Innovative Assessment of Collaboration 2014
Panel 4: Statistical Models for Process Data

Alina von Davier, ETS

This is a very nice validation for all of us that this is a topic of interest. I also want to welcome you all and also perhaps there are some of you who are new to this conference today. Welcome. Thank you for joining us this early, and I'm looking forward to the next two panels and discussions around psychometric models. So yesterday we've been hearing all of our colleagues commenting on the challenges around the measurement issues of complex systems such as collaborative interactions. Today we'll hear from our colleagues about various attempts to address those challenges.

I also want to remind everyone that one of the co-organizes of this conference is Mengxiao Zhu who couldn't be here, so I'm bringing up her name because, as I said, she can't be here, so what I want you to remember, that she is one of the organizers together with Pat and with me. Thank you and I wish all of us and me a very nice conference today.

Patrick Kyllonen, ETS

So we're going to have four panel participants and it should be a very exciting, very interesting, very informative set of presentations. So we're going to begin with Peter Halpin. There he is, and Peter is an assistant professor of applied statistics at Steinhardt School of Culture, Education & Human Development at New York University, and he received his PhD in Psychology from Simon Fraser University. And then we're going to, I just saw him, I know he's here, we're going to hear from Ron Stevens who is

professor at UCLA School of Medicine and is a member of the UCLA Brain Research Institute. And then after Ron, we're going to hear from Yoav Bergner who is a research scientist in the Center for Advanced Psychometrics at ETS and who comes with a theoretical physics background at Harvard and MIT. And then we're going to hear from Jiangang Hao who is also a research scientist in the Center for Advanced Psychometrics at ETS who also received a PhD in physics and a Master's Degree in Statistics from the University of Michigan. So, with that, I would like to begin with our first speak, Peter.

Peter Halpin

I'll be talking about ..... collaborations and point processes and related statistical procedures. This is joint work with Alina von Davier. We've been working on collaborative problem solving for a couple of years now, and that was originally based on work I had done on dyadic interaction. The data we'll be talking about are due to Alina, Jiangang, and Lei Liu's extensive efforts to collect some collaborative assessment data over the last year. So thank you to my collaborators.

The topics that I've been discussing with Alina lately is how can we get something like a collaboration index. I put that in scare quotes because yesterday we heard a lot of different interpretations about what collaboration is. It's an in umbrella with a bunch of subsidiary skills. It means a lot of different things to a lot of different people, and in this context I'm really just trying to phrase the problem of how can we disentangle the complexity and interaction down to the individual components. And index is also in scare quotes because one of the things we can do is count kind of indicators and in

chat(?) data we can count things like word count or number of chats, but these aren't exactly indicative or exactly what we mean by collaboration. So a person can talk a lot, send a lot of chats, not necessarily interacting with our partner. In a similar way, you can respond to a lot of items on a multiple choice test without necessarily getting them all correct, although the number of items you respond to might be correlated with the number you get correct, it's not exactly what we mean by an index is ability. That's generally the problem that I'm working towards, and something I won't talk about is basketball data. That's in the abstract, something we've been working on for a couple of years because we haven't had a really good educational collaborative task to deal with. I'm going to talk a little bit about the Tetralogue data. I won't get into too much of the details because that's something Jiangang will be talking about much more extensively.

Thinking about process data, basically approaching it from a time series perspective, you have a time axis, you can bin it and within each bin you can observe some outcome. So, if the bins are time units, the next site was 1, an example would be whether or not a chat was sent. Holy smokes, that goes fast, whether or not a chat was sent in that time interval. The bins don't have to be time unit, the bins can be, for example, item responses or other units of analysis. Basically, any uniform task component that you can organize sequentially fits into this modeling framework or this time series framework. And there's tons of stuff on how to analyze this data, and the stuff I'll be talking about today isn't novel in any sense, although the application might be a little bit.

So process data and temporally complex tasks, so we have a collection of events. A simple task is one in which there's no dependence on what happened in the

past. So a multiple choice exam, it has this characteristic or the way we treat multiple choice exams we do it in this way. So the order in which you answer the items doesn't affect this correlated(?) event. On the other hand, a complex task is one in which the events, the sequence, or the timing of the events does matter. So in a dialogue or an interaction, obviously if we commute, the turn is taken, we still have the same number of ....., we still have the same number of questions, statements. We still have all the individually or aggregate descriptive data the same, but what's been changed is something that I'll refer to agnostically is the complexity of the task. IRT(?) treats(?) tasks that simple, and I shouldn't beat up on IRT(?). There's lots of complicated ways of dealing with violations of local independence in IRT, but the basic model is premised on this local independence system which means that your response to any item is independent of your response to any other item given the underlying trait.

So, given that the basic workhorse of psychometrics is not exactly suited to the problem of temporally complex data, what else can we do with this. This measure here is Kullback-Leibler divergence, kind of avoided the reference distribution. But it's basically to say whether or not the joint distribution of the data is different from the individual time points. The denominator there is what if you just randomly permuted(?) all the events and the numerator is what if we preserve the sequence. Is there any statistical dependence or difference between those two models. And that iteration down to the item level is just a basic application for further notation. We can then treat this as a function of lag(?), so we can progressively move more up the history of preceding a time point and say at which point in time does an event sequence no longer depend on

its past. And then we can aggregate over lags such that the dependence is on zero. That's basically a rough way of talking about the area under the curve.

Instead of looking at formulas, let's look at some graphs. This is an example of Tetralogue chat. This is equation 1 basically binned over the sequence of the events. The gray area is where we do not have evidence of more dependence than would be expected by chance or plus-on(?) process. And the spikes are basically where we're seeing a lot of complexity, a lot of historical dependence in the chat data in this task. So I think this can be used to inform task design, obviously. So, if we're expecting to see more dependence in certain parts of the task, this would be a way of seeing if that actually manifested itself in the data we have. And we can also use this to go back through logs and find out the kind of dependence, the kind of chat that's happening at the points that seem to be dependent. So this process, in addition to providing an index or this approach in addition to providing(?) an index, also it helps us kind of work with task design and think about or find what kind of, what's happening in the event log at the times at which complexity is being registered by this approach.

This is equation 2. That's basically the lag over 24 second bins, so up to about 80 seconds there's some dependence in this event history, and equation 3 is a rough way of talking about the area under that curve, and that could be the total amount of dependence in the data.

At this point, we haven't separated out the total complexity in 20(?) of its components and this is really the main challenge. So the first thing we can do is separate it into the two individuals involved in the chat. Obviously, this could be any other bivariate process, it could be a multivariate process. But the example for dyadic

collaboration would be the purple and the green represent the chat times of different individuals in this context. Something I was working on, I'm pretty sure this has been proven also in the literature, I just couldn't find it, is that you can break down the overall complexity into two parts, one corresponding to each individual, and then you can break down the parts of each individual into one which depends on their own past which is referred to auto, basically an auto-correlation and the part that depends on the history of the partner. So this is pretty intuitive, so you can break down the history in the dyadic interaction into two parts. That depends on each person and each person you can break that down into another two parts, one that depends on their own history and one that depends on the history of their partner. And this involves no assumptions of independence and is(?) model(?) free(?) or nonparametric.

So, in summary, we have four parts. What are those four parts interpreted as, how the actions of a person depend on their own past, how the actions of the person depend on the past of their partner. In the context of how these processes apply to chat, I don't know if you could see the purple in those equations, but that corresponds to two parameters that we can estimate in the model based context. In the context of chat with the ..... process, we're basically talking about the proportion ..... of chats that are in response to X's own past chat behavior. So, when you're chatting sometimes, I'm sure you guys know this, you can send a long iteration of messages. So hi, how's it going, what are you doing, without getting a response from your partner. That iterative non-baseline part of the process is really just due to X, it's not due to Y. Y has to actually say something because X can be responding to that. And so those are the two parts that are

being separated. The interpretation for Hawkes process is given a lot more detail in the references below.

So here's how that looks in the Tetralogue chat. So similar kinds of, I've shown similar plots for a lot of different types of complex dyadic or team-based events. In basketball the lag is something like two seconds. In email it's something like to hours, so chat is somewhere in between basketball and email, if that is insightful for the amount of lag. Also, the cross-dependence is how the two partners are responding to one another. That's symmetrical. That's often the case in interactive processes rather than asymmetrical ones such as Twitter. And the auto-dependence is much less than the cross-dependence which is what we would expect to see if people were actually interacting.

