WIDA November 2016 TAC Minutes

Contents

1. Review of ACCESS 2.0 400 administration

Presentation of powerpoint (Shu Jing)

Interruption issues

Lyle – any point in re-estimating scores from S400 verification study & recalculating student scores?

Gary – WIDA has done this; this was the data used for standard setting

Differences between 400 transition year and going forward: yes, CAL/WIDA will receive early spring data extractions. Going forward, embedded L/R FTs will mean that sample size for L/R will be available earlier (not in early spring).

July data – what percentage of students filtered out due to interruptions? Gary: 10-15% in earlier states; proportion is smaller with later data draws as the interruption issue is mitigated

Randomness of interruptions – Kei; that the interruption is treated as random, in terms of her statistical model, although there are systematic trends across locations

Pre-equating sample

Lyle – how were the fourteen states for the pre-equating study selected? Gary – these were the earliest states to test. Lyle; concern about sample bias? Gary -- sample size is large and this provides confidence; Kei -- the sample is similar to historical distribution; Dorry – CAL has always sampled from early returns. Carol – that the early return states likely do not represent the diversity of home languages in WIDA states.

Scaling

Lyle – does putting S on the L scale and W on the R scale presuppose some relationship in the construct across the domains? Dorry – that it doesn't, but allows Rasch model to create vertical scale; Shu Jing – that what is key is the rank ordering of students. Lyle – you may be building in dependencies that affect the interpretation of the composite. Important to frame the information and presentation in terms of the assessment use argument and the claims & evidence.

401 Implementation--Speaking FT

Tim – do Ts/Ss know which speaking task is the FT task?? In his experience there were tasks that Ts had identified as FT tasks which were not always completed. Aki—any location effect of FT folder? Dorry – JIC there is a fatigue effect it's actually better for the FT folder to come at the end.

Suggestion – look to see if there are significant numbers of students who are not completing the FT tasks.

In general, WIDA does not communicate to students/administrators that there are FT items on the test; Lyle – that it is not ethical to have students complete items that are not scored without telling them. Vince – what do the professional standards say on this point?

Carol – do S FT items have to be appended as a folder or could they be placed into existing folders?

Carol – how does the scoring of FT performance items differ from OP items? Dorry – DRC raters are trained on task-specific scoring, this is not possible for the new FT tasks; DRC raters didn't begin scoring S FT until a large quota of OP items were scored. Implications – the FT scores on perf. items are used for item selection but not for scoring OP items, as there is early return data from spring that is collected to produce OP scoring tables.

Writing FT; are there consent/disclosure issues with W stand-alone? Are there incentives (school may receive WIDA PD)? Gary – hope is to move toward embedded FT for W.

Tim – could you set up direct feedback on student performance or instructional purposes as a FT incentive in cases of standalone FTing?

Discussion of time needed for performance tasks & WIDA/CAL's future plans to examine this.

Carol – Rating efficiency study in which raters rate the first 100 words, and then rate full sample, and look to see if there is any difference.

Follow up items:

a) Look at FT item performances for location effects (kids not taking the folder)
b) Look at other ways of spiraling items so that they aren't all at the end
c) Look at the "ethics" of where the FT folder is placed and whether students should be informed of this - report back at the next TAC

2. Standard Setting Report

Presentation of powerpoint (Gary/Dorry)

Phase I: Panelists' choice for Phase I cuts L/R was much higher than a 6.0; this was reduced to 5.5. Tim—how did you know how much to reduce? Gary – this was a preliminary decision that fed into phase II. Gary – also because we knew that CCR standards would require an increase in proficiency. Tim – continue to be concerned that the policy based decision here is not well motivated.

Discussion of the "stretching the rubber band" metaphor.

Tim – if the PL are now at a higher bar, should the descriptions of the performances not also be adjusted? There is a concern that while PL cuts have changed, the description of what kids have to do have not changed – why not? We are now expecting more of students at the different categories.

Concern over using means rather than medians in case of outliers (in phase II)

Re: Composite scores – has there been empirical checks on whether the a priori weights are actual weights and that intercorrelations between domains don't affect the a priori weights?

