INVESTIGATING MULTIDIMENSIONALITY IN STUDENT RESPONSES TO THE COMPREHENSIVE ASSESSMENT OF OUTCOMES IN STATISTICS

V.N. VIMAL RAO
University of Minnesota
rao00013@umn.edu

ABSTRACT

The Comprehensive Assessment of Outcomes in Statistics (CAOS) is a test designed to measure students' conceptual understanding of basic learning outcomes after a first course in statistics). Research utilizing CAOS has sometimes included an analysis of items by topic, tantamount to the utilization of subscores. To date only one study of multidimensionality has been conducted utilizing CAOS data. In this paper I extend previous analyses of dimensionality of student responses to CAOS by applying multidimensional item response theory (MIRT) methods to 14 years of CAOS data in an attempt to provide validity evidence for the use of subscores. Results indicate that a bifactor model fits the data better than a unidimensional model, providing evidence of multidimensionality. However, further analysis is required in order to fully establish the validity of the use of CAOS subscores.

Keywords: *Statistics education research; CAOS; multidimensional item response theory; multilevel modeling*

1. INTRODUCTION

The Comprehensive Assessment of Outcomes in Statistics (CAOS) was developed as part of the ARTIST project (delMas, Garfield, & Chance, 2003; Garfield, delMas, & Chance, 2002) to measure students' conceptual understanding of basic learning outcomes after taking a first course in statistics (delMas, Garfield, Ooms, & Chance, 2007). The test was designed to focus on students' reasoning about variability, thought to be the main learning objective of a first course in statistics that follows the *consensus curriculum* (Cobb, 2007). This construct, reasoning about variability, was thought to include reasoning about distributions, comparing groups of data, sampling, and sampling distributions (delMas et al., 2007). Items on the test were designed to address both students' statistical reasoning and statistical literacy (delMas, Garfield, & Ooms, 2005).

To date there has only been one study examining the dimensionality of the latent traits governing student responses to CAOS. delMas (2014) conducted a confirmatory factor analysis (CFA) of a unidimensional model on data collected between 2005 and 2013 and found strong evidence suggesting that CAOS measures a single construct, which was thought to represent students' statistical understanding of concepts taught in introductory statistics courses.

Despite a lack of statistical evidence of multidimensionality, CAOS subscores have been reported and analyzed in teaching and research. Instructors utilizing CAOS receive item analysis reports that summarize student performance on the test and include the distribution of students' score by topic. Empirical studies have also included analyses of student scores by topic (delMas, 2014; Tintle et al., 2014).

Without analyses of multidimensionality of latent traits measured by CAOS, the validity of subscore analysis, and whether or not such an analysis provides any meaningful summary information upon which inference can be based, remains undetermined.

2. BACKGROUND

2.1. CAOS TOPICS

While not including the calculation of subscores, delMas et al. (2007) included a discussion of students' gains in understanding by topic area. This discussion partitioned the 40 CAOS items into 10 topic areas representing data collection and design, descriptive statistics, graphical representations, boxplots, the normal distribution, bivariate data, probability, sampling variability, confidence intervals, and tests of significance.

Tintle et al. (2011) utilized CAOS to assess the effects of different curricula on students' understanding of statistics. This analysis included an evaluation of student performance by topic. The 40 CAOS items were grouped according to the framework presented by delMas et al. (2007) with the exception of the normal distribution, resulting in the use of nine total topic areas (Table 1). These nine topics were similarly utilized in subsequent studies (Tintle et al., 2012; Tintle et al., 2014).