At this point, we have these four parts and the middle part, I should mention, is basically the transfer entropies(?) or the nonlinear granger causality which has been discussed in other contexts. We have these four parts, and we can cut them up any way we want to get some kind of index, but if you're going to take a time series approach and you're going to talk about the individual contributions to that time series, those the four parts you have to work with. You might be able to break them down further. That's basically all the dependence in the process data, so what do you want to do with it. And the first thing I thought is let's take the overall dependence and take each person's, take the total, take each person on either side of the total, call that a proportion of dependence due to each person. I think this was a really good idea at first. It's got lots of drawbacks though. Quite honestly, you can't compare across dyads anymore because you've normed out difference in dependence. You might be able to treat this as

a task design issue, so to have a fixed amount of chat, for example, in a complex task. But it also has other drawbacks so I norm my performance based on that of my partner, we can then come up with kind of tricky ways to avoid actually doing any collaboration and both got the exact same score on the task. So there's two reasons not to like that approach. One approach is to norm over dyads, something like basically just subtracting the mean, and now we can compare over dyads, and this is what's becoming increasingly clear to me when I do these analyses, we don't really have a parameter for the task. So what is the measure of how complex the task needs to be in order to be performed? If we had that, that would be a task-based or criterion reference way of norming.

But rather than taking any of those approaches, here's the raw information index. It's normally distributed. That's nice, human achievement stuff typically has been ....., so is random error(?). It's not really that assuring, but at least it's something we know we can work with. The relation to the number of messages is pretty weak. That correlation is about .3. We're not just counting the number of messages sent, we're not just counting the number of words sent, and the relationship between a partner's index and your index, that correlation is around .5 which I think the parameters themselves are not necessarily correlated. So that I would be happy to interpret at as an empirical correlation.

But this is the slide that's kept me up for the last four days and this is the slide that's about one of the questions is what are(?) the relation to task outcomes. And, in this case, the relation, that correlation is about zero. So we're talking people had an initial responses, they chatted with each other, then they had a chance to change their

response. This is a difference between their initial and their revised responses. And there's no association with the chat. I was really discouraged by this, but Noshir Contractor's talk yesterday where his team effects were kind of a substantial proportion of the task variance. Basically, we're analyzing the team effects with this approach, so I feel like hopefully this is a task issue and we can keep working on that. There's some conclusions and thank you very much.

PK: Our next speaker is Ron Stevens.

Ronald Stevens, University of California – Los Angeles

Okay. Now we're going to switch gears a bit and start looking at the neurodynamics of team organizations. And this work was supported by Darpa and NSF and special thanks for Tricia Galloway, other collaborators, and especially the sailors and the staff at the Submarine Learning Center where this research was done.

What's our rationale? Well, it's brain entrainment in a cross-brain synchronizations. So, if we were to start a signal in this room that oscillated 10 times a second, after a couple of minutes, if we measured your brain rhythms, there would be a major 10 Hz component in just about everybody's head, and this is a well documented phenomena, entrainment of the brain by external stimuli, whether it's visual flashing, whether it's auditory. In 2004, Uri Hasson at Princeton extended this by having a group of subjects watch the segments from the Good, the Bad, and the Ugly. And, as they did, he looked at brain activations in different parts of the brain, and he found that there were pretty good correlations with different scenes across different people watching the same movie at a different point in time. So this suggests that there is something in the

structure of our wiring that responds similarly to events as they unfold. So our idea is that teamwork is a lot like watching a movie. The only difference is that the team members can actually manipulate the movie and have it change directions. It's interaction of one complex system with many other complex systems, including the environment.

So I guess the basic question is are teams entrained by the teamwork task? The task we're talking about is submarine piloting and navigation. It's required for graduation by the officers at the Submarine Learning Center. It begins or the goal is to pilot the submarine safely in and out of a harbor and a sample trace is shown to the right. There's a briefing that lasts 15 minutes where the goal is laid out. There's a scenario of an hour to two hours where the situation involves(?). There's predictable events like marking the ships position every three minutes, as shown by the green squares. And there's a lot of traffic density and environmental changes, and eventually there's the debriefing.

And so we studied 24 teams, and we generally have between five and six different team members with the headsets on, so we're modeling six person teams. And we have sensor locations at 10 sites on the brain, and we collect the EEG into 40 Hz(?) bins from 1-40, but we'll talk more about these frequencies shortly. So you end up with pretty large datasets. So you have these 40 frequency bands x 6 persons x 10 electrode combinations, and one of the challenges is to either simplify it during the analytics or during the visualization. And early on we chose to use a symbolic representation of the team rather than working with six numeric streams. And so what we do is create snapshots, as shown in number 1 in the upper right-hand corner. You

can see this. So this shows what the state of the team member is at one particular EEG frequency at this second, and team members 2, 4, 5, and 6 are expressing high levels. Team members 1 and 3 are expressing low average numbers. By collecting these across the whole performance, we can create state spaces which show the most representative combinations of this EEG marker across different members of the team. For simplicity, I'm showing you 25 here. We have used up to 400 different symbols in other experiments. So, in the performance then, you end up with this data stream of these 25 symbols, and the idea is that these have information about the current status of the team, the past status of the team, and we would hope that something in the long memory process would allow us to predict future states of the team. These data streams have temporal structure, as shown in the lower right-hand corner. If we plot this symbol at one point in time with the symbol at the next point in time, we get a diagonal. So there is some persistence in the expression of these symbols. And, if we randomize the data, as shown to the right, the structure disappears, and we use randomization of the symbols doing anything to them as a control.

We can then take these symbols, as shown on the left, and so what we're doing is plotting the expression of the 25 symbols over a performance which has a brief scenario and day(?) debriefing. And you can see that the expression is discontinuous. Here in the debriefing, there is a whole series of symbols that are not expressed that were expressed earlier on in the scenario. And the same thing at various points here, there are gaps where some symbols aren't being expressed. So what we do now is take the symbol stream and then calculate the entropy over 100 second moving window, and each second we add one more symbol, take another one off. And the idea is that data

streams with a lot of mix of the symbols will have high entropy values and places where there's more than normal persistence, then we're going to see downward fluctuations as the entropy decreases. And this is the plot of the entropy for this one team performance here.

Now, this is an example at one of the 40 different frequency bins. So now what we do is we take all 40 of them and sandwich them together like this and then look down from above, and now we can get a brain wide view of the performance over at the different frequencies.

So, at the right, we have our 40 different frequency bins. This is the time of the simulation, and these dark contours are valleys. This is where the team is undergoing increased synchronization or organization, and we can relate these to events within the task. So, for instance, at 10 Hz, there's this intermittent band that lines up nicely with the periodic component of taking the ship's rounds or lining up the position of the ship. There's an increasing dark patch from about 1600 up, and this was when the submarine was left of track and right of track, the soundings weren't matching up, the GPS was going. Any of these individually would not result in the problem, but combined with a little set in drift, it resulted in a grounding. And so then the simulation was paused right around this area. There was a discussion with the instructor and then it picked back up again until debrief. So, with these types of team coordination maps, we can look at points in complex scenarios like this where the team was going into a persistent stage and we can tell which frequencies that this is actually occurring.

This is similar to the drawing at the bottom. It's plotting the frequency versus the organization for the debrief scenario and briefing. For reference, delta is involved in

suppression of external stimuli. Theta(?) has to do with your spatial navigation in your own personal space, memory. Alpha is where there's a lot of social coordination markers, and we see a lot of synchronization in this area. Beta is in the pre-motor cortex area, the so-called mirror cell. It overlaps somewhat with the mirror cell activity. Then sliding down into the brief and scenario, it's more in the gamma which is more representative of voluntary and involuntary attention.

So now we're able to correlate these activities with an instrument that the submarine floors(?) has developed, evaluating a team resilience. And I'll just say it goes on a scale of 1, 2, 3, 4 and we can talk about it later. But the team synchronization is correlated with team resilience, and in the briefing, the most resilient teams were the most synchronized. Whereas, in the scenario, the most resilient teams were the least synchronized. So there's something in the briefing that may be important, and there's something different in the scenario that may be important for distinguishing teams with higher and lower resilience.