Tim: That although the proportion of kids being reclassified next year will decrease, once kids have had time to work through the extra couple of years in program that will be required by the higher cuts, the proportion of kids being reclassified should even back out to baseline levels.

Discussion of box & whisker – discussion around the question of the difference between ELP and ELA.

Lyle: How does the standard setting relate to the assessment use argument – and specifically, what are the consequences of the standard setting decisions for assessment use. Will provide notes on some ways that the info already presented could be packaged/framed in AUA terms.

Follow up items:

a) In writing the standard setting report for peer review, ensure that there is an explanation for why bookmarking went high to low; it may be interpreted or assumed that this decision contributed to why the PL is very high.
b) In writing the standard setting report, be very clear about why you chose the PL values you did for Phase 1 cuts (i.e., 5.5 for reading and listening, and 4.5 for speaking).
c) In writing the standard setting report, it might be helpful to look at inter-rater variance at Phase I, consider looking at differences between teachers and policy makers.
d) Consider how you will position the standard setting activity in the AUA

3. Writing Score Issues

Presentation of powerpoint (Shu Jing)

Rubric Refinement Study

Lyle – AUA framing; study will collect evidence for claims of meaningfulness for claim 3 and consistency for claim 4

Description of current work at CAL on "plus" rating description.

Carol – are the rater-analysts comparing to the 3 or the 4?

Carol – expressed concern that there were only two raters who looked at the plus tasks to detail their characteristics. Dorry noted however that when the plus level descriptors are produced, they will be produced across clusters so therefore there would be five pairs.

Lyle – has there been any examination of whether the general rubric can be used by raters in the same way as the task-specific information. Gary -- How should the distinction between the general rubric and the grade-level expectations be communicated to the field?

Lyle – need evidence that the grade-specific and task-specific additional notes don't change the construct

Considering phase 2 (of the Rubric Refinement Study) with additional raters:

Carol – would be helpful to do this with a separate set of papers so that any bias in the sample isn't reproduced across phase 2. Would additionally like more samples. Who will the raters be? CAL or WIDA raters? DRC raters? Gary – CAL could work with grad students and train them on same materials as DRC raters see. Carol – would prefer to see DRC raters.

Follow up items:

a) None. The TAC did express some skepticism about the utility of the study.

4a. Online and Paper Comparability Issues

DAY 1

That the score on test A means the same as test B; that the proficiency category from test A is the same as test B; and that the constructs are the same.

The interpretation the ACCESS 2.0 (Online ACCESS) assessment is comparable to the Paper form of the ACCESS 2.0 (Paper ACCESS) if

- *The Online and Paper ACCESS are appropriately equated*
- *The Online and Paper ACCESS are administered consistent with the test administration guidelines*
- *Each student is administered the appropriate tier for Paper ACCESS*

The Online form of the ACCESS 2.0 (Online ACCESS) assessment is comparable to the Paper form of the ACCESS 2.0 (Paper ACCESS) if

For the Paper ACCESS speaking test

- *Each rater is properly trained using materials provided by WIDA for local scoring of the speaking test,*
- *Each school/district provides sufficient time and training to prepare raters for rating the speaking test,*
- *Each school/district provides the appropriate resources raters need to rate the speaking test, and*
- *Each school/district routinely monitors the rating of speaking tests and evaluates inter-rater reliability indices*

Lyle – is their backing for all of these claims (paper speaking)? You can't make the argument if you don't have the evidence. But procedural evidence is okay.

Vince – no information is provided to teach schools how to maintain IRR with local scorers. Dorry – check with Margo on this. Could WIDA produce standards/ procedures for schools to monitor their IRR? Carol – could routine monitoring include observation of test administration? E.g. video monitoring that is then externally reviewed.

Lyle – could the framing be that local educators must do x,y,z and if not, WIDA cannot vouch for the consistency? WIDA is ultimately responsible for the assessment.

Seon Hwa – is there any way the paper speaking could be scored centrally?

