down.

PK: I wonder if I could add there's been a pretty dramatic change in our recognizing the importance of some of the skills that before haven't been recognized. I'm thinking for example of Roger Weissberg, University of Illinois-Chicago Collaborative for Academic, Social and Emotional Learning actually has, there's legislation in Illinois, state standards for social and emotional learning. In other words collaborative learning is recognized as something states should teach. NBC News just put up a website last weekend that tells, it's called a Parent Toolkit on how to develop not just your child's math skills and their English language arts skills, but also their social-emotional learning skills. So it's a very significant change. This wasn't possible three years ago and it's happening very rapidly. So I think that the idea that we're in it ourselves, and we work on math and reading and that's it is changing, and so I think you're going to see

increasing openness toward the idea of learning additional skills besides the ones that have historically been the focus of our public education system.

AG: Do you think bullying had, was part of the stimulus there or the economic?

PK: Bullying is very important. In other places, for example in Singapore there's the recognition that Patrick was kind of hinting at that their students have already achieved the highest standards in the world in terms of standardized tests, but the economy isn't following. What's missing? And so there's recognition that there are these other skills, these collaborative skills and other skills that are important, and that's why they're out there working on this before we are and we'll be trying to catch up with them in collaborative problem solving in a few years.

AD: And I also think that if we make progress in disentangling the individual skills from the team contribution, then that will help in turn both teachers and parents. So if we are able to show who was contributing significantly versus those who did, you know, social …., then parents and teachers will feel more comfortable with the assignments and things.

AG: Steve?

R: I wanted to revisit a point that Pat made earlier about the high correlation between the reading, the math, and the science and the problem solving, and whether or not there really is something different. And earlier this year David Udall(?) and colleagues at Northwestern published a meta analysis on spatial thinking, and from that they made an argument that spatial thinking is something that's trainable, but it's not always something that is assessed in a lot of these tests and they linked it to STEM education and STEM learning. And I don't know the degree to which PISA problem

solving has anything to do with spatial thinking. I don't know the degree to which the collaborative problem solving has anything to do with spatial thinking. And if not I suspect that if you add that, then you're more likely to get some different correlations than you currently are.

AG: Do you want to address any of that?

PB: Yeah. In fact, some of the tasks, some of the scenarios did involve a fair amount of spatial thinking, laying things out in a two-by-two grid or planning people's movements to various destinations, so there was some spatial element at least in a couple of the scenarios that people were confronted with.

PK: And just to add, that's a very good point and David Lapinski, Camilla Benbow and Harrison Kell have published a couple of meta analyses showing that very strong spatial ability is a nice predictor of achievements in STEM fields above and beyond the math scores, fairly recent work.

AG: Yeah, Carolyn?

R: So this might seem like a kind of out-from-left-field kind of a question, and I've asked this question in a different setting once before, but I wonder now that there are countries of people practicing up for this test, and thinking about the ranking and all whether there's been any press about this, and I would think that you would be aware of that. I guess one thing that's always made me nervous about this test – maybe I'm just paranoid – is the idea of what would be the political implication when rankings of countries in terms of collaborative skill started to come out, and when people were doing secondary analyses where they're breaking it down by maybe gender, but by religion or other demographic characteristics as well. Is anybody concerned about what

that could lead to in the end? I mean I'm as excited as the next guy that people value collaboration enough that it will be tested and hopefully that makes more space for it also to be taught, but I have been concerned about also this possible negative, so I wonder.

AG: How about some policy comments?

PK: Well, you will be able to analyze those datasets in about a year or two years. Those will be publicly available, available for secondary analysis. Of the variables you mention religion is not going to be on the questionnaire, I can tell you that. That's been excluded from all the PISA questionnaires. But some of the other, gender certainly, you would be able to look at the data by gender, by schools, by countries, by a lot of other background, personality, a lot of other background characteristics.

R: I guess the question really though is anybody concerned about what somebody might do with that information once the analyses are out? That was my real question.

PK: Well, it's not traceable back to an individual.

AG: I think she's wondering about the broader consequences in the media and politics, and policy people like yourself might have some insight onto that.

AD: Well, it will be abused as the current data. I mean you can't really do much about that and all the scientists have been pointing out that perhaps all this classification and rankings do not make much sense. They never take into account the standards that are associated with those group means that are listed there. So there are a lot of other features that the media doesn't take them into account despite the fact that people like us are trying to be conscientious about that, and be honest and try to be

precise. When we try to be precise in bringing the equations, we are told we are boring and therefore the media will go on with the ranking which is much more of interest. So it will be abused as all the other ones and it will lead to teaching to the construct. As Patrick mentioned that's probably already taking place given that those countries have a chance to see what the test is about. Those countries in which test preparation is very important, they are already probably preparing their kids for that, for this type of test. So I don't think we have a way around it except to keep our common sense and critical reading of all the news. I think that's what we can do. And on our side to be as transparent as possible about the results.

