Indiana Validity Study Report Outline

V. 1.0

Validity Study Number: 6 **Short Title: Comparability of Paper-Based and Online Assessment**

Lead Author: Derek Briggs

Study Overview: A key issue for states that use online assessments for most but not all students is how comparable are the results of the assessments given on paper to those administered online? This is important to study both for considering the policy issue of whether universal online assessment should be used, as well as whether any adjustments to students' scores should be made since the ISTEP+ test results are used in school and in educator accountability.

Methodology—The IDOE plans presented to the ISBE in May 2015 included "Paper/pencil and online comparability studies" for completion in August 2015. We suggest that DB review the plans for the study or studies, and provide his reactions to them this summer, and then monitor the conduct of the study or studies over the summer, and finish this by reviewing the results of the study or studies.

Study Data Needs and Information Supplied

Documentation Sought	Documentation Provided
A. Information on the design of the comparability studies planned or conducted.	CTB Response for IDOE 10.20.15_FINAL.pdf 2015 ISTEP+ vertical scaling Memo Sep 11.pdf
B. Documentation of results from comparability studies conducted.	Mode_Study_Draft_10 02 2015v2.pdf CTB Response for IDOE 10.20.15_FINAL.pdf Mode_Study_2015_ISTEP_Oct_23.pdf

Summary of Documentation

My initial review began with the document "Mode_Study_Draft_10 02 2015v2.pdf" that was sent by Cynthia Roach on 10/13/15. This draft document was missing a considerable amount of important information about the design that supported CTB's evaluation of mode effects. It also contained some information that raised some flags about the process that CTB used to estimate the magnitude of mode effects. I provided feedback about this over email on the evening of 10/13/15. This led to a conference call with SBOE staff along with Ed Roeber and Wes Bruce on 10/15/15. Concerns were relayed to CTB and IDOE that same day (see below), and we received the document "CTB Response for IDOE 10.20.15_FINAL.pdf" on Tuesday, 10/20/15. Lastly, we received the document Mode_Study_2015_ISTEP_Oct_23.pdf on Friday, October 23rd.

Discussion

I raised the following concerns in an email after reading the initial mode study draft "Mode_Study_Draft_10 02 2015v2.pdf". The crux of my concerns were about (1) the validity of the approach that was used to place paper and pencil (PP) and online (OL) items onto a common scale, and (2) the validity of the approach (propensity score matching) that was used to create equivalent groups of students before estimating the effect of mode of testing on student performance.

> "1) It comes as news to me that the PP and OL items were scaled using concurrent calibration. I'm rather nervous about this approach because there is probably good reason to believe that it would introduce an additional source of dependence between items over and above that which is caused by the latent construct that is the target of measurement. So I would expect to see that, at a minimum, some exploratory factor analyses were conducted prior to conducting the concurrent

calibration.

2) Almost everything about this investigation hinges upon the ability to create equivalent groups of students using PSM. Unfortunately there are a lot of important details missing about how this matching was conducted. First, Table 2 indicates that students were being matched on the basis of 2015 test performance. If so, that's a huge mistake!! You can't match students on the outcome of interest! They need to be matched on the basis of prior year test performance in 2014. I'm hoping this was just a typo. Second, there are many different ways to match students after propensity scores have been estimated, and the key criterion is evidence of balance along the covariates used to estimate the propensity score. None of this evidence with regard to balance has been presented, nor do we have any sense for how many students in each group couldn't be matched.

I raise points 1 and 2 above because there is in fact good reason to worry about a mode effect in favor of PP over OL. I've just recently seen the preliminary results of two high profile testing programs finding what appear to be rather large mode effects. So if the mode effects in IN are trivial, it would come as a surprise to me. That could well turn out to be the case, but I would at a minimum need to see better answers to (1) and (2) above before I believe it."

The documentation provided by CTB in response (CTB Response for IDOE 10.20.15_FINAL.pdf) helped to clarify the design that supported the concurrent calibration approach that was used to place PP and OL items onto a common scale. What had not been evident to me was that with the exception of a small minority of IN students, all students were given a common block of PP items in "Part 1" of their test. This is indicated in the table below, pulled from page 2 of the CTB response document.