And we can go down and actually break it down now to see which frequency bands is this occurring in. And, in the briefing, it is primarily in the 25-40 Hz band which is where there's voluntary and involuntary attention. So maybe paying attention during the briefing, maybe teams that do that will have a better outcome and the submarine force is going to look into this. In the scenario, the low resilient teams were most organized in the so-called pre(?) sensory motor cortex where you're looking at other people and anticipating their motions and moves, and it may be in the novice teams there's a lot of looking around as people try to see what's everybody else doing, what are we supposed to do next? And this may be absent in the more resilient teams.

From this and other studies, we're beginning to develop a framework for teamwork here which compares outcomes of their performance versus cognitive organization which is our neurodynamic measures and versus flexibility. And so very, very novice teams are random. They're completely disorganized, and on the other end of the spectrum, teams which have very, very high organization also have poor outcomes. And this is probably because they're thinking in a very rigid state. Expert teams, up at the top. So they have the flexibility to deal with new situations but when crunch time comes, they can drill down and get organized and handle the situation.

So the what, where, when, why, and how of team synchrony, the conclusions generally occurred when teams need to focus. The frequency, magnitude, and duration is different with task elements, situation complexity, and team resilience. The why I won't get into because it's more theoretical and we can talk about this offline, but it has to do with the interesting relationships to the right where low entropy equals greater synchronization equals more organization equals more information transfer among members of the team. Thank you. There are some references and email, and you can find movies of some of these things on our website at teamneurodynamics.com down at the bottom. Thank you.

PK: Thank, Ron.

Yoav Bergner, ETS

So this talk is kind of a methods talk but I'm not going to get into really dry details of the methods. I'm going to try to present a framework and show an application in two really different ways to the same dataset which is kind of collaborative. It's a peer

tutoring interaction. I hope that this sounds like somewhat of a response to Eduardo's observation number 8 which is that people do different things at different times in teams. And so, to some extent, what we need are models that account for different things happening at different times, and these are simple versions of those kinds of models.

I'll say a little bit about hidden Markov models and how they fit into dynamical(?) Bayes nets, in case people don't know much about those, and then I'll talk about the dataset, the Automated Peer Tutoring Assistant which was developed by Erin Walker who's a collaborator, and then I'll present these two different models. And they're quite different models in terms of how they(?) interpret the data.

So the first thing about hidden Markov models is that they're getting kind of old. They've been around for a long time. They model discrete sequences in terms of some hidden state that undergoes a Markov process, and that means that the state at a certain time only depends on the state at the previous time. So getting back to people doing different things at different times, within this framework of a hidden Markov model, it only depends on what they were doing at the last time, but it can sort of present an evolution. If a finite mixture model is familiar to you, then you can think of hidden Markov models as a dynamic generalization of a finite mixture model that kind of has this Markov property. And a fairly modern view, certainly not the one in the 1950s, of hidden Markov models is that they're special cases of Dynamical(?) Bayes Nets. And what that means is that they can be extended in lots of ways, one of which I'll talk about but one which I won't talk about is they can be extended to non-discrete states such as continuous states. For example, the Kalman Filter can be seen as a Dynamical Bayesian Network.

They've been applied broadly in the field of computer science and been very successful in speech recognition and computer vision. And there's this kind of attempt for the artificial intelligence to recognize some sort of hidden message through observable things like sounds and pixel maps. But, in education, they're also starting to be applied, and one of the better known examples of the use of Bayesian Knowledge Tracing and Intelligent Tutoring Systems where the hidden state is the mastery level of the student and what is observed are their correct and incorrect opportunities in practice problems. And Bayesian Knowledge Tracing is actually the kind of inspiration for one of the models. Knowing sharing in groups, so specifically applicable to the kinds of things we're talking about today was illustrated Amy Solar(?) and Ron Stevens who's sitting to my right, so a while ago. Russell Almond has talked about Markov Decision Processes which are quite related, and Christie Boyer and collaborates at NC State have looked at tutoring styles which is also related to one of my models today.

So let me talk about the dataset. So what happens here is these are pretty much high school age kids. They come in to do an experiment a few hours long, and it involves using an Intelligent Tutoring System for algebra, but rather than having the individual student work in front of the intelligent tutor, which is actually what it's normally designed to do is to provide them with help, they don't get help directly from the tutor, but they have a peer tutor who can see what's on their screen and they have a chat window. And they can get help from the peer tutor through his chat window. The peer tutor actually has access to hints(?) from the tutor which solves the problem of the peer tutor not knowing how to solve the problem. And the design of the experiment was really to help promote peer tutoring behavior. So that's what Erin Walker was doing. That was

what she was coding for automatically to provide the peer tutor with feedback about the value of giving more elaborated hints rather than kind of un-elaborated help. So they were coded to trigger that kind of feedback. This is sort of this work as a secondary analysis of that.

So this is a picture of the first model, the one that I said it kind of inspired by Bayesian Knowledge Tracing. And this isn't actually a vanilla hidden Markov model. This is called an input-output hidden Markov model. It has this input layer which is shown on the top, and the inputs and the outputs are observed. But in the middle there's this layer that isn't. So the treatment of the use of this model in this application is that the peer tutor is seen as providing inputs and the peer tutee is seen as doing things which are outputs. And somewhere in the middle is this idea that the tutee's capability is being altered by the utterances of the tutor. Now, this really doesn't work the same way as Bayesian Knowledge Tracing, if you are familiar with that, because that's kind of a long haul thing. Actually, Vincent Aleven showed some examples of learning curves in tutoring systems yesterday. But the data for this model is sliced very small between times that the tutee actually makes a mistake and then corrects the mistake. And so you could say what this model is trying to do is figure out which kinds of things that the tutor says ultimately help the tutee overcome those obstacles. The outcome of this is an assistance parameter.

So this is how data gets sliced into this model. So there's this interaction where on the left side you see what the tutee does and on the right side you see what the tutor does. And so the tutee might make a move that is automatically recognized as an incorrect step. And, by the way, this isn't a final outcome. The steps are scored in the

tutor, and the tutor says you did that last step wrong. You need to divide those sides by R+B. So you see that actually there's a lot of telling going on and that's kind of important in the generalization of how useful this was in this application. So no, I didn't and then they undo, and then the tutor says, "No, listen to what I'm saying," and there are these codes that are on the right that are things like help after incorrect and starters with no help. If you see starters, that refers to scaffolding that was put in place to try to encourage the peer tutor to press buttons that say like ask why or explain. There's a lot of evidence in the literature that you need to kind of scaffold computer supported collaborative learning to get conversations to be on task. Anyway, all of these things get coded into that sort of input layer that you see on top. And the output layer in this case is just observations of correct and incorrect. And what I want to say is a couple of things about the results from this model, nothing about how you actually estimate it. For the most part, the assistance results from this model were kind of self clustered into two groups. There were sort of assistive and not very assistive. I mean, it could have been anywhere on the continuum but they fell into kinds of hits and misses. And they were mostly, when you looked at them, they seemed like those made sense, things like no prompt after misconception is not helpful. But there were a couple of false positives and false negatives that we were able to uncover why they happened, either because of issues of missing data where two things always occurred at the same time or because the codes themselves that were automatically generated turned out to be wrong. So we actually were able to see that the model said, well, this looks helpful and we looked in the codes that said that's not helpful, and we saw lots of helpful things. So that helped encourage us to keep looking.

The other good result, and it's a little hard to unpack this, but everything above the dashed line is kind of what I'm calling hits and below these dashed lines is what I'm calling misses. And this is an attempt to compare students who did very poorly on the pre-test compared to students who did well on the pre-test and see whether they have differences. And there were two differences, and these are what they were. And this was a little bit surprising but it kind of made sense after the fact. So high level help means elaborated as opposed to saying just divide by K, dang(?) it. So why is it that the bottom half of pre-test scores benefit from high level help and the top half done? Well, if high level help is kind of telling you you're supposed to solve for X and I don't even really know that, then it turns out to be important. High level help isn't that common in a dataset but it does happen, and it turns out that it's actually more important for the students who came in basically scoring zero on the pre-test. Help after incorrect, one explanation of it is that if you're basically getting everything wrong because you have no idea what you're doing, then it's not particularly helpful to have that pointed out to you, you made a mistake. But if you sort of know what you're doing and you make a mistake every once in awhile, you have the basis to sort of construct knowledge on top of that with feedback. So help after incorrect was useful to the top scorers but not so useful to the bottom scorers.