Lyle/Carol – would like to see an agent made responsible for monitoring/auditing for the comparability evidence for paper speaking.

Tim – take care with the implementation communication. In 1.0 Speaking test, there was no evidence provided on local IRR etc. Don't want a conversation about who is "cheating" on the test and negative job consequences for educators. Frame conversation around enhancements in industry best practice.

Carol – what kind of evidence can states provide about appropriately assigning students to tiers? (Gary – will return to this question) Dorry – one way is to check frequencies of tiers to check for capping out on A or B.

Conversation around what is the responsibility of the state and what is the responsibility of the consortium; how to communicate this to (with) states.

Tim – is ACCESS 2.0 comparable to 1.0 by the line of argument you are using? Gary – No, further, WIDA is not providing growth info for test for this year or next

Review of slide on Score Equivalence Evidence (Cohen's D): Dorry – Listening early grades could be a tier capping effect – kids are assigned lower tiers because of literacy abilities but could perform at a higher level on Listening

Kei – looked at historical gains across years; found a greater than expected jump in year to year gain for speaking.

Vince – is there a concern in a paper based state that all of the stats for scoring my students are based on online student data? We would potentially like to compare our scores to other states. Carol – could a state argue that local scorers are more valid because they have the experience of being teachers? Lyle – is it possible the kids actually perform better with an interview? Dorry – not F2F interview; tape and a test booklet.

DAY 2

Current state of affairs: specs for online and paper are the same, however refreshment plan is different.

Is the same set of specifications sufficient for comparability? Lyle – yes. Carol – could you look to see if items shared across the assessment have similar difficulties (or at least similar location in the rank order of items).

Lyle – how can states continue to use a non-refreshed test? Gary – intent is to sunset paper, but no planned date as yet. Lyle – recommend that as a matter of policy, WIDA plans for a timeline for retiring paper (except for an accommodation), or at least for maintaining comparability.

Tim – could WIDA create a pricing structure for states which would like to maintain comparability between two forms? i.e. would continue to refresh paper. Any research on use/re-use of testing over time and the time frame that would be inappropriate?

Carol – how important is comparability? Can the argument be that the tests measure the same construct and standards can be interpreted in the same way? But would need different cut scores? There will in any event need to be support and evidence for the validity of the paper test, if it is to be used as an accommodation.

Discussion of capping issues on paper vis a vis online

Shu Jing – what about an online routing test that could predict students' tier placement for paper more accurately? Then a more efficient ordering process for paper? Gary – what about screener? Lyle – if they can't do a computer-based test, how could they do a computer based routing test?

Seon – what does WIDA already provide in terms of resources to predict tiers? Gary – there are materials from WIDA to assist states/local educators, look at scores on past ACCESS, screener, model.

Carol – what kind of research can be done to assist in prediction? Gary – planned research includes looking at relationship between screener and future ACCESS. Tim – ensure that when this relationship is considered, separate paper and online.

Concern expressed over "gaming the system" so that kids can't exit (overallocating to Tier A) for funding reasons.

Carol – could overage rate of paper ordering be increased?

Tim – recommendation that there is communication directly to states in cases where there appears to be excessive mis-assignment to tiers.

Follow up items:

a) The claim for routine monitoring for IRR for local speaking scoring (above), needs to be stronger so that schools/districts take action if monitoring shows problems.
b) The claim for paper speaking comparability should add the type of evidence that needs to be collected to undergird the claims
c) WIDA should provide IRR training to support LEAs administering the local speaking test

d) WIDA stated that they will do a follow up comparability analysis looking at students with matched current year domain scores - report back at the next TAC
e) The TAC recommends that WIDA raise the issue of when to sunset the paper form of ACCESS 2.0 with the Executive Committee ASAP

4b. Interruptions

Presentation on the final status of interruption issues (Kei)

The session was primarily an update.

Follow up items:

a) None. This was an informational presentation.