PK: If I could, countries are very interested in this information for policy purposes. Countries are very interested in where they stand with respect to other countries, and that's why there's very significant investments, you know, public sector investments in these, supporting these kinds of assessments. It wouldn't exist if country policymakers didn't want the information provided. It's been a major instigator of educational reform in a number of countries, not in the United States by the way. We don't really pay much attention to PISA scores in the US. We have fairly low participation. It doesn't get that much media attention. But in other countries there's a lot of attention given to it. I was in Santiago when PISA scores were released last December and there was tremendous interest. There was an hour-long special on the meaning of Chile's scores, the cycle, and how it looked like Chile was still the highest-scoring Latin American country, but Brazil was catching up rapidly. And so this is the kind of information that policymakers around the world are very interested in. There is a lot of discussion about the fact that educational reform in Germany was heavily

influenced by the reporting of PISA scores. And so my, and there are other countries as well. And I heard informally – this is not official – but someone from one country told me that they did not want their scores reported because they did so well that they thought that that would put a stop to efforts at educational reform. And so these are taken very seriously around the world and my guess is that if the collaborative problem-scores were significantly different from the math, reading, science and problem-solving scores, the countries that didn't do so well would take that information – unless it's the United States in which we didn't even care – but if other countries saw that they were not placing as well as they did with the other scores, my guess is that that could be a big motivator for educational reform in promoting collaborative skills. And that's why I think some of us were hoping that the assessment would reflect collaboration the way we're talking about it in this meeting. We'll see how that plays out.

AG: One point I should make is since I've been involved in PISA I've noticed it in the press more, and I don't know whether it's because I was involved in PISA. Have you documented any changes in press coverage of PISA in the United States in recent years?

PK: I don't know.

AG: I would be willing to bet 38 cents there's been an increase. Okay, yes.

R: Hi. I think for once the mic might actually be at my height, which …. So on the note on educational reform one of my concerns; I work at the Educational Policy Improvement Center. It's over in Eugene, Oregon, little(?) Oregon, and we're kind of approaching collaboration and other skills from the other direction, more from the learning environment, hoping that we could kind of build up and meet in the middle with

assessment, which is why I'm here. But one of the concerns that we have is given the great benefits to distinguishing problem solving from subject areas or domain areas in terms of measurement, and even in terms of comparisons across countries, I have a little bit of fear for what that would translate into given the top-down nature of education right now. And one of my biggest concerns is that teachers or educators will take that as we have math class, science class, reading class, collaboration class and so forth, not knowing how to integrate it within the domains. So I don't know if that's something you've thought of or if there is any ideas?

PK: Well, I think that; I mean in countries that are doing fairly serious collaborative problem-solving interventions it seems to me, at least from my observations, that it's very embedded in the curriculum of other subjects, a little bit like writing. So all the reform on increasing the amount of attention that writing gets in the schools tended to come not as a special writing class, but as embedded within the curriculum. It's not the only model, but, and I've seen that in collaborative problem solving as well. So, yeah, there's always this kind of idea that there's going to be some disembodied kind of topic called problem solving, but that didn't happen with PISA when PISA in 2003 brought in a problem-solving test. I don't know of any school systems around the world that all of a sudden started teaching problem solving, but so I'm not sure that that's a real, something to be really concerned with. I think it's much more likely, again following programs that I know in different countries that have actually tried to improve collaborative problem solving, it has been drills, and it's been exercises, and it's been reforms to the actual subjects themselves, such as literature, or science or something along those lines.

R: Is that a process? In terms of the way assessment has evolved in this country is that something you could foresee happening where the subject areas specifically will start incorporating it as opposed to another statewide assessment that includes domain-independent problem-solving tasks?

PK: Yeah.

AD: Yeah.

R: Well, thank you.

AG: Thank you. Patrick, did you want to end with a few words?

PK: Oh, yes.

AG: I know he was going to prepare this huge 20-minute summary that he summarizes in the word collaborative problem solving.

PK: Right.

AG: But only a few words.

PK: I would like to just end by saying this has been a really terrific day. We had three panels, outstanding panels with excellent presentations, well prepared, and I think that at least speaking for myself, I learned a lot. I'm a lot more aware of some of the issues, challenges and findings in the field than I was this morning and I hope that this was true for all of you too. And tomorrow we're going to have, we're going to turn our attention towards some of the measurement issues that Alina alluded to and started hinting at in her presentation, and so we look forward to that. Now notice in the calendar that we start a lot earlier tomorrow because we have a lot to pack in before we leave at noon or so, and so I believe the starting time tomorrow is 8:30 with the breakfast a little bit before that, but we actually begin presentations at 8:30. Is that correct, Alina?

AD: Yes.

PK: Yes, so make sure to be here at 8:30. But with that, thank you very much, everybody.

END OF PRESENTATION