So sort of the main findings I think I talked about them and I'll just skip over the summary slide and talk about model 2 which is quite different. So, in model 2, instead of thinking about the tutor and the tutee as inputs and outputs, we want to think about both of them as kind of parts of a system. The system itself undergoes a state that evolves and we don't have direct access to what that state is, but we have access to

observations that come from the tutor and the tutee. And one of the things that we were corrected for in model 2 is the fact that we had those miscoded chats which were going to mess up our inferences, and so these are all using humanly coded chats for what the tutor said. And they were actually coded not just for cognitive characteristics but also affective characteristics. So you see where it says, “No, listen to what I’m saying,” impoliteness and rudeness were actually now tagged by Amy Ogen(?). And we’re still only observing things like incorrect moves and undos on the tutee’s part, but we’re observing their ability to talk back. Those weren’t coded for affective characteristics, so that’s why we don’t have more fine grained subdivisions of those. But this is actually a more vanilla model in terms of the hidden Markov model.

What we want to do with this model is use it as a classifier because there was a pre-test and a post-test, and we want to see which dyads that work together ended up kind of really gaining something after the three hour experiment on the post-test. So we train one of these classifiers for the low groups and we train one for the high groups, and this is very kind of related to work that Amy Solar and Ron Stevens did and we then used those two classifiers to try to take a new dyad and predict whether they’re going to have gains on the post-test. And it turns out that it does quite well and it does quite well compared to really naive models where you take all of the things that we observed, all those codes like rudeness, and corrects, and undos, and you just count them all for that dyad. There were hundreds of observations, and you say what if we build a giant kind of multiple regression sort of logistic(?) model and that doesn’t do much better than chance. If we do the best you can with aggregates which is doing kind of some forward, backward, stepwise, AIC based model selection and then apply it, we do better. That’s

the best logistic. But this hidden Markov model which takes into account the fact that things follow in a certain sequence. Certain sequences are observed, not just overalls, is much more predictive. And the classifier turned out to have eight hidden states, and I just want to briefly flash this. This is the kind of thing you then try to interpret at the end and it's hard. But you see large numbers and large numbers tell you some things like persistent states or persistent oscillations between state 4 and 7. And then when you want to know what is persistent state 5 or this oscillation between 4 and 7, you have to go to something called an observation matrix which looks like this. But it turns out that you might be able to make sense of this and that persistent state 5 is undo. So, for example, the model discovers that when a student undoes, they're likely to undo a lot of times, and so the system is like, oh, we're in an undoing state. And the other state was actually an off topic chatting state, so where the tutor and the tutee are getting off topic and they're getting into a conversation and that kind of likely to persist in this oscillation.

So there's still a lot of work to be done to kind of make more interpretability out of it, but this was a little bit of a proof of concept using these two different models, one of which really treats the tutor and the tutee as input and output and one of which treats them as parts of a holistic system. In terms of future work, we'd like to improve the quality of data more and because this maybe relates to some of the questions that we're supposed to answer in the panel session, I'll sort of put that off as to the kinds of things we might do to kind of score process data. So thank you.

PK: Thanks, Yoav.

Jiangang Hao, ETS

Okay, thank you. My name is Jiangang Hao from ETS. Today I'm going to introduce our project called Tetralogue. This is a project that was designed to assess the CPS skill, and I think Alina yesterday and Peter just introduced a little bit about this project, and I will give you some more details.

Before I start with my own slides, I would like to share with you one note I got yesterday. This is from Eduardo. "What's good for research is not necessarily good for practice." I think that's very important, and when we talk about the complex things like CPS, we need to be very clear about what can be done in a more operational stance(?) and what is the ultimate goal, to understand the underlying mechanism. So this will help us to structure our plan and structure our resources.

As I'm from ETS, I would have think things more from the educational testing perspective, so I normally have ..... question. What's the probability you won't get sued if you report a CPS score(?). So whenever I have an idea I always think, okay, so if I eventually report a score, what's how likely I got into trouble. So I think this will be helpful for you to think when you are trying to plan something.

Some practical considerations for CPS assessment, we care about a few things. I list three of them here. The first one is repeatability, so that means how stable and repeatable the results you report are. So this is actually the golden rule of science. So, basically, your results need to be repeatable. And also we care about the generalizability, so whether the findings based on one task can be generalized to other situations. For example, you've got a lot of findings about the team interaction or collaboration based on basketball. And can you apply those findings to football or

baseball. So, if you cannot, then you have to make it very clear when you publish the results that your findings only apply to a particular subclass of all possible collaborations. So I think I'm trying to classify those kinds of different tasks so that we can find a concrete result within each subclasses. And one more important thing I need to know is that we have limits in time if we want to do a real assessment. You cannot ask a team, a test taker to take a test to assess CPS for a year.

There's different levels of CPS assessment. So the first level is a group level assessment. Basically, we do not care about the individual teams. We do not care about the individual persons. We care about statistical(?) properties of the CPS from many teams. And then this will depend on the member's properties and also we need to assign the members into the team in a random way. This kind of assessment is very typical for, for example, NAEP. And therefore, the individual level assessment and suppose you have a CPS assessment for videos(?) and you give them a CPS score. And if they got a high score, probably everything is fine. If they got a low score, they will have a few complaints. For example, maybe they'll complain, okay, the task sucks and it's not suitable for me. For example, it's a football-based task. Maybe you say, okay, I'm not football player, I don't know or a second complaint they may have is that my partner sucks, very easy. "My partner just screw up the collaboration and we cannot get any meaningful results." So the solution, one of the possible solutions is to use different tasks and also switch different partners, so that we can map out the CPS in a grade(?) of these two dimensions. An idealist(?) will be a uni-model(?) distribution but more likely it's such ..... stuff(?).

The CPS is very complex, but given relatively well designed tasks, there are some things you can directly measure. What are the directly(?) variables(?). The team response for each item in the task, individual response for each items, and also we have time stamped communications during the collaboration, the ..... video, audio adjuster(?) and something else. And also we can gather team members' other information outside of the task, for example, their personality, hobby, and their knowledge, and the skill, whatever you can name. All the story about CPS are essentially based on this directly(?) variables(?).

To start with some actual collaboration(?) of all these kind of questions, we started this Tetralogue project since last year. The goal is to try and explore the possibility of marrying cognitive skills and CPS skill in a psychometrically rigorous way. And we have a lot of members(?) in our projects, and I highlight Alina, me, and Aladio(?), we are the major person who work on this project for a longer time. In this project, we are developing a simulation task to assess CPS, and also we are aiming at getting a seldom(?) ..... responses from the Amazon Mechanical Turk.

Here are a list of the major key research questions. For example, we want to identify the CPS construct that is reliable enough for psychometrically rigorous assessment. And also the amount of prototype from game-like environment for assessing CPS, we want to find out the relationship between the cognitive and the social skills in CPS task. I will not repeat each of them and let's go out to the next one.

So to address all these kind of research questions when you have a relatively broad experiment design. So in our ..... project, we have five main components. The first part is two simulation tasks about volcano science. One of them is a single player

version. Another is a CPS version. So, in the single player version only one person complete a task, this as a baseline model. And then we have general science knowledge task, consist of 40(?) multiple choice and we have questionnaires for the demographic information and the questionnaires for the personality information. And also we have after collaboration survey. So the data collection is through the Amazon Mechanical Turk, and we ..... getting(?) 1,500 total participants. They also take a B, C, and a D and 500 of the participants take the single player version of the simulation, this is a baseline model, and another 1,000 participants will form teams of dyads to take the collaborative version.