Table 1. Items by topic as analyzed by Tintle et al. (2011)

	Item Numbers
Data collection and design	7, 22, 24, 38
Descriptive statistics	14, 15, 18
Graphical representations	1, 3, 4, 5, 6, 11, 12, 13, 33
Boxplots	2, 8, 9, 10
Bivariate data	20, 21, 39
Probability	36, 37
Sampling variability	16, 17, 32, 34, 35
Confidence intervals	28, 29, 30, 31
Tests of significance	19, 23, 25, 26, 27, 40

DelMas (2014) found that IRT models poorly described student performance on item 32, and therefore excluded it from analyses. Furthermore, since CAOS contains many testlets, i.e., a group of items developed as a single unit to be administered together

(Wainer, Bradlow, & Wang, 2007), delMas compressed items belonging to a testlet in a manner consistent with a graded response model (GRM; Samejima, 1969) to control for significant local dependence of items affecting model fit. This resulted in the reduction of 40 total items to 24 locally independent items, and prompted a revision of the partition of items by topic, since many of the original topic areas were left with only a single composite item. delMas thus collapsed the original 10 topic areas into six topic areas, as is detailed in Table 2, which were used to analyze the stability of student performance by topic across time.

Table 2. Items by topic as analyzed by delMas (2014)

	Item Numbers
Data collection and design	7, 22, {23, 24}, 37, 38
Variability	{14, 15}, 18
Graphical representations	{1, 2}, {3, 4, 5}, 6, {8, 9, 10}, 33
Bivariate data	20, 21, 36, 39
Sampling variability	16, 17, {34, 35}
Tests of significance	{11, 12, 13}, 19, {25, 26, 27}, {28, 29, 30, 31}, 40

*items belonging to a single testlet are grouped by brackets

2.2. SUBSCORE VALIDITY

Standard 1.14 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) states that "When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated" (p. 27).

While these standards typically apply to large scale high-stakes testing and evaluation, the use of CAOS subscores in research may influence curricular design while its inclusion on instructor reports may influence classroom instruction. Therefore, although a less rigorous standard may be acceptable in this scenario, the necessity of establishing subscore validity is not wholly obviated.

Reliability Haberman (2008) proposed a criterion for determining when subscores can be useful, based on the reduction of root mean squared error, to assess whether the subscore provides more meaningful and reliable information on the construct it purports to measure when compared to the total scores. The method can be reframed in terms of the relative estimates of subscore reliability and total score reliability. However, Sinharay, Haberman, and Puhan (2007) note that rarely does this criterion result in a determination that subscores provide meaningful information worthy of reporting.

Simulation studies conducted by Sinharay (2010) found that as the number of items per subscore decreases the number of subscores that add meaningful information by Haberman's criteria decreases, since the reliability of scores decreases as the number of items decrease. The smallest number of items per subscore Sinharay simulated was 10, much larger than the number of items belonging to many of the topics on CAOS as utilized by delMas (2014) or Tintle et al. (2011). As such, it is unlikely that nine or even six subscores of CAOS will provide meaningful information under Haberman's criteria.

Feinberg and Jurich (2017) extended Haberman's criteria to include consideration of statistical significance, and use this framework to make recommendations when generation of subscores would be harmful and misleading. Yet, they note that even this slightly relaxed criterion is still unlikely to be regularly met in practice, especially when a test has been constructed to achieve unidimensionality.

Similarly, Meijer et al. (2017) note that unless a test is intentionally designed to provide meaningful subscores, it is rare that subscores will be rated as useful under Haberman's criteria. While CAOS was designed with a single statistical reasoning construct in mind, reasoning about variability, it was believed that this main construct has several distinct components in addition to various content emphases (delMas et al., 2007). However, the inclusion of specific subcontent areas alone may not be a sufficient basis for the reporting of subscores (Biancarosa et al., 2019).

Therefore, it may be that these methods of establishing validity evidence based on analyses of the reliability of subscores are inapplicably strict with regards to CAOS and its typical use. As such, I instead focus the identification of validity evidence for subscore use on their distinctiveness.

Distinctiveness A common method for determining the distinctiveness of subscores is an analysis of the correlations between subscores and their correlation to the total score (Lyren, 2009). Haberman and Sinharay (2010) found that subscores generated from multidimensional item response theory (MIRT) models were more reliable than raw subscores but also more highly correlated, or less distinct. Reise, Moore, and Haviland (2010) argue that bifactor models are particularly effective in establishing validity evidence of distinctness for the use of subscores as by definition secondary factors are uncorrelated – in bifactor models, all items load on a general common factor, with subsets of items clustered into orthogonal secondary factors that account for additional variance not accounted for by the common factor (Holzinger & Swineford, 1937). Similarly, Li, Jiao, and Lissitz (2012) explicitly link analysis of MIRT models in the context of establishing subscore validity to the general task of examining multidimensionality in student response patterns on tests. Therefore, I focus the investigation of distinctiveness on the analysis and establishment of multidimensionality in student responses to CAOS through an evaluation of candidate bifactor models.