4c. Speaking Summative Scoring Comparability Study

Lyle -- That the research plan, as part of the AUA, is a rebuttal argument. Where does it fit into AUA? Suggest claim 3 and that the warrant lists facets of measurement design as sources of potential inconsistency, and that research weakens evidence of rebuttals. Ditto claim 2, a rebuttal on potential evidence against impartiality of raters. IOW, comparability itself is not the end; the end is to provide AUA rebuttal arguments against threats to the AUA.

Slide 4; Carol – what are "scoring conditions"? Shu Jing – all of the elements not captured by other items on list. E.g. local rater cannot re-listen. Aki – what about an educator effect? Teachers may be rating their own students. Aki is working on research on difference between f2f and recorded oral proficiency (however raters do not know the test-takers in his research). Noted that f2f educators appeared to count mispronunciations more against students.

Gary – central raters cueing on rubric and on student response; local raters cueing on rubric and on everything that they know and believe about student. Which is "correct"? Seon – the factor is that the educator knows the student (not that the raters are educators vs non-educators)

Shu Jing – Aki, are you proposing that we conduct an educator effect study? Discussion of the practicality of whether it's possible for schools/districts to require that the students not be assessed by educators who know them?

Lyle – possible to conduct a study in which local responses are recorded and scored, and then centrally rated and scored? Carol -- Educator effect study to parse out difference between highly trained raters who are or who are not educators? Seon – is there a difference between scores assigned by teachers who do or who do not provide instruction directly to the kids? Lyle – there is research in the language

testing literature which suggests that training is more important than background (i.e. a well trained rater likely to be more consistent than untrained but ESL professional).

Lyle – recommendation that WIDA recommends that teachers don't test their own classes.

Vince – FL has a law that says only certified teachers may administer. This is so that there can be consequences for malfeasance; teacher's certification can be withdrawn.

Tim – there is both unintended and intended educator biases; we have to be careful with communications so that those who are acting with unintended biases may face the consequences of those acting with intended biases. We should rather make recommendations (don't test your own kid) in the format that this recommendation is intended to protect teachers from being in an unfair and awkward position. That the state should be accountable, rather than the teachers. FL once had a law that stated that only certified teachers can administer. State law provides for possible criminal charges and fines for violations of test security.

Recommendation from TAC – that having teacher score their own students puts teacher in a position of conflict of interest and potentially compromises test scores. States/districts/schools should ensure that students are rated by a qualified rater who is not the teacher of the student.

☺☺☺☺☺Spontaneous celebration and congratulations occurred!!! Congratulations New Grandpa Dorry!!

Carol – what does the local educator scoring quiz look like? Dorry – quiz is a set of scoring exercises. Seon Hwa – what are the checks to ensure local raters complete all of the training thoroughly and also the quiz? Vince – FL reports that some schools did not provide adequate resources for teachers to have time to complete training.

Carol – calling the local rater cert requirement a "quiz" seems inaccurate; implies multiple choice. Dorry – this choice was made to reduce affective filter for teachers.

Lyle – Warrant for claim 1; that this research study will look for evidence that the scores between local and central are not significantly or meaningfully different?

Shu Jing – explained detail on data collection on educator qual quiz from DRC?

Carol – what will you know about DRC raters? Educators vs non-educators? ESL vs non-ESL experience? Dorry – not sure about what exactly could be collected by DRC. Lyle – could do analyses on the facets that you have access to, thinking directly about the kinds of changes that might be made based on the results.

Aki – how exactly will this research answer the questions/reservations about the central/local scorers? Specifically, where do the observable differences between central and local scorers come from? What explains them? Lyle – the dependent variable is score on the speaking test, not the score on the cert

quiz. Better study would be recordings of local educators (with attention to who is or is not scoring their own student)

Vince – do quiz participants get feedback on performance on quiz? Dorry – no.

Tim – running the study will allow one source of variation potentially to be ruled out. Seon Hwa – training likely more important than simply passing the quiz.

Dorry – claim is that central raters and local scorers who take the quiz are both sufficiently trained to potentially provide the same results on student assessments. Carol – including rater characteristics might be very revealing. Lyle – could you collect any qual info? cog labs or questionnaires regarding how educators experienced the training.

