Investigation of Methods for Improving Reliability of Claim Scores

The SBAC assessments report not only a total scale score, but also content-specific claim scores for each content area for each student who took the test (Connecticut State Board of Education, 2016). The content-specific claims were developed in alignment with Connecticut Core Standards, and the claim-level scores were intended to provide information about the knowledge and skills students demonstrated on the assessment. Such diagnostic information is desirable and important as it allows teachers to improve instruction to meet the specific needs of individual students and allows parents to understand their children's performance in alignment with Connecticut Core Standards. First, validity and reliability evidences were gathered to support the use of students' total scores. However, there is legitimate concern about the reporting of claim-level scores as reliability evidence for claim-level scores is lacking. Because claim-level scores are based on small numbers of items, they are necessarily less reliable than total scores. This unreliability has implications for classification decisions. Student reports on SBAC tests include classification into one of three performance levels for each claim (Below Standards; Approaching Standards; Above Standards). The classification decision is made by first computing an error band for each student's claim-level score to take into account the conditional standard error of measurement. If the band falls fully within one of the performance levels, the student is classified in that level. However, in many cases the error band is so wide that it crosses two levels, with the result that a large proportion of students fall into the indeterminate middle level. The large standard error and corresponding low reliability of claim-

level scores therefore makes the utility of such classifications questionable. Additionally, it makes the standard setting for claim-level scores more difficult.

An obvious but impractical solution to the problem of the unreliability of the claim score is to increase the number of items measuring each claim. The purpose of this study was to investigate an approach to improving sub-score reliability without increasing test length. This approach is referred to as the augmented score method (Wainer et al., 2001).

The basic idea of the augmented score method is to "use ancillary information to increase the precision of estimates" (Wainer et al., 2001, p. 346). More specifically, augmented scores are obtained by using collateral information from the performance of comparable group of examinees and weighting this information by the score reliability:

$$\theta_{aug} = r\theta_o + (1 - r)\overline{\theta_o}$$

where θ_o is the observed score, $\overline{\theta_o}$ is the group mean, and r is the sample estimate of reliability. However, in the SBAC context, the claim-level score is not an observed score, but rather is an IRT estimated score. Because the IRT estimates of claim-level scores are already shrunk toward the mean in the estimation process, there is an extra "unshrink" process before they can be used as observed scores. The unregressed score θ_o is obtained from the estimated claim level score for claim *s*, denoted as θ_s, as follows:

$$\theta_o = \frac{\theta_s}{r_s}$$

(Wainer et al., 2001, p. 367). The sample estimate of reliability r_s in the IRT framework is calculated as

$$r_s = \frac{Variance(\theta_s)}{Variance(\theta_s) + Average(SE^2(\theta_s))}.$$

Augmented sub-scores can then be computed by

$$\theta_{aug} = \overline{\theta_o} + B * (\theta_o - \overline{\theta_o})$$

where B is a matrix that is the multivariate analog for the estimated reliability, derived from

$$B = Cov_T * (Cov_o)^{-1}$$

The off diagonal elements of Cov_o and Cov_T are equivalent, but the diagonal elements of Cov_T are the true score variance while the diagonal elements of Cov_o are the observed score variance.

The data used in this study was from the 2016-2017 Smarter Balanced ELA test. The augmented scores were calculated for Grade 3 – 8 students.

Results

The reliability coefficients for the augmented scores (Aug) in each grade for Reading (R), Speaking and Listening (S&L), and Writing (W) are shown in Table 3 in comparison with reliability coefficients based on original reported claim-level scores.

Table 3. Comparison of Original and Augmented Sub-score Reliability Coefficients

	N(R)	R	Aug_R	N(S&L)	S&L	Aug_S&L	N(W)	W	Aug_W
GRADE3	15	0.83	0.92	8	0.71	0.88	16	0.86	0.92
GRADE4	15	0.79	0.89	8	0.73	0.88	16	0.83	0.91
GRADE5	15	0.80	0.92	8	0.73	0.90	16	0.85	0.92
GRADE6	14	0.80	0.90	8	0.68	0.89	16	0.84	0.91
GRADE7	15	0.82	0.91	8	0.69	0.88	16	0.83	0.91
GRADE8	16	0.82	0.91	8	0.67	0.87	16	0.84	0.91

After augmentation, the reliability coefficient of each claim was improved on average by .12. Reliability of the claim of Speaking and Listening and improved the most. The S&L claim has the smallest number of items, hence its original marginal reliability was the lowest among the three claims. By incorporating the information from other claims, the reliability of the claim that provides the least information originally is improved most.

Table 4 shows the average SEM based on original claim level scores and average SEM based on augmented scores. The standard errors decreased for all sub-scores, with substantial improvement for Listening & Speaking.

Table 4. Comparison of Original and Augmented Sub-score Average Standard Error

Grade	NR	SEM R	SEM_ Aug_R	N(S&L)	SEM S&L	SEM Aug_ S&L	N(W)	SEM W&I	SEM Aug_W
3	15	0.54	0.39	8	0.88	0.35	16	0.47	0.45
4	15	0.63	0.42	8	0.89	0.42	16	0.53	0.50
5	15	0.64	0.40	8	0.89	0.36	16	0.51	0.51
6	14	0.65	0.45	8	0.97	0.33	16	0.54	0.54
7	15	0.61	0.47	8	0.94	0.38	16	0.59	0.49
8	16	0.61	0.47	8	1.02	0.39	16	0.58	0.52

Table 5 shows the comparison of classifications using the original claim level scores and augmented scores. As can be seen in the table, a smaller proportion of students were classified as Approaching Standards due to the improvement of reliability and precision.

For Reading and Writing, 80% of the students across grades remained in the same level, around 9% of them were classified to a lower level and 11% of them were classified to an upper level. For Listening, 75% of the students across grades remained in the same level, but in the grades for which reliability was lower the percent of students remaining in the same level was lower (e.g., in Grade 6 only 70% of students remained the same level). Of the students who changed levels, 11% were classified to a lower level and 14% were classified to an upper level.

Table 5. Comparison of Original and Augmented Sub-score Classifications

Grade		R			S&L			W		
		L1	L2	L3	L1	L2	L3	L1	L2	L3
3	Original	29	42	29	16	62	23	29	41	29
	Augmented	34	30	36	28	38	34	34	30	36
4	Original	22	48	31	19	57	24	27	45	28
	Augmented	30	34	36	29	38	33	32	31	36
5	Original	23	45	32	17	59	25	27	41	32
	Augmented	30	30	40	28	33	39	31	29	40
6	Original	25	49	27	15	64	22	27	45	28
	Augmented	30	33	37	26	34	40	31	32	38
7	Original	24	45	31	18	64	18	25	47	29
	Augmented	30	33	37	27	38	35	30	33	36
8	Original	26	44	30	14	65	21	28	45	27
	Augmented	31	32	37	26	38	36	32	32	36

In each grade, a few students changed to two levels higher or lower, which made a great difference in their score interpretation. Closer examination revealed that these students all had unbalanced performance levels on the subscales. For example, students who were classified two levels lower in Reading after score augmentation all had the highest performance level in Reading originally, but low performance levels in both Writing and Listening. Students who were classified two levels upper in Writing after score augmentation all had the lowest performance level in Writing originally, but high performance levels in both Reading and Listening. Because the augmented scoring method assumes the subscales are correlated with each other, the performance levels across subscales after augmentation are more similar.

Conclusions

In general, the augmented scoring method improved the reliability and precision of claim level scores. However, in cases where performance is very uneven across subscales, the procedure may over-correct.