CAOS dimensionality While no study has comprehensively examined and investigated multidimensionality in student responses to CAOS, factor analyses of similar assessments have generally failed to establish evidence of multidimensionality. Sabbag (2016) found that while the Reasoning and Literacy Instrument (REALI), designed to measure student statistical literacy and reasoning simultaneously, was hypothesized as measuring multiple constructs, after accounting for model parsimony and validity of subscores, a unidimensional model was more appropriate than a bifactor model to explain students' responses on the assessment.

Similarly, Allen (2006) found that a multifactor model fit data collected from 295 students taking the Statistical Concept Inventory (SCI), developed to assess students' conceptual understanding, better than a unidimensional model but at great cost to parsimony, and determined that a unidimensional model fit the data sufficiently well.

While neither Allen (2006) nor Sabbag (2016) found strong evidence of multidimensionality in their tests, both studies identified possible multidimensionality when analyzing the fit of bifactor models and, perhaps due to paucity of sample size, were unable to detect multidimensionality with any power and statistical significance. The largest study examining dimensionality for CAOS was conducted by delMas (2014) and included a sample size of over 23,000, but did not include examinations of multilevel models.

2.3. RESEARCH QUESTION

In order to inform the use of subscores on CAOS, and to examine validity evidence for their use, I aim to answer the following research question: Is there evidence of multidimensionality in students' responses to CAOS items that can be modelled by a bifactor model?

Although the process of establishing validity evidence for the use of subscores includes further analyses than simply establishing multidimensionality, this is a prerequisite first step in any such analysis. Therefore, in the present study I simply focus methods on examining the dimensionality of student responses to CAOS in order to provide justification for further inquiry.

The bifactor model performed well in modeling similar assessments as determined by Allen (2006) and Sabbag (2016), and due to its structure, requires secondary factors to have a zero correlation, thus increasing the chances that subscores generated by the model meet criteria for distinctiveness (Haberman & Sinharay, 2010). Bonifay, Lane, and Reise (2017) note that bifactor models often overfit data. However, this may be controlled by the use of multiple model comparison statistics that penalize overfit, such as the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). Henson, Reise, and Kim (2007) found that while the AIC still preferred models that overfit data, the BIC preferred models that underfit data. Therefore, it may be that a bifactor model preferred by both the AIC and BIC is one without inappropriate overfit. Thus, as a first step in validating subscores, the bifactor model may still be an appropriate model to fit to data before further analyses are conducted to ensure the usefulness of subscores (Reise, Moore, & Haviland, 2010).

3. METHODS

3.1. CAOS DATA

The CAOS test is comprised of 40 multiple choice items and is administered online in a forced choice format (delMas et al., 2004). There are nine testlests in the test which includes a total of 24 of the 40 items. The median completion time is approximately 27 minutes, with the middle half of all students completing the test in between 20 and 35 minutes. Median end-of-course performance on the test has risen modestly from a score of 50% on tests taken between 2006–2008 to a score of 52.5% on tests taken between 2015-2017.

Since its first offering in the 2005–2006 academic year, CAOS has become one of the most widely used assessments of statistical reasoning. As of the 2017–2018 academic year, it has been taken by over 48,000 students and has been used by 239 different instructors from 167 different institutions across the United States.

All students completing CAOS between 2005 and 2018 were initially considered for inclusion in the present analysis. In order to ensure a homogenous population, only students at tertiary-level educational institutions were included. Furthermore, in order to ensure effortful responses from all students, only students with completion times between 10 minutes and 60 minutes were included. These inclusion criteria match those utilized by delMas (2014). A total of 41,209 student responses met these criteria.