Table 1. ELA Calibration Design

Part 1 Mode	Part 1 Form	Part 1 Data			Part 2 Data		Group Name	Group Number
					MC OL/TE	MC PP/TEP		
PP	1	XXXXX XXXXX			XXXXX		PP1OL	1
						XXXXX	PP1PP	2
	2		XXXXX XXXXX		XXXXX		PP2OL	3
						XXXXX	PP2PP	4
OL	1			XXXXX XXXXX	XXXXX		OL1OL	5
						XXXXX	OL1PP	6

- XXXXX indicates blocks administered to the given group
- PP: Paper-pencil; OL: Online
- TEP: Converted PP item from OL

This common block of PP items supports the use of concurrent calibration to place PP and OL items on a common scale. Furthermore, CTB was able to show that the OL item parameters estimated from either a separate or concurrent calibration are almost perfectly correlated. A lingering threat to the validity of a concurrent calibration is the possibility of secondary and tertiary dimensions that correspond to PP and OL item formats. Results from exploratory factor analyses conducted by CTB in response to this concern indicate some evidence of multidimensionality, particularly for the ELA tests. However, the first dimension plays the dominant role in explaining inter-item covariation, and the results from this EFA are not far outside of what I have seen on other state tests. Hence while I think this is something that might be important to monitor as a possible source of item level bias (i.e., DIF), I don't suspect that it presents a problem that fundamentally undermines the evaluation of mode comparability.

One important comment in regard to a statement made in the CTB document. On p. 1 they write that "the equating design allowed for student scores in Math and ELA to be made equivalent across paper/pencil and online modes." I think this is a potentially misleading statement because it implies that mode effects have been removed in the equating process. But as we see below, that is not the case because when we form equivalent groups of students on the basis of 2014 test performance, we see instances of significant differences in test performance by mode, typically favoring students in the PP condition. I think it would be more accurate to say that the "equating" design makes it possible to place all OL and PP items onto a common scale, which is in itself to small feat.

The CTB response also helped to establish more comprehensively the approach that was taken to create equivalent groups of students by mode condition. Doing so is important because in their response document, it is clear that in general ("II.C S2014 Test Performance Summary" on p. 103), students who took the test in OL mode (i.e., PP1OL, PP2OL, OL1OL, OL1PP) tended to be have significantly higher mean scores on tests taken the previous year in 2014. Because of this, in order to estimate a mode effect by grade and subject, it is necessary to make a statistical adjustment to ensure that the two groups of students have a similar profile in terms of variables such as prior academic achievement, socioeconomic status, race/ethnicity, etc. before we compare their 2015 ISTEP+ test scores.

In their initial draft document, CTB indicated (see Table 2, page 3) that they had used 2015 test scores as covariates in a logistic regression used to estimate the propensity (probability) of each student taking a test in a particular mode. This would represent a serious flaw, because 2015 test scores are the outcome to be compared. It is critical to estimate propensity scores on the basis of variables collected prior to the outcome of interest. Furthermore, it was not made clear in the draft document how students in each grade/subject/mode were matched according to their estimated propensity scores.

In their response and in the final version of their mode comparability report, CTB has clarified that (with the exception of grade 3) they are using 2014 test scores to predict the propensity of taking the test in an OL mode. (Whether it was always the case that 2014 scores were being used or whether this was done in response to the concern I raised is not clear.) They have also clarified the approach taken to match students—they use a nearest neighbor method with replacement, the default option in the MatchIt procedure available in the R computing environment.

PSM is a complex approach, and its use as a way to estimate a causal effect (the effect of mode of test performance) depends upon the specification of the underlying logistic regression used to compute propensities, evidence that covariate balance has been obtained, and the way that subjects are matched by propensity scores. It could be argued that many variables that would help to predict why students do or do not end up taking the test in an OL mode are missing from CTB's specification: in particular, school-level variables such as mode of test taken in previous year, demographic composition and achievement profile seem highly relevant. It could also be argued that nearest neighbor matching with replacement is not the best approach to take—we have no sense for the sensitivity to the finding to choice of matching approach. And as is noted in the report, the matching approach was not always successful in producing acceptable balance among the covariates that were used to estimate propensities (see "Summary and Discussion" on page 13 of final report).

However, on the whole the approach CTB took to create equivalent groups of students by subject in grades 4 through 8 is defensible, and serves as a reasonable first order approximation of the magnitude of mode effects in these grades and subjects. We see that for ELA, the mode effects (PP-OL) are consistently positive (though often rather small when expressed in effect size units). In MA, the mode effects in grades 4-8 do not always favor PP—though small, the effects favor the OL mode in grades 5 and 7. The relevant tables with results provided in CTB's final mode comparability report are pasted below. Mode effects by grade for each subject are shown in effect size units in the last column.