So, here, I give you a screen shot about the simulation. The left-hand side ..... the single player version ..... the CPS version. These are all based on previously developed simulation called the trilog(?) out of ETS. We add this additional collaboration layer and also we make them lab-based, so that we can collect data through them on Turk. So, here, this is CPS version. There's the two major windows. These windows ..... 16(?) prompts. Basically, we supply some facilitation information for the two participants to complete the task. And then we have this chat history. This basically is two people will communicate with a chat, and all this will be recorded into a log file and then later on we can analyze their conversation.

To get more information from this task, we have a highly scaffolding(?) facilitation. So, for each item in the simulation, the two participants are prompt to respond separately first. And then after they respond separately, they are prompt to discuss with(?) each other. This is the collaboration part. And after the discussion, each participant will be given an opportunity to revise their initial response, and then finally

one participant will be chosen as a representative to submit the team response. So these are facilitation of the collaboration.

Most of their collaboration is through their chat message, and we have a framework to classify the conversation into four different categories. This is developed by my colleague, Lei Liu. She is also in the audience today, and she, based on the PISA and also the CL/CL(?) research, developed this kind of CPS skill framework. We use this framework to classify the conversations into different social skill categories. So here's the project timeline. We started last year. Last December we got 500 responses and by now we got 400 responses from 400 dyads. And also we have this scoring Rubrics developed and the data reduction pipeline, some preliminary analysis.

So here I show you some results, very interesting results. So we have, in the simulation tasks, the first seven items. They are selective response. That allow us to score them in automatic way, and the y axis here is a sum(?) score. And the group 1 is a single participant from last year's data. These people, they take the single player versions ..... the baseline population. And the group 2 is the team individual initial. That's the initial response of each of the participants in each team. As you can see, this initial response is almost ..... the same as last year's individual versions of response. So that means we can use this initial response as a proxy of the individual work. And then the interesting points, group 3 and group 4. Group 3 is still the team individual response but is based on their revised answer. So after discussion, something happened so that their performance has been improved. And the final one is the team response, the response from the team representative. We can see, if you compare this one and group 2, there's about a 4.5 ..... increase. So that means the collaboration do have some

effect. And frankly speaking, I've been worried about this for a long time if this points(?) actually here(?) what we can do. But, luckily, the kind of scaffolding in terms of facilitation and also the task do elicit more collaborations during the task, so that you can see there's a positive effect.

Another one is about the collaboration process, and here the Y axis, members'(?) words in the communication. So the members'(?) words(?) can be a proxy of how collaborate they are but they may not be a good indicator of this. The X axis are different topics within the simulation, and, as you can see, there's a clear change when there's collaboration going on. At certain point they have more communications in terms of words and some places have very few of them.

So here I show you one interesting conversation. Before I go to that one, I show you this first one, so introduce each other. At the very beginning, they need to introduce each other, and most of the team, people do not spend a lot of time on this. But this particular team, they spend a lot of time. The one I'm going to show you is corresponding to list(?) here. So I'm particularly pleased by these two sentences. "I was just about to ask what should we do, click next?" Another person said, "I will wait a bit." That's so ..... and also they start talking about something else, about have you ever seen a real volcano or something like it. Here is one animation to show the positive(?) on(?) active(?) outcomes. The Y axis is defined as the revised response, subtract, minus the initial response. So, if their revised response is higher than the initial response, that means the collaboration has some positive effect. And, as you can see, on different teams they follow different paths. And here, each of the frame whose(?) animation is based on the total, if you sum all this together, this will be the total increase

or decrease of the change. And for different change, you can see different patterns of this revised response minus the initial response. These are just the preliminary analysis and we are now in the process of modeling this process.

So what's next? Next year we are developing a platform or frame(?), so that we can plug in a lot more other tasks easily and also we can make the facilitation more adaptive based on what people type in. And also we will enable some multi-modal(?) communications. This year in the task, we only allow the task communications but we will also allow video audio chats. So regardless(?) of CPS frame, they can help us to test a number of different tasks and also switch different partners easily. That's basically where we are right now. Thank you.

PK: Thanks, Jiangang. Okay, that was a very interesting set of presentations. We heard a lot of very nice ideas on new methods, pretty extensive use of technology, different technologies. Plus, there's an incredibly diverse set of presentations which I think is in keeping of the theme of this meeting even within this one panel. I think that a lot of us, as we're watching through these presentations, might have thought about yesterday's presentations and how the applications of some of these methods potentially is very promising. So I wanted to just open up the Q&A with the panel now with this set of questions. So the first one is, and I think all the panelists really had a take on this first question, how do we score process data. That's basically what it was all about. I want to just go through each of the members and let them say something they might not have had a chance to during their presentation. But, really, for the audience what the take home messages are in terms of what did you do with the process data, where is it right now, where do you think it needs to go, and at what point

can we transition these methods off to Eduardo and other potential users. Let's just start with Peter.

PH: Sure. So I think what I've been trying to do with the process data is find some way to, if there is information in the sequential aspect of the data, then we need a relatively generic way of describing it to apply it to different types of tasks. And I think Yoav's demonstration nicely showed that there is some added benefit from having that sequential information. How can we use that to describe the performance of an individual or multiple individuals, or multiple facets of the task as they're unfolding? This is the opposite problems with aggregating from one team up to another team which I think is the relational events systems approach. It's how to go from one team and kind of peel that back down to the individuals. That's the basic conception, the basic goal of what to do with this process data. And I think this can basically be done right now, but what the shortcomings are is exactly which parts of a complex task are the parts that are relevant for understanding the outcomes. And this is where the teamwork data that we heard or teamwork literature that we heard from yesterday in the first panel, is about 40 years ahead of the assessment community. So we clearly have a lot to learn in terms of basically, what kind of covariates do we want to put on these events and how are we going to, what a richer description of the kind of event process. Given that we have a richer description of the event process, we can develop tasks that will elicit those events and then we can quantify the dependency in the process data. So I think it's ultimately a question of task development and then relating the types of process data that we're interested in to the types of outcomes that we're interested in. And that's question, I think, 4. So I won't go into that.

PK: Before we get to you, Ron, because yours is a little different direction, let's hear from Yoav on that same issue of how do we score process data.

YB: Okay, thanks. So I'd like to see two things happen at the same time and informing each other which is sort of the top down and the bottom up approach. So I think from the top down sense I think we have a lot to think about from organizational psychology and work on what are the markers of a collaborative process that we want to look for and design a task and score it and so on. From the bottom up approach, there are these kind of in-the-wild collections of data, and if you can look at those processes and try to explore the data and understand what is it that different processes have in common. Do they sort of self organize? Do they cluster? Are there features that we didn't know to look for because they weren't sort of theoretically given to us in the top down approach but we find them in the bottom up approach? And then these two approaches I think need to meet in the middle and sort of work back in this iterative cycle. So that's sort of more of my thoughts about this are kind of like how we as a whole multi-team system can work together and inform each other from both directions.

PK: Jiangang, if you want to weigh in on the process data scoring.

JH: So, basically, if you want to describe the process data, it's really when you design the task, you need to have a sum(?) scoring plan about how are you going to score this data. So, in the CPS task, there are two types of scoring you can do. One is social skill, another is the cognitive skills. So I think for the social skill we need to have certain scoring Rubrics and then we have human raters(?) to read them, and based on the human rater(?) score, we can train some automatic scorer engine to do the scoring.

But, for the cognitive part in the task, there's a relatively clear traditional standard to say, yes or no or correct or incorrect, 1, 2, 3, 4, 5 levels.

PK: Okay. Ron?

RS: Well, right now we're still trying to define what our process data actually really is. And when we started the project, we were looking at over timeframes of a minute, something on the order of that, much like Steve, you already talked about with guitar duets, short-term changes. Then as we began modeling, we saw these very large differences when the task changed at the briefing and the debriefing. And so now our analysis of process data jump to about two hours, but yet within this two hours there's still shorter segments where we see these entropy fluctuations that go from anywhere from two minutes up to 15 minutes. The more recent data where we start to correlate it with performance has now given it another jump. The last slide I showed I didn't describe, but the X axis was 10^7 seconds, and that's about how long the SOAC training course is at the Submarine Learning Center. And that scale extended all the way from one or two seconds where there was decision making, up through communication, up through larger dialogues and so forth, through tasks, through series of tasks. And so now there's the possibility that team synchrony at a neurodynamic level may actually be able to scale over pretty long periods of time and may actually have some training significance over long periods of time. So that's the challenge where we're at right now is what exactly does team synchronization mean, so that's currently where we're at.