3.2. Candidate Models

delMas (2014) determined that a testlet model fit CAOS data better due to the presence of local dependence of items when compared to standard unidimensional item response theory (IRT) models. Thus, a unidimensional testlet model will similarly be utilized as a base model in this study. A GRM will be fit to testlets (Cook, Dodd, & Fitzpatrick, 1999) while a two-parameter logistic model (2PL; Birnbaum, 1968) will be fit to all other items. In order to establish validity evidence for the use of CAOS subscores, each topic must be distinct, and I consider three candidate bifactor models utilizing a confirmatory modeling approach (Gibbons & Hedeker, 1992; Joreskog, 1969).

Although in this analysis I take an IRT-based approach, this method is mathematically equivalent to non-linear factor-analytic approaches and thus is a valid extension of previous research examining the dimensionality of CAOS data and similar assessments of statistical reasoning (Kamata & Bauer, 2008; Reise, 2012). All models will be fitted by utilizing the mirt package in R (Chalmers, 2012) with a convergence tolerance of 0.01.

The first candidate model to be evaluated, the six-topic bifactor model, is based on topics as analyzed by delMas (2014). Since some of the topics in this categorization have as few as two items, I also consider a more parsimonious three-topic bifactor model that categorizes the CAOS items into three separate topics representing data collection, descriptive statistics, and inferential statistics, as enumerated in Table 3.

Table 3. Items by topic in the three-topic bifactor candidate model

	Item Numbers
Data collection	7, 22, {23, 24}, 37, 38
Descriptive statistics	{1, 2}, {3, 4, 5}, 6, {8, 9, 10}, {14, 15}, 18, 20, 21, 33, 36, 39
Inferential statistics	{11, 12, 13}, 16, 17, 19, {25, 26, 27}, {28, 29, 30, 31}, {34, 35}, 40

*items belonging to a single testlet are grouped by brackets in this and subsequent tables

Furthermore, the field of statistics is often dichotomized into descriptive statistics and inferential statistics. Therefore, the final and most parsimonious candidate model that will be considered will be the two-topic bifactor model. The two topics are intended to represent inferential statistics and descriptive statistics. Items will be partitioned in this model in a manner similar to the partition of items in the three-topic model but with the data collection and descriptive statistics topics combined.

3.3. Model evaluation

A popular method for assessing item-level fit for IRT models is the $S\text{-}X^2$ test statistic (de Ayala, 2009; Kang & Chen, 2008; Orlando & Thissen, 2000). However, this fit index is highly sensitive to sample size, resulting in inflated Type 1 error rates in large samples (Chon, Lee, & Dunbar, 2010). With a sample size of over 41,000 in the present analysis, this measure is unlikely to provide any meaningful information when applied to fitted models. As such, the root mean squared error of approximation (RMSEA) for each item will be used to assess the magnitude of misfit (von Davier, 2008). Items with good fit should return an RMSEA below 0.05 (Kunina-Habenicht, Rupp, & Wilhelm, 2012).

The G^2 statistic (McKinley & Mills, 1985) and its associated *p*-value in addition to the RMSEA (Maydeu-Olivares, Cai, & Hernandez, 2011) are two common methods to assess model fit of IRT models (de Ayala, 2009). Models with good fit should return a non-significant G^2 *p*-value and an RMSEA below 0.05 (Browne & Cudeck, 1993).

The AIC, BIC, and the likelihood ratio test (LRT; Neyman & Pearson, 1933) are three common methods to compare IRT models (de Ayala, 2009). Each of the candidate models will be compared against the unidimensional model by these criteria to determine if the candidate model provides a better fit for the data and thus evidence of multidimensionality. The preferred model will be one with lower AIC and BIC values and that returns a statistically significant LRT. In order to unambiguously determine which model is a better fit, I will require that a candidate model outperform the unidimensional model on all three model comparison methods.