Table 4. ELA Mean Differences and ES for OL and PP based on PSM Approach

Mode	Test	N Before PSM		N After PSM		PP*		OL*		PP SS-OL SS	ES
		PP*	OL*	PP*	OL*	Mean	SD	Mean	SD		
OL1OL Vs. PP1PP	EL03	12609	1127	928	1127	460.76	48.87	452.24	47.84	8.52	0.18
	EL04	9556	1085	957	1077	479.93	48.03	476.99	52.84	2.94	0.06
	EL05	8144	1189	953	1110	503.43	46.56	497.46	50.36	5.97	0.12
	EL06	10688	2426	1908	2400	528.30	51.62	526.09	55.12	2.22	0.04
	EL07	11026	2830	2174	2807	543.78	55.58	541.33	57.74	2.45	0.04
	EL08	8911	3089	2145	3052	559.19	62.66	555.35	64.02	3.84	0.06
PP1OL Vs. PP1PP	EL03	12609	26061	12609	8995	450.20	50.30	449.57	48.92	0.63	0.01
	EL04	9556	25848	9452	6030	476.04	51.79	475.80	51.92	0.24	0.00
	EL05	8144	27542	8026	5659	500.07	47.51	496.73	48.26	3.33	0.07
	EL06	10688	23522	10554	6815	521.16	52.88	517.86	53.91	3.30	0.06
	EL07	11026	23639	10859	6714	535.68	56.46	529.99	58.15	5.68	0.10
	EL08	8911	27758	8786	5786	553.60	64.00	545.98	62.88	7.62	0.12
PP2OL Vs. PP2PP	EL03	12134	23558	12134	8414	452.75	49.57	452.35	49.38	0.40	0.01
	EL04	8869	23941	8798	5787	479.76	52.01	478.29	50.23	1.47	0.03
	EL05	9063	24243	8954	6249	504.08	49.67	501.15	50.18	2.93	0.06
	EL06	8808	22872	8708	5629	521.01	55.42	518.41	57.02	2.60	0.05
	EL07	9047	23599	8937	5635	533.56	57.16	531.70	55.69	1.87	0.03
	EL08	8197	25772	8082	5348	553.20	67.43	544.42	64.28	8.78	0.13

*OL indicates Part 2 OL form; PP indicates Part 2 PP

Table 5. MA Mean Differences and ES for OL and PP based on PSM Results

Mode	Test	N Before PSM		N After PSM		PP		OL		PP SS-OL SS	ES
		PP	OL	PP	OL	Mean	SD	Mean	SD		
PP1OL Vs. PP1PP	MA03	12615	27361	12615	9109	432.06	56.15	433.74	53.60	-1.67	-0.03
	MA04	9247	27027	9145	5904	468.09	51.74	466.14	51.14	1.95	0.04
	MA05	7972	28806	7855	5547	498.79	49.94	494.88	49.66	3.91	0.08
	MA06	10537	25999	10404	7011	520.56	46.80	517.95	48.86	2.61	0.06
	MA07	11067	26574	10897	6863	535.48	50.96	530.98	47.72	4.50	0.09
	MA08	8785	30905	8659	5829	553.32	48.33	550.15	47.45	3.17	0.07
P2OL Vs. PP2PP	MA03	12092	23753	12092	8496	434.98	55.13	438.48	52.56	-3.50	-0.07
	MA04	8653	24093	8574	5632	468.92	50.38	467.95	48.91	0.97	0.02
	MA05	8868	24285	8758	6060	500.52	51.89	502.22	50.96	-1.70	-0.03
	MA06	8893	22944	8793	5750	520.91	49.55	520.80	51.23	0.11	0.00
	MA07	8899	23709	8788	5523	531.93	51.81	534.23	46.79	-2.30	-0.05
	MA08	7979	25822	7862	5284	553.49	50.85	551.05	49.18	2.44	0.05

Table 6. SC/SS Mean Differences and ES for OL and PP based on PSM Approach

Test	N Before PSM		N After PSM		PP		OL		PP SS-OL SS	ES
	PP	OL	PP	OL	Mean	SD	Mean	SD		
SCG4	16272	45986	16107	10391	419.37	56.00	415.13	55.53	-4.25	-0.08
SCG6	16666	44755	16474	10866	480.95	67.91	485.25	69.41	4.29	0.06
SSG5	1844	7369	1825	1499	500.67	73.25	505.50	73.84	4.83	0.07
SSG7	2564	6919	2538	1927	508.95	68.65	507.89	68.18	-1.06	-0.02

