# iPAL Meeting Notes

December 9, 2020

## General Updates

- We have had some delays with getting the iPAL website transferred to BC. However, progress now seems to be moving in the right direction and we hope to have the website up soon.
- Jim, Henry and Rich are pursuing an opportunity at U. Illinois Chicago.
- Kai noted that he has been modifying his PA (water case).
- Prior to the meeting, Rich circulated the scoring framework from the Frontiers paper (Braun et al., 2020), as well as scoring rubrics from Henry, Julian, Olga and Huanhuan. It will be important to develop some guidelines for the development/refinement of iPAL scoring rubrics in relation to the framework. Of course, the framework itself is subject to modification.
- Colleagues have found that students' CT skills are at a fairly low level when using comprehensive scoring rubrics. We may have to modify the PAs to make them more manageable in the time usually allotted to the exercise. The scoring rubrics will have to be modified accordingly. (See further discussion below.)
- In an instructional context, PAs can be modified to provide scaffolding for learners. We need to be more explicit about what is expected from the student. Model responses would be helpful.

## Main Topic 1: Inter-American Development Bank (IADB) Opportunity

- The meeting began with a presentation from Julian and Natalia about a potential opportunity with IADB (see document attached). The bank is similar to the World Bank but with a regional focus on Latin America and the Caribbean.
- Julian's team contacted IABD following our April 2020 iPAL meeting.
- IADB has been working on a 21st century skills project. Their goal is to provide free, high-quality assessments.
- They are interested in assessing critical thinking with the goal of issuing certificates of accomplishment that holders can present at job interviews. They are particularly interested in helping individuals who have not had access to higher education or advanced training opportunities.

- This effort from the IADB is focused on the public good.
- IADB has already signed agreements with some agencies, including KERIS, a Korean agency. They have asked us to consider joining their consortium.
- IADB is willing to finance automated scoring (and perhaps some aspects of task development). Automated scoring is essential to the agreement.
- IADB would share all data resulting from this overall project.
- What is IADB's understanding of performance assessments?
- They have seen examples of performance assessments from us. Therefore, their understanding of PAs is, by that nature, aligned with ours.
- What could we offer to IADB? The "Migrants" PA that was originally developed by Heidi and Auli, has been adapted by Julian's team to Colombia (language and context).
- Julian's group is collaborating with additional universities to pilot the adapted scenario. They have already collected substantial data.
- There were some concerns among the group about using the PTs for certification purposes, given that our task was designed for undergraduate students and they have not been performing particularly well.
- Would the broader population of interest to the IADB have the necessary basic literacy skills to complete the PTs? Could print-to-speech reduce the reading burden?
- We might need to give students a preliminary, basic literacy test. Julian said this type of assessment is available in Spanish. Julián also mentioned that the Colombian adapted "Migrants" PA includes some multiple-choice questions that assess basic reading skills.
- We have to recognize that there are substantial variations in the Spanish of different countries. Adaptation will likely be needed.
- Should the Migrants PTs be disaggregated into multiple mini-PTs focused on specific facets of CT skills?
- Benefits of this project/collaboration with IADB: This project could push us into (1) developing the automation needed to scale up iPAL assessments and (2) disaggregating the larger tasks into smaller tasks.
- We would like to continue with a feasibility study, including cognitive labs and exploration of the possibility for adaptation.

- We need to partner with a company with expertise in automated scoring to address automated scoring. Julian has been in contact with someone (Quantil- Diego Jara).
- The automated scoring development and calibration requires a full range of responses at various ability levels. This can present practical difficulties given the current experience with responses received.
- IADB wants to provide a certificate of CT skills for students "successfully" completing the PT. (It is currently unclear how they are defining "success.") They also want this certificate to represent a relatively high skill level. They do not want a watered-down credential.
- One member was concerned about what receiving a certificate means in terms of what that credential says about a person's skills. What take-aways will people have based on the certificate? What is the interpretive inference that can be supported?
- One member suggested that the true value of the certificate will probably be determined by the market. If recruiters start to use it, then it will have more value.
- Could we make an argument to the IADB that there are other criteria that should be considered? For instance, there could be important, verifiable biographical indicators (e.g., volunteering) of students' skills and job-readiness.
- Is it possible to see how the automation of the scoring would be impacted by different approaches (e.g., Google translation versus human translation) to the translation/adaptation to other cultural contexts and languages?
    - Perhaps down the road we could explore this as one of many research projects.
- We need to be careful about distinguishing iPAL's responsibility from IADB's responsibility. iPAL's task is to create a construct-valid, reliable assessment. IADB's task is the certification process and validation of the task for certification.
- We need to clarify to IADB who is responsible for certification and validation.
- IADB would like to get this assessment out ASAP. However, we should tell them that we need to do a pilot. Then based on the pilot we "might" need to make revisions.

- When building the assessment system, it might be helpful if the IADB ensures the system has the capability to capture auxiliary information (e.g. overall response time to complete the task, how much time students spend on a page, etc.).
- At first, IADB wanted to have the assessment in their system by April, but now they are saying by June.
- We should share the results of the AHELO study with them. IADB should collect data for about a year and make refinements. They need to think carefully about certification and whether they could do more harm than good.