The explained common variance (ECV) statistic is a measure for assessing the degree of unidimensionality in bifactor models (Bentler, 2009; ten Berge & Socan, 2004). Unfortunately, no clear practical standard exists for determining meaningful thresholds for interpretation of the statistic (Reise, 2012). However, high values of ECV, approaching 1.0, can generally be interpreted as signifying that the general factor in the model explains a larger proportion of the variance in the data than the secondary factors and that a unidimensional model may be appropriate.

Reise (2012) notes that it is generally advisable to perform exploratory bifactor analyses in addition to confirmatory bifactor analyses. Despite taking a confirmatory modeling approach, I will also conduct exploratory bifactor modeling analyses to determine the optimal partitions of items for either two, three, or six factors using the

Schmid-Leiman orthogonalization method (SL; Schmid & Leiman, 1957; Waller, 2018). I will compare bifactor models utilizing the SL partitions of items to their heuristically defined equivalent models, as well as the unidimensional model, in order to explore the degree to which the content-based heuristic partition is valid.

Reise (2012) also argues for an analysis of the invariance of the general factor in any bifactor model to confirm the validity of the model and the general factor as a reflection of the true common variance shared by all items. A general factor that is invariant will result in similar factor loadings regardless of which subset of item content domains are included. However, Reise does not suggest a formal statistical test for invariance. Rather he proposes an examination of results from models fitted to multiple random subsamples of items. Factor loadings of items onto the general factor in item subsamples that are consistent with item factor loadings in the full model are considered to be an indicator of validity evidence for the model. Therefore, an examination of general factor item loadings from subsamples of items will be compared for the best fitting bifactor model to ascertain the level of invariance.

4. RESULTS

In general, none of the fitted models exhibited significant misfit for any item, as is summarized in Table 4, with all RMSEA item fit statistics well under the 0.05 threshold for good fit. Using the RMSEA as a relative fit index, both the unidimensional model and the two-topic bifactor model appeared to generally provide for better item level fit than the three-topic bifactor model or the six-topic bifactor model. However, this may also be a consequence of the choice of the 2PL and GRM models for item responses, and may not necessarily reflect the structure of the data itself.

Table 4. RMSEA item-fit statistics by item and model

	Unidimensional model	Two-topic bifactor model	Three-topic bifactor model	Six-topic bifactor model
{1, 2}	0.012	0.011	0.015	0.039
{3, 4, 5}	0.013	0.013	0.013	0.016
6	0.010	0.009	0.013	0.027
7	0.018	0.016	0.019	0.019
{8, 9, 10}	0.015	0.014	0.016	0.014
{11, 12, 13}	0.013	0.013	0.014	0.015
{14, 15}	0.005	0.005	0.006	0.014
16	0.010	0.011	0.017	0.036
17	0.010	0.010	0.011	0.012
18	0.009	0.008	0.010	0.011
19	0.006	0.005	0.007	0.018
20	0.009	0.011	0.011	0.042
21	0.009	0.009	0.009	0.015
22	0.005	0.005	0.005	0.004
{23, 24}	0.009	0.008	0.010	0.018
{25, 26, 27}	0.025	0.025	0.025	0.025
{28, 29, 30, 31}	0.006	0.006	0.008	0.020
33	0.012	0.011	0.011	0.019
{34, 35}	0.010	0.010	0.011	0.025
36	0.006	0.006	0.008	0.013
37	0.021	0.019	0.022	0.020
38	0.003	0.003	0.004	0.004
39	0.016	0.014	0.015	0.018
40	0.007	0.007	0.010	0.018

Similarly, all four models exhibited strong overall fit as summarized by Table 5. All RMSEA model fit statistics were, once more, well under the 0.05 threshold for good fit, and p-values of the G^2 test statistic well above thresholds for identifying a statistically significant model misfit.

A comparison of all models using the AIC and BIC criteria suggests that the two-topic bifactor model is the best model to fit the data, as can be seen in Table 5. The LRT, used in order to assess the degree to which the two-topic bifactor model may fit the data better than the unidimensional model, results in a *p*-value less than 0.0001, indicating that indeed the two-topic bifactor model fits the data better than the unidimensional model, and contributes a statistically significant improvement in model fit over the simpler unidimensional model. This result is consistent with model comparisons utilizing the AIC and BIC criteria. Similar LRTs comparing the two-topic bifactor model to the three-topic bifactor model and the six-topic bifactor model imply that model fit is not improved in extending the number of topics to three or six, with *p*-values well above 0.20 in both cases.