I am most concerned about the validity of the mode effects estimated for grade 3 MA. Here because there no prior grade test scores available (since no tests are given to students in grade 2), CTB instead used 2015 IREAD3 scores as a covariate in the estimation of propensity scores for both ELA and MA. As can be seen in Table 3 (page 4), the correlation of IREAD3 scores with ELA and MA 2015 ISTEP+ scores is .78 in ELA, but only 0.67 in MA. In contrast, for all other grades the correlation of ISTEP+ with prior year math scores is 0.80 or higher. Because of this, I would take the findings of mode effects favoring OL for grade 3 MA with a huge grain of salt. My hunch is that this is an artifact of not successfully creating equivalent groups via PSM. Unfortunately, I don't think there is much more that can be done to create more equivalent groups in MA.

I disagree with the conclusion stated on p. 13 that "In summary, no evidence of mode effects or issues with comparability across modes was found across contents and grades."

The tables shown above do indeed indicate the presence of small mode effects. CTB argues that the effect sizes are small and hence not practically significant in the sense that none are greater than 0.2 and few are greater than 0.1. According to Cohen's conventions, these are small effects. But this interpretation is not so sensible in the present context. In the way these test scores are being used in support of accountability decisions, even very small effects could have a big impact. It is true that we have uncertainty about the true magnitude of these mode effects for some of the reasons posed above about the PSM approach that was employed and the availability of key covariates for use in the PSM approach. But in the end, CTB has to stand behind their best possible estimate of grade by subject mode effects and make recommendations from on this basis.

I also disagree with the statement on p. 14 that "Although there are some items that showed mode differences for ELA and MA, this is not an issue for reporting scores, including students' scale scores and IPI scores. This is because the scale scores and IPI scores are based on the equated (mode-specific) item parameters, which account for the potential mode effects through the calibration design." If forms were successfully equated, then students (and schools) would be indifferent as to which mode was used to administer the ISTEP+. It follows that if randomly equivalent groups of students took the ISTEP+ in each mode, we should expect to observe the same mean score beyond differences due to chance variability in random assignment. The point of conducting a PSM is to approximate random assignment. To the extent this was successful, it does not appear that students/schools would consistently be indifferent to the mode in which the test was administered. Now to be sure, some of the observed differences are small enough to be explained by chance, but obtaining unbiased standard error estimates in PSM is not straightforward, and none have been provided by CTB in their analysis, so we can't evaluate this formally at the present time. But other observed differences are clearly of practical and statistical significance given the magnitude of effect size and relative sample sizes for each group (EL05, EL07, EL08). From a policy perspective it seems important to communicate to stakeholders that they will not be disadvantaged because they were "early adopters," even if it is true that some of the adjustments in questions are incredibly small and could be

explained by chance.

Recommendations

My short term recommendation is to, at a minimum, examine the potential consequences of mode effects on accountability decisions. This could be done by adding the mode effect to the scale scores of each student and then running Indiana's growth model again with these adjusted values to see if it leads to any changes in school classifications. If any schools shift upwards, it would seem wise to give them the benefit of the doubt. As a concrete example, for students taking the test in the OL1OL condition for EL05, the mode effect is 5.97 scale score points (for an effect size of 0.12). So for every student taking the test in the OL1OL condition, I would recommend adding 6 scale score points to their score, and then use this adjusted data set to feed into the state's growth model.

With respect to grade 3 MA, I would base an adjustment on the average mode effect detected in grades 4-8 where a stronger case can be made for successfully creating equivalent groups. So for example, in the PP1OL mode, the average effect for grades 4-8 was .068 favoring PP. I would translate this into scale score units for grade 3 and then apply the same scale score adjustment as described above.

A policy decision will need to be made about whether it would be sensible to apply the same adjustment approach to the few remaining grades/subjects in which there is a mode effect in favor of OL. A good case could be made for always making an adjustment based on estimated mode effect (whether it favors PP or OL), or for only making an adjustment when students/schools would be disadvantaged by taking the OL mode. The latter policy creates an incentive for more schools to move to the OL format in the future.