PK: Great. I want to turn to the next question. The next question really has to do with the design of the tasks and the analysis and what comes first, basically. And so all of you have been involved, in some sense, in an analysis kind of a process. And I

guess a question is how should we design tasks to get data back that will be more informative in the analysis task or is it a little bit of back and forth where we do some analysis and then we do some design, we do some more analysis and that informs a design and it's iterative(?). I guess the question is how do we best extract data through a design process that's most informative and moves the science forward the most. And maybe we can start, Ron, with you this time.

RS: Sure. Mine's short. We don't design any tasks at all. We go into the wild, so nothing to comment there.

PK: All right. Peter.

PH: I think that task design is, I'm in favor of a top down approach there because clearly defined tasks make it easy to do the statistical analysis, basically. So, when you're analyzing basketball, you don't have to figure out which components of the basketball game are the relevant aspects of the process because the game is pretty well defined in that regard. So things like passes, rebounds, shots. At least on the offensive side of it you don't have to really grapple semantically with what are the relevant parts of the process. If you don't have something like that to hang analysis on, then you really are just taking a shot in the dark which is a little bit of what happened with un-filtered chat data in the Tetralogue, at least in the analysis I ran without screening based on were they having technical issues which showed up or were they actually on task versus other stuff. So not having enough task design is a frustrating position to be in for an analyst, that's for sure. At the same time, the tasks have to be designed to reflect the interests of the people who study and theorize collaborative problem solving and the people who require(?) those skills. I think as a thing that came

up already, it's a collaborative endeavor, so there's multiple skills and there's multiple groups of people that need to be involved to make these tasks work. And yesterday we heard a lot of interesting stuff from the task designers from PISA 2015, and you get the impression that that's a pretty sophisticated part of the process. I mean, once those two things come together, then the analysis, I think, can be a lot more fruitful.

PK: Okay, thanks. Yoav.

YB: So, in the example that I talked about I think is a good example of not using evidence in a design and then thinking about how you can improve it. I don't want to beat myself up too much about it. The purpose of the experiment was originally to do something else, and we're sort of doing an alternative analysis. But in that peer tutors, it was actually an interesting collaborative dynamic, I think, because you're trying to get students to learn from each other. You want them both to learn and to learn to be better peer tutors. So there's a good reason for doing it. There's a good motivation for doing it. But if you just let them do it, they fall into a lot of behaviors that we don't think are productive like just telling each other what the answer is. And so that you realize is happening and it's not only compromising your analysis, which is important to me, it's compromising the whole point of the thing. So think about designing a way for them to not do that. Just one way to do that is incentivizing it in a fun way like, you know, the tutor will earn a certain number of points if they lead the tutee to the answer without telling them, but, of course, over time, to accumulate more points, they do want them to get other opportunities, so if they're really stuck, you can tell them the answer. You won't get any points but you'll get another chance. So we're talking about how we would design this experiment again to elicit the kinds of behaviors that show this positive

interdependence between the tutor and the tutee and then we can model that. So I definitely think it's an iterative process, and it has to be an iterative process. And the saddest thing about some of these kinds of experiments in social research are that they don't iterate.

PK: Great. Yeah, thank you. And Jiangang, you're going to embark on another round of data collection with the science tasks. Do you have some ideas on design?

JH: Yes. So, in general, when you do any kind of scientific research, definitely design should drive the analysis, not vice versa. And that's in the situation that you cannot design experiment in a way to facilitate your analysis. For example, if we want to study the stars in our universe, and basically we cannot design them. They are there. We have to use the available analysis methodology to drive our design of the experiment. But for the assessment, in general, we have the actual control. We can basically control what we want to assess and then we can design everything. So, from this perspective, I think if we can control the design in a way so that the analysis is very simple and straightforward, why should we try to design a task in a way that analysis is(?) very complicated and involves complicated models. You can solve the problems with a single analysis. Why you go to very complicated modeling. The fancier modeling won't make our ..... problem significant. I think this is a very interesting thing.

PK: Yeah. Sometimes we're stuck with analyzing an existing system or an existing universe of stars or whatever it is, and sometimes we have these choices. In the cases where we're going in the wild, as you put it, others put it, in cases where we're going in the wild, is there something that can be done to augment the wild? Is there some design elements that we can put into that situation that help us draw more

lessons about teamwork and so on or are we just basically stuck with? Yeah. Go ahead, Ron.

RS: Well, we were very fortunate that the submarine, the span has three segments: a briefing, a scenario, and debriefing. And the original temptation was that all of the action was going to be in the scenario, so let's focus data collection on that. Instead, we collected from the time people walked in the room and we could get the headsets on till they walked out. We model over the entire performance, and what this does is it gives you a built in baseline level of the highs and the lows of normal conversation of high stress. It's all contained within there. And so subsequently when we go into a new environment, we always want some type of just chatting, or pauses, or things like that. So the task should have an inherent structure to it, and you should try to use that as much as possible. The other part is just hard work. You've got to get transcript logs from six people. You've got to code them up and make the associations. Jamie Gorman at Texas Tech is going through and analyzing the semantic content of a lot of these transcripts where we're seeing changing dynamics. And that's just a lot of hard work and there's no way around that at this stage.

PK: Okay. Anyone else want to weigh in on that one? Okay. Most of you have some control over the data that you're looking at. Let's switch to question 4. This is the issue of, because I think we already covered evidence-centered design in a way, unless someone else wants to take a crack at that.

PH: Sure, I'll just jump in there for a second.

PK: Okay, yeah.

PH: This is something that also came up yesterday, but designing the task to elicit a particular type of evidence for a particular type of construct is obviously an important thing to do before analyzing the data. But one thing I forgot to mention, I guess, between 2 and 3 is that at this point, we don't really have characteristics to describe the tasks the way that we do in a, like we have item parameters and item response theory. We can say that this item is difficult in the sense that people with a certain level of ability answer it correctly 50% of the time. And we can say that's relatively to a ..... population. That's in a characteristic of this task. And in a collaborative framework or in a process framework, it's not really clear what those parameters are supposed to be. And if we don't have them then what are we really measuring? Like what's on the item side of it I guess is really the ECD part, I think. Anyway, I just wanted to throw that in there.

PK: Great. Suggesting that we're limited with our current standard use tools for testing and test theory and item response theory and it really will require some new tools and some new ideas on how to characterize these contexts in collaboration.

PH: Yeah.

PK: Any other comments on that, Jiangang? Okay. So the next question has to do with in long process data, what's the role of being conscious and unconscious in teams and tasks. For example, if students are consciously ignoring something versus unconsciously ignoring things, is there any sense here in the data that you looked at that there is a distinction such as this kind of conscious versus unconscious distinction? Anyone?

PH: That one's for Ron.

YB: Well, it's obviously for Ron so he should answer last.

RS: Yeah. Over the last couple of months we've been trying to think about this idea of team consciousness, and is there such a thing as team consciousness and how can you measure it. With individual consciousness, one of the breakthroughs was when you would give these stimuli which were barely recognizable but it gave a brain signal that was registered, but consciously the person didn't really register or couldn't verbalize that they actually saw something. And this started the whole bandwagon on individual consciousness. In our experience, we've had a submarine simulation where there's a fishing boat out front. The officer on the deck sees it, the radar operator sees it, the scope operator sees it, and they run it over. And this to me would be an example of an unconscious team, that at no point in time was it registered. We've also seen this in high school students with map tasks where they're trying to map up markers and draw lines around markers. And sometimes there are duplicates on the giver's and the follower's map and sometimes they're offset. And we've seen times when the teams are just going along like gangbusters and completely going around the wrong markers. And so the giver isn't correcting the follower or doesn't know what the follower is doing as they go through this maze of landmarks. And that would suggest to me that the team is not conscious about what they're doing. Where does this fit into situation awareness, team cognition, team macro-cognition? Jamie Gorman had some interesting work that are reported at HFES on speech. There was an abstract there, and I won't go anymore into it but if you can get hold of the abstracts there, it's a speech perspective of the same type of thing.