**Main Topic 2: Mini Assessments**

- Could NGSS-related assessments be an example of how to break down a larger assessment into smaller mini-assessments?
- A somewhat different situation, but some insights possible. Note that a team at MSU has been working on automated scoring for NGSS responses.
- In her Bayesian analysis work, Pat has identified the following steps toward argumentation: (1) Summarizing article, (2) Compare/Contrast, and (3) Argumentation - Synthesis.
- Rather than breaking the assessment into summarizing, compare/contrast, and synthesis, I was thinking of mini assessments in moral reasoning, etc. (i.e., components of CT).
- Perhaps there is a three-stage process we should keep in mind. Stage 1: Understand key points in an article. Stage 2. Mini assessments. Stage 3: Tackle larger tasks.
- We could break down tasks in one of two ways: (1) Teach students CT using an incremental process (i.e., similar to teaching parts of a musical piece before performing the full score). (2) Decomposing a task that you already have.
- Olga shared some prior research in which the construct was conceptualized as a multifaceted process broken into smaller tasks (Task 1: Searching facet. Task 2: Evaluating information facet). The easier tasks worked well. However, with more complex facets students had difficulty in demonstrating proficiency. Cognitive labs revealed students did not have enough time to demonstrate their ability.

- We want to have example practice tests available to students taking the IADB performance assessment. We could test new tasks here, and randomly assign students to experimental conditions for research purposes.
- Students' writing skills are also an issue. Also, controversial issues have multiple correct answers. How do we score this?
- Scoring could be problematic. It is possible to score these, but it is challenging. Perhaps hold off on moral/ethical issues. Suggested that Julian decompose BC's legacy admission scenarios into multiple smaller skills, in order to preserve Migrants for the main assessment.
- Heidi has developed a scoring rubric for moral reasoning. However, she does not think it would be possible to automatically score this construct. Also, she analyzed data from the CLA and found low correlations between constructed response questions and multiple-choice items. Heidi and her colleagues think the low correlations are due to the lack of time available to complete the CLA+ items. This time pressure leads to a high amount of guessing and/or skipping questions. Additionally, Heidi believes the low correlation is due to the different skills measured by PT tasks and multiple-choice questions. To reduce the time needed to complete the overall assessment, Heidi suggested using the same documents for the PT and multiple-choice questions. We cannot ask students to read too many documents. (Another member agrees that the same documents should be used for PT tasks and multiple-choice questions.)
- Multiple-choice questions might be necessary but not sufficient for doing well on a PT.
- Instead of multiple-choice, perhaps ask students to jot down notes about specific prerequisites needed to answer the full PT (e.g., can you identify the weak sources?).

**Main Topic 3: Discussion of Scoring**

- It is easier to train people to score more objective aspects of students' responses. In some cases, holistic rubrics can lead to higher agreement. It might be easier to do automatic scoring with a holistic scoring procedure. However, it is still somewhat unclear.
- We should think about automated scoring when designing the PA and the scoring rubric. The number of responses needed to calibrate automated

scoring might be higher depending on a holistic versus analytic scoring approach.

- It is important for the automated scoring team and the test development team to work together to get a good degree of construct validity while assuring feasibility of the scoring algorithm.
- How many documents are too many for students to really understand the documents? Be careful about picking a reasonable number. One member suggested perhaps 3 documents (might depend on length of the docs).
- One member reminded us of the DBQs in the AP History exam. The College Board had to collapse some of the scoring categories because students could not get to the most advanced level.
- One member reminded us of Sam Wineburg's criticism of the DBQ items. There are some problems with DBQs.
- Computer-based automatic scoring and human scoring require different approaches. It is important to keep in mind the Student Model (ECD). For automated scoring, you need very detailed descriptions of every process.
- We cannot necessarily know every possible route a student might take. Worried about modifying the task to align with the scoring rather than the construct.
- The deep-learning approach, where the algorithm essentially decides what to do, is the opposite of what Olga described.
- Someone outside iPAL team would be helping with the IADB automated scoring, uses neural networks.
- Deep learning requires more data. Deep learning is more about developing prediction algorithms, rather than providing interpretable scores.
- We want our scoring system to predict the scores that humans have provided. It is a different process than what linguistics might use.
- Could we use a Bayesian approach to scoring? If a student cannot do the most basic task, then they probably would not need to be scored on more advanced tasks.
- Should we even have a generic rubric?
- What level of family resemblance do we have among iPAL tasks? If they are not closely enough related, then perhaps it does not make sense to have a generic rubric.

- Our tasks are probably not close enough yet for a generic rubric. Think about how one generic rubric might be easily specified for a family of sibling tasks (intentionally designed).
- ETS's GISA reading comprehension task had simpler tasks along the way connected to a larger task.
- We need to be very careful about what skills we are focusing on when scoring. Students' skills vary across tasks.
- What is our model of being able to do the more complex tasks? Is it compensatory or conjunctive?
- If we use a Bayesian decisions-based approach, we need to think about which skills are absolutely necessary and which can be compensated for.
- Students surprise us all the time. Their skills might not follow along the hierarchy we, as assessment developments, have in mind. Very worried about establishing any sort of cutoff point.

**Main Topic 4: Funding**

- It is challenging to get funding for higher education research. Could we circumvent this by focusing on the high school to college transition?
- These tasks might be good for general instructional purposes. Under certain instructional contexts, the time limit would not be as much of an issue. Consider using these tasks for entering first-year college students. UIC has "Freshman Seminars" that could serve as an appropriate setting for CT assessment.
- This is sometimes referred to as "assessment as learning." BC's Lynch School of Education and Human Development has a freshman course that is specifically designed to develop students' CT skills while helping them acculturate to the demands of a liberal arts college. Using another measure of CT, he and colleagues have shown that students make substantial gains over the course of the year (no control group yet).

**Next Steps**

- Julian (with support from iPAL) will continue discussions with IADB.
- Julian will continue conversation with Heidi/Auli re: Migrants.
- Another iPAL webinar will be scheduled for January or February 2021.

## Happy holidays and a good new year to all!