The ECV of the two-topic model is approximately 0.82. Factor loadings of items in a unidimensional model are consistent with item factor loadings onto the general factor in the two-topic bifactor model, as seen in Table 6. Reise, Moore, and Haviland (2010) argue that when this is the case, it may be that a unidimensional model sufficiently explains the variation in the data. This is consistent with the interpretation of the ECV. Therefore, while there is statistically significant evidence suggesting that a bifactor model is a better fit for the data than a unidimensional model, further evaluation must be conducted to determine the practical significance of this improvement.

In comparing the SL and heuristic two-factor partitions of the 24 composite items of CAOS, 14 items are aligned similarly in both approaches, as displayed in Table 7. Bifactor models based on SL partitions of items in three factor models and six factor models do not result in improvements over the heuristically determined two-topic bifactor model. However, a comparison of the SL two-topic bifactor model outperforms the heuristically determined two-topic bifactor model when compared under AIC, BIC, and LRT criteria. Therefore, further analysis is warranted to examine the common traits of items aligned according to the SL procedure, which may lead to the creation of more meaningful subscores than the heuristic partition of items.

Table 5. Absolute and comparative model fit statistics by model

	Unidimensional model	Two-topic bifactor model	Three-topic bifactor model	Six-topic bifactor model
G^2	610,106	607,606	608,357	608,593
p-value for G^2	> 0.999	> 0.999	> 0.999	> 0.999
RMSEA	< 0.001	< 0.001	< 0.001	< 0.001
AIC	1,484,029	1,481,577	1,482,328	1,482,564
BIC	1,484,573	1,482,327	1,483,078	1,483,315

Table 6. Factor loadings onto the general factor by item and model

	delMas (2014)	Unidimensional model	Two-topic bifactor model
{1, 2}	0.150	0.174	0.167
{3, 4, 5}	0.547	0.634	0.629
6	0.429	0.544	0.549
7	0.285	0.461	0.464
{8, 9, 10}	0.495	0.533	0.537
{11, 12, 13}	0.391	0.525	0.525
{14, 15}	0.465	0.515	0.511
16	0.517	0.599	0.608
17	0.357	0.398	0.406
18	0.221	0.352	0.328
19	0.346	0.465	0.458
20	0.178	0.511	0.416
21	0.163	0.296	0.268
22	0.289	0.356	0.353
{23, 24}	0.247	0.311	0.302
{25, 26, 27}	0.439	0.492	0.489
{28, 29, 30, 31}	0.381	0.399	0.387
33	0.234	0.275	0.287
{34, 35}	0.297	0.326	0.329
36	0.364	0.446	0.445
37	0.302	0.388	0.404
38	0.344	0.395	0.397
39	0.225	0.323	0.332
40	0.163	0.473	0.469

Table 7. SL and heuristic item partition for a two-factor model

Factor 1	Descriptive Statistics	Factor 2	Inferential Statistics
{1, 2}	{1, 2}		
{3, 4, 5}	{3, 4, 5}		
	6	6	
	7	7	
	{8, 9, 10}	{8, 9, 10}	
{11, 12, 13}			{11, 12, 13}
{14, 15}	{14, 15}		
		16	16
17			17
18	18		
		19	19
20	20		
21	21		
	22	22	
{23, 24}	{23, 24}		
		{25, 26, 27}	{25, 26, 27}
		{28, 29, 30, 31}	{28, 29, 30, 31}
	33	33	
		{34, 35}	{34, 35}
36	36		
	37	37	
	38	38	
	39	39	
		40	40

Examination of general factor item loadings from models fitted to student responses to three different randomly selected subsamples of 15 items, as displayed in Table 8, shows consistency across the subsamples. Furthermore, unidimensional models only including items for each of the two hypothesized content domains of the two-topic bifactor model, descriptive statistics and inferential statistics, also generate similar item factor loadings. Thus, the criteria set forth by Reise (2012) for general factor invariance in a bifactor model appears to be satisfied.