PK: Okay. Anyone else?

YB: I thought Ron was going to say that he has direct access to the unconscious so it's really not an issue for him because it's inobservable(?). But I was going to say, and this is speculative, but I think that those of us who aren't doing neuroimaging at all might still some access to the unconscious insofar as Vincent talked about eye tracking and eye tracking is sometimes telling us things that the student is doing without knowing that they're doing it. And I think Art would probably agree that textual linguistic traces that they leave behind can give us information about things that are unconscious. So, if my tutorial dyads are talking about being hungry and annoyed at this experiment that's going on for a long time, that is conscious about something in their affective state but it's unconscious in terms of how willing they are to keep plowing through a math problem. So, by working together with people who can really do deep NLP and kind of like experts in multi-modal analysis, then I think we can make some hay out of the sort of distinction between conscious and unconscious.

PK: I'm also reminded of Saad's presentation yesterday where he talked about a lot of the kind of, the so-called ..... signals that we send and communicate through, another kind of form of unconscious behavior. And a lot of you are dealing with communication through a chat window and a question is, is there a difference between that kind of a communication where some of those signals are not really apparent versus maybe a face-to-face version of the same tasks that you're doing. Some of the data collection, I know, for example, on the Tetralogue is more or less for convenience sake. It's easier to collect data on Amazon Mechanical Turk where everybody is alone in their room joining you. But would there be a different situation if the two people were

together, and might there be some communication there that we don't see in the chat windows. Jiangang, you might have.

JH: Yes. What Pat mentioned for the Tetralogue when you text(?), that's for convenience, that's right. And we do not use the other model multi-modal things. There's also other reasons. The multi-modal things, they will bring in new information, but also they bring in a lot of confounding factors. So, for example, you can see each other, then the gender play a role and how the ethnic groups may play a role. And the accents of the pronunciation(?) with(?) English will play a role. So there are so many confounding factors. I think it's kind of a trade-off. I think there should be a kind of three spot between additional information you introduce and additional knowledge(?) you introduce. So, right now, in the past few months, a colleague, Saad and a summer intern, Diego, and they work on using the Tetralogue task. The same time when the two player participant work on the task, they also ..... their facial expressions and also their voice. So they are still analyzing the data and hopefully we are trying to see what kind of a trade-off we can get.

PK: Great. Thank you. Anyone else? Okay, I want to just give everyone a chance to kind of recap their assessment of their correlation between process data and outcomes. Some of you mentioned it during your talks but it went by quickly. But I think that's a really important point. What is that link? Is there a link? In Peter's case, I don't think there was one.

PH: No.

PK: But is there a link? And, if there isn't, tell us what you think that link could be with more data collection. There must be a link. There must be something about

what people do engage in when they're working together that relates to how well the thing turns out at the end. You can speculate, too. I mean, you have some data that you looked at and some speculation. Yoav, let's just start with you.

YB: Well, I mean, yes. I think there has to be a link and we found a link insofar as we could sort of predict the better performing post-test takers in this example. But I would actually caution that we don't want the link to be perfect either because then we're not really learning anything new. There have to be different ways of cracking the same nut, and we're sometimes interested in those individual differences and strategies. So they may come out with similar outcomes but do it by different routes, and that should be an interesting end in and of itself because it's part of understanding the sort of diversity of human experience. So we shouldn't be trying to push the correlation to point 9, but we wanted to get off of zero. So I guess that's all I want to say about that.

PK: All right.

PH: I think, yeah, I mean that's the whole belief, by doing the process data analysis that it's going to tell us something about the outcomes. And if we didn't think that then there wouldn't be much point in collecting the process data. Whether or not the performance markers are supposed to validate in the sense of you expect a process that has certain characteristics to correspond to a performance that has other characteristics in a numerological(?) network type sense, like a good collaboration should lead to these kinds of outcomes. I mean, in that sense it's really part of the reliability and validation of the measure. But there's also the concern I think that Yoav already touched on that if we're only getting the information from the process ..... getting

in the performance, then we don't really had any reason to make the task more complex by looking at the performance or the scoring of the task more complex. So, clearly, there's processes there. We acknowledge them. We now have, I mean everything is unfolding in time. Even if you're taking a multiple choice test, that's a complex process. We just don't encode that information when we make a decision about the person's ability. And, if we're not going to get anything from encoding that information, then I mean that's the open question. That's what has to be kind of addressed empirically. It's not something you can design or something you can account on. That's the sciency part of the CPS stuff, so.

PK: Great. Ron?

RS: When we began working with the sub teams five years ago, we were fortunate enough to have novice and expert teams. Expert teams were those who had been at sea as navigation teams and came off of the boat and we got them in the simulator. That was very useful because when we asked the instructors at that time at the end of a simulation, how did the team do and it was always average, above average, okay, below average. So we never really had much from the instructor because they were so focused on their own individual training goals. So we were able to get preliminary novice expert discriminations, both neurodynamically and Jamie Gorman and Terry Dunbar have ..... (?) correlations with the transcriptions as well, the semantic data. In 2012, there were two collisions and after that the submarine force got very serious and did a review, and one of the main problems was the inability of watch teams to work effectively together. And, at that time, Jerry Lamb and his associates at the Naval Submarine Medical Research Center started developing this submarine team

behaviors toolkit where they have X submarine captains, two of them, reviewing each performance and they do ratings on them. And our recent data is now using this outcome as a performance measure, and that is making our life much, much easier. But, as you saw in the correlations, at one part of the segment, it's a positive correlation. At the other part of the segment, it's a negative correlation. When we did the correlations initially there was no correlation, and so you have to drill down and even when you start drilling down to individual parts of the brain, you can see correlations go up and down. But, at the same time, it's telling us an awful lot about what across team brain to brain communication actually is. There's some rambling.

PK: Thank you. Jiangang.

JH: Yeah. I think the process data can be highly correlated with outcome or can be highly negatively correlated with outcome, can be no correlation with outcome at all because it really depends on the specific task. So the CPS is a very complex thing and it also depends on specific task and also the population taking the task. So you cannot have a general finding, general conclusion whether the process data will affect the outcome, oh no. You need to classify that ..... to different specific types of task and a specific population take the task. Then you can make reliable and repeatable claims. Otherwise, we just, in the literature some team will report, okay, there's a correlation. Some team will say there's no correlation. They're all right. They are using different tasks, different populations. So that's my.

PK: Okay, very good. Thank you. We'll now open the floor to any questions. Anyone have a question for the panelists? We have Steve coming up the microphone.

R: First, thanks for a great and fascinating panel. First I had a comment and then I have a suggestion. The comment is never talk about consciousness because we really don't even know what that means, and I think what you really mean are the constructs that have to do with awareness and lack of awareness or implicit or explicit. So there's a lot of data in cognitive psychology and in organizational psychology that's talked about implicit and explicit knowledge with regard to performance. So I really encourage you not to say team consciousness. The second thing though is in looking at this process data methods and thinking about some of the talks yesterday, what I'm wondering is if we could take a lesson from other disciplines like computer science where they have these competitions on a common dataset. So Saad Khan talked about the AVEC dataset and their analysis associated with that where they had a competition to see whose methods could better predict affect using computer vision recognition systems. So yesterday we hear about the ATC 21 and the PISA models of collaborative problem solving. My colleagues and I have the macrocognition and teams model collaborative problem solving. So I'd like to encourage funders here to maybe consider funding kind of a grand challenge or a competition where we take a common dataset and we use these methods to try and see what model is most predictive of collaboration because I think that's really what we're reaching for, even though we're looking at it more piecemeal.

RS: I'll make just one comment. One comment on that. David Pinkus at Chapman University in association with the Nonlinear Dynamical Society is hosting a two day workshop where they're going to have people running all of the different nonlinear dynamical model analyses. And they want people to come and bring their

data so that it can be run through all of the different procedures. So if you feel that your data has very high nonlinear dynamical properties, David Pinkus, Chapman University. I think it's in March or April.

PK: Great. Thank you. Next question, please.