Follow up items:

a) TAC somewhat aghast that some teachers might be scoring their own students' speaking tests. Consider (strongly) communicating to SEAs and LEAs that having teachers score their own students' speaking tests is placing them in a conflict of interest and is a potential compromise to test scores

4d. Proposed Keyboarding and Handwriting Responses Study

Lyle –that this research falls under claim 3; warrant that the scale scores are not different across different modes of response

Lyle – that this is could be potentiality threat to the claim of impartiality

Dorry – that WIDA should be able to provide supports so that they can tell which students should not keyboard.

Slide 5 – not just younger students potentially disadvantaged. It's students who can't keyboard.

Gary -- Do scores on HW writing task and scores on KB writing task have the same meaning? Lyle – okay, and if you find they don't, need to find out why not. Gary – need to figure out if the construct is different. Lyle – any evidence on whether raters rate hw vs kb differently? Carol/Gary – yes, evidence in literature. Gary – if we can demonstrate that kids who keyboard poorly perform significantly less well, we would have evidence to be able to say that kids must have kb skills in order to participate in kb test.

Lyle – if you find mode effects and then enhance procedures to counter, these are evidence to back warrants.

Gary – suggestion to include school effects in covariate model.

On rater bias HW vs KB: Dorry – look at samples at same score point which are KB and HW to investigate potential rater difference between HW and KB? Gary – create typed versions of HW assessments and examine them? Lyle – what about thinkalouds with raters?

Carol – consider looking at differences in fit stats between HW and KB group? Rating scale category statistics. Are the scale points being used in a similar manner across modes?

Shu Jing – is WIDA considering an assessment of keyboarding skills?

Follow up items:

a) Several recommendations offered to enchance the proposed study: 1) frame the study under AUA claim 3, 2) include school effects in statistical model (to mitigate potential location effects), 3) in analyses look at item fit statistics between HW and KB groups to address whether scores are similarly situatuated across the scoring scale

5. Composite Conditional Standard Error of Measurement (CSEM) Study

Lyle – that the very fact that we combine scores across domains seems to speak to an assumption that relatioship between domains are oblique not orthongal. Gary – if we assume we are measuring oblique dimensions, what does that mean for an additive composite and its standard error? And further, what is the effect of the weights?

Aki – recommend MIRT approach assuming orthogonal relationship.

Aki -- Bootstrapping model.

Lyle – run MIRT models fo r each composite Gary – don't need to, just need to assume they are orthogonal. Lyle – running compensatory/oblique model reflects the claim that the domains are aspects of a single construct, whereas running orthogonal model does not back that claim.

Dorry – how are people using standard error estimates? Gary – in value added models and in student growth models.

Lyle – more important to worry about substantive issues of score interpretation and the AUA. Need to do the "right thing" – either approach could be the "right thing." Most convincing will be to model composites that match up with score based interpretations and the meaning of the scores. Gary – how do we understand and impute meaning to the error of the composites?

Shu Jing – look at distribution of composite scores, and sample distribution of the scores that are found across domains within a composite score point. Benefit – not model dependent, and easy to do sampling. Lyle – this is appealing.

Kei – how do you compute test reliability for the composite if you don't use an (IRT) model?

Follow up items:

a) The TAC questioned whether a compensatory (oblique) MIRT model was appropriate. They suggest using an non-compensatory (orthogonal) MIRT model instead. The same composite combination approach could then be applied.

b) The TAC also suggested looking at creating pseudo-CSEM values by looking at the variance of each composite score with census data. Using a pre-defined (and theoretically sensible) central tendency approach, calculate the standard deviation at each composite score point. This becomes the pseudo-CSEM. This will be done at the end of each testing cycle and applied forward to the next year's scores and score reports. The pseudo-CSEM could take the form…

$$pseudo - CSEM = \sqrt{\frac{\sum(Score_i - CentTend)^2}{N}}$$

where, $Score_i$ is the observed composite score for student *i* and *CentTend* is the selected central tendency statistic. The goal would then be to examine the differences between the Rasch CSEM, the MIRT CSEM, and pseudo-CSEM. With a priori assumptions about the purpose of CSEM, select the most appropriate and report back to the TAC on the decision.