R: Thank you for an excellent session. My question is an extension to Jiangang's presentation, but it's open to everyone on the panel. Now, when we talk about practicality an important consideration, and I think many would agree, is cost-effective. So has the panel given any thought about the scalability of such analysis? So we are not talking about 300 people, we are talking about 300,000 people? Thank you.

JH: I think that's a very good question, and I remember a few months ago I ..... to talk with my colleagues, Bob Mislevy ..... at ETS about the scalability of this simulation-based task. I think our conclusions that when you want to assess some new skills that cannot(?) be measured(?) reliably using traditional items, you should seek for those kind of new assessments. If what you are trying to measured can be well measured by the traditional methodology, what's the point to do this kind of a more expensive thing. Yeah, I think for all these kind of simulations and ....., it really cost a lot of money and take a long time. We know in the traditional item development, you allow people to develop bad(?) items, so that's pretty normal. After you take the test, you say, okay, this item is bad, you take them out. But when you spend $40,000 and develop simulation, ..., you find that no, it's not working, can you say, okay, I throw it away? That's really a very good question. I think it's not only a question for me, it's a question for a lot of the people.

YB: Just a brief thing I'll say. I think that question kind of shoots an arrow right into the heart of the face-to-face versus computer mediated collaboration issue because if you're satisfied that computer mediated collaboration gets at what you're trying to get at, then the scaling issue isn't an issue. But, if it isn't and that's a big open question, then if you need to see someone actually participate in a real face-to-face team in order to evaluate their contribution to that team, then that scaling problem is not going away and it's just a reality.

PK: Thank you. Next question, please.

R: I have a question about a lot of the data that you've discussed has been what you come in with and what your outcome is. And since you have process data, so I work with data that goes at one second intervals and I work on creative outcomes for collaborative team scientists or artisan(?) scientists, and so I'm wondering about these oscillations when they're working together. And instead of saying were they successful at the end, how to look at this data, are there precursors to a creative event, and can you look at it from those types of data points since you do have process data versus yes, they were creative or they weren't creative. And that kind of starts to pull apart what we can look at as far as resilience within the team as well.

PH: I was going to give that to Yoav because clearly that's what his models are doing. But, I mean in that case, so then the analysis that I showed for the chat data only considered the chat data of two people. If you wanted to add into that undifferentiated set of events, an outcome, and look at how the outcome is also part of the process, that's clearly something that can be done. It wasn't really feasible with the Tetralogue example. It has been more feasible with the basketball data I've looked at. But, in

Tetralogue there was only seven questions. It became more difficult to treat that in a sequential time series. But, presumably, if you're collecting the process data, you also have time signatures on the things that you think are outcomes, so task performance related properties of the task. You can definitely do those types of analyses and it just naturally falls out of the methods that we've been talking about, I think. I would encourage you to look into it, definitely.

PK: Yoav, comments?

RS: With Nia Amazeen and Aaron Likens at ASU, we've gone ahead and looked at the used(?) wavelets to look at the fractality of the neurodynamic entropy streams, and they're multi-fractal. And, from the scaling exponents in these fractal patterns, you can start to determine whether the behavior at any point in time is exploratory or corrective. And so just from doing the analysis of the data streams, you can start to see long and short-term patterns in there. This was *Social Neuroscience* last year.

PK: You want to add to that?

YB: I think there is this kind of dynamic chunking problem that you have when things are happening over time and sometimes they'll take longer periods and shorter time periods. And so that's like a really interesting problem in itself which is identifying the chunks. And, in the examples I gave, I took kind of simplistic approaches to identifying chunks between the mistake and the correcting of that mistake. In Ron's example, he had chunks because he knew what was happening with the briefing period and the debriefing period. But if you don't, if you have this kind of organic thing, then

that's one of the problems that you should try to address using process data analysis is how to identify chunks, and there's a lot of different ways to do that but we can talk.

PK: Thank you. Next question, please.

R: I think one of the concepts down the line of after analyzing process data and kind of thinking about a measure for collaboration in these different contexts is coming from the(?) world or the orthodoxy of measurement, one of the central concepts is reliability. And, Jiangang, you mentioned repeatability and generalizability as really important in having a measure of collaboration. Now, reliability requires these independent measures, so how do you compute reliability on these tasks? Is it a useful concept even? Do you need more data or do you feel you have enough with just that one task?

PK: Great and very central question for this whole panel.

JH: Yeah. So I'm trying to answer this question, but I'm not sure I can answer that in a very satisfactory way. Basically, let me say something can be, some measurement of results can be repeated and they're(?) unreliable(?). There are two things. The first, the construct itself is kind of stable for a long period. If the construct itself is changing, changing, changing, there's no way you can get repeatable measurements. The second is the methodologies ..... So, if you have a stable construct but you use a very terribly structured methodology, then you've got fluctuating(?) results. So, in terms of this collaboration, if you want to tell anything about the repeatable reliability, it basically means you need to do this multiple times. If you do that once, you cannot talk about repeatable something. So that's why we are developing a kind of a platform that will allow us to plug in to multiple tasks relatively easily, and also

we can change different partners relatively easily. When you have a number of this kind of response to different tasks with different partners, then you can define the kind of reliability in the sense of how consistent the results are. So, yeah, that is mine.

PK: Anyone else want to take a shot at the reliability question?

RS: Just real quick, again we were fortunate that the navigation task ahs a periodic component built in which is every three minutes they fixed the position of the boat by a rounds process. And so we can go in and aggregate these together over a performance and actually look at the dynamics as the team organizes during the last minute, marks the position of the boat, and then begins to become flexible and loose again as they move out of the actual marking. But it has to be internal. You have to have some type of internal measure, and this is a periodic one in the teams, fortunately.

PK: Okay, thank you. I thought, Liz, we're going for another five minutes until 11:20, right?

Liz: Yes, ...

PK: Okay. so we'll take one more question.

R: I think some of you might be familiar with the definition of productive failure in collaboration. So I wonder, I think this is related to the last question in the panel discussion about the correlation between process and outcome. So it seems like if that happens, meaning there will be negative correlation with outcome, then how would we treat that in scoring of the collaborative process and then the outcome data? Should we combine these two, different scores in some way or using some psychometric models to do that? Any thoughts on that?

PK: The process scorer disagrees with the outcome.

R: Yeah, so meaning they may get the wrong solution but then actually the process data is very good. They have very good collaboration, meaning are they actively(?) ..... prior knowledge and also differentiate the prior knowledge very well but then they still got the wrong answer in the end?

PK: Right, not unlike having a great proof of something that is actually wrong, right?

R: Right.

PK: Okay.

YB: Yeah. So I think that's absolutely really important and so it goes to the whole kind of there are different ways of working, different ways of doing things. And we don't want to design tasks so carefully that we prevent that from happening. So I think it's important in the designing, coming back to the design, that failure is an option and that you should be able to distinguish failure in a productive way from failure in a non-productive way with just kind of conflict and no attempt to resolve conflict and that kind of thing. So I think that's absolutely something we need to be thinking about.

JH: I think when you have something very good and you are sure it's very good and also you have something not very good, also you are very sure it's not good from the same group, that means you are marrying two different dimensions. So you don't have to take away or throw away one of them. You just keep both of them. You say, okay, here's this dimension, this is another dimension. Yeah, that's what I think.

PK: Thank you.

PH: To chime in on that, it's definitely the case that you can have good collaborations that don't lead to great outcomes in a lot of the kind of process oriented

“non-cognitive” skills such as resilience and motivation don’t necessarily entail that you have a positive outcome to the task that you’re engaging in. But, if you don’t anchor the meaning of collaboration or successful collaboration to the task outcome, then you need another set of constructs to kind of build up that meaning. So you can look at various task components, but how do you know that that’s telling you about positive collaboration rather than some other set of skills or things that could be demonstrated in the process data. So there needs to be some criterion against what(?) you validate the interpretation when it is indeed a good collaboration, and if it’s not the outcome then it has to be something else.

PK: Great. Ron? Well, that’s it. That’s a great question to end on because we’re having a great collaborative experience here together, but we don’t yet know the outcomes and so we can’t tell whether it’s successful. With that, we’ll break for another 20 minutes and then reconvene for our last panel that Alina will lead. Thank you.

END OF PANEL