6. New DIF Study

Lyle – what are other potential sources of DIF? Length of time in program? Gary – potentially interrupted formal education? Discussion of difficulty in looking at non-Spanish home languages.

Carol – is the Bayesian approach more sensitive? Shu Jing – yes. However may over-identify some B level items.

Aki – how do you impose the prior? Shu Jing – papers assumes a normal distribution. Impose normal MH.

Aki – sees DIF as data description, not searching for hidden patterns in the data. Bayesian approach might be "too much" … ?

Carol – are there any other methods for detecting DIF in small samples in CATs? Shu Jing – not aware. Dorry – likely a one-time thing, sampling issues to be fixed.

Lyle – what about non-uniform DIF? What is more critical – over or under identifying DIF?

Follow up items:

a) None.

7. Validity Study Plans

Lyle – suggest framing studies in terms of the claims they are intended to support.

Study of how ACCESS 2.0 materials are used for classification and placement purposes (consequential)

Discussion of what it means to have "sufficient" information when screener is one part of multi-part decision making

Lyle – important to brainstorm unintended consequences. Possibly, misclassification. What mechanisms might be in place to look again at cases where students are misclassified? Gary – just published tool for reviewing misclassified students.

Carol – who are the stakeholder groups that may face unintended consequences? Lyle – intended consequences for stakeholder groups should be stated in claim 1.

Study of the the use of accommodations on ACCESS 2.0 online (construct)

Seon Hwa – can you track which accommodations are matched to which students disability categories? Gary – No. Also, we can tell if an accommodation was allowed, but not whether or not it was used.

Study of the underlying factor structure of ACCESS 2.0 online and paper (construct)

Hypothesis that there are four first order language domains and two second order factors.

Lyle—that there should be a third order factor; language proficiency. Dorry concurs.

Receptive vs productive – confounded by multiple choice vs non multiple choice tests.

Shu Jing/Dave – if you flip the model order, you get a bifactor model.

Carol – will these factors be looked at by grade? Gary – yes, by cluster.

Study of the relationship between the online screener and annual summative ACCESS 2.0 assessment (criterion related)

That WIDA is not responsible for paper screener. This is WiCEPS.

Carol – how can you look at whether screener identifies accurately? Gary – we have an instrument that looks at academic language in the classroom which will be helpful. Dorry – if a state has a process for re-deciding after the screener placement, could that be considered as part of this research?

Any others?
Vince – what about an independent alignment study? Gary -- In a couple of years, a study looking at alignment between standards and test specs?

Follow up items:

a) Frame these studies in like of the AUA.
b) In the construct validity study, look at a third order factor (potentially identified as academic language)
c) When doing the screener validity study, make sure you look at classification accuracy. It may also be useful to look at classification accuracy by state, i.e., by different ACCESS scores.

8. ACCESS 2.0 401 projection and contingency plan

Carol – are there good assurances that test interruptions will not be seen? Seon Hwa – I'm surprised that we continue to have contingency plans or issues that we have experienced and that expect to have been fixed. Would like to see very clear signs of improvement.

Follow up items:

a) None.

9. Yearly Review and TAC agenda

Lyle – operational to include progressive development of AUA and fitting ongoing studies into that development.

Carol – that TAC responsibilities might include creating more long term research agenda, and identifying researchers outside of current circle.

Lyle – TOEFL COE developed a matrix from validation framework to identify gaps.

Discussion of when TAC would review ATR. Carol – perhaps committee could split up the report to review?

May 1 & 2, 2017 – spring meeting in San Antonio, TX

November – to co-incide with veteran's day (observed)

Vince – are state members rotating members of the TAC?

Follow up items:

a) Scheduled next TAC meeting to be after the NCME conference in San Antonio, TX.
b) As part of the "formal" TAC agenda, introduce an agenda item on long term psychometric, research studies, potentially identifying researchers outside of current TAC, WIDA, CAL circle.