Table 8. General factor loadings in the full model and models fitted to subsets of items

Item	Full Model	Descriptive Items	Inferential Items	Random Item Subsample 1	Random Item Subsample 2	Random Item Subsample 3
{1, 2}	0.167	0.164	-	-	0.158	-
{3, 4, 5}	0.629	0.629	-	0.614	0.621	-
6	0.549	0.561	-	0.563	-	0.523
7	0.464	0.461	-	0.474	-	0.436
{8, 9, 10}	0.537	0.554	-	-	0.538	-
{11, 12, 13}	0.525	-	0.469	0.528	-	0.471
{14, 15}	0.511	0.527	-	0.506	0.509	0.505
16	0.608	-	0.604A	-	0.604	0.610
17	0.406	-	0.428	-	-	0.395
18	0.328	0.344	-	0.295	-	-
19	0.458	-	0.468	0.466	0.441	-
20	0.416	0.494	-	-	0.353	-
21	0.268	0.290	-	0.285	-	0.257
22	0.353	0.352	-	0.367	-	0.335
{23, 24}	0.302	0.299	-	0.306	0.284	-
{25, 26, 27}	0.489	-	0.518	0.492	-	-
{28, 29, 30, 31}	0.387	-	0.421	0.399	0.371	-
33	0.287	0.279	-	-	0.309	0.266
{34, 35}	0.329	-	0.298	-	0.338	0.338
36	0.445	0.445	-	0.422	0.446	0.460
37	0.404	0.377	-	-	0.434	0.397
38	0.397	0.386	-	-	0.399	0.413
39	0.332	0.327	-	0.346	-	0.290
40	0.469	-	0.474	0.473	0.461	0.481

An examination of the consistency of models in this analysis compared to that by delMas (2014) indicates that factor loadings were generally consistent for all items with the exception of items 7, 20, and 40. Items 7 and 20 were identified by delMas (2014) as potentially representing content not included in instruction for courses utilizing CAOS. Noticeable increases in the factor loadings was seen on all items identified as having factor loadings less than 0.3 by delMas (2014) except the testlet item 1 and 2, as shown in Table 6. As this analysis utilized MIRT models to estimate item factor loadings, while delMas

(2014) utilized a CFA, differences in item factor loadings may be due to improved item fit with the use of a 2PL model.

Furthermore, the reliability of the unidimensional model in the present study was estimated to be approximately 0.784, nearly equivalent to delMas's (2014) estimate of 0.75. This was also consistent with the two-topic bifactor model estimate of the general factor score reliability of 0.780. This replication of results from delMas (2014) adds further validity evidence for the comparison of past and present analyses.

5. SUMMARY

5.1. DISCUSSION

In order to establish evidence of multidimensionality in students' responses to CAOS for the purpose of justifying the use of subscores, I examined three candidate bifactor models and compared them to a unidimensional IRT model. A two-topic bifactor model outperformed the unidimensional model on all model comparison tests considered. This suggests that there may indeed be a multidimensional structure to latent traits governing student responses on CAOS.

However, while there is statistical evidence in support of this claim, further analysis needs to be conducted to determine whether this difference is of any meaningful or practical significance. Furthermore, additional investigation must be conducted to fully establish validity evidence for the use of subscores from CAOS. Nevertheless, the establishment of multidimensionality is an important first step in this validation process.

Although there is evidence of multidimensionality, the best fitting model only includes two secondary factors, which are intended to represent content and reasoning related to descriptive statistics and inferential statistics respectively. Therefore, it remains doubtful that there is sufficient evidence to justify the use and analysis of subscores aligned with either six (delMas, 2014) or nine (Tintle et al., 2014) topics.

Additionally, further study is required to validate that the two topics do indeed represent descriptive statistics and inferential statistics constructs before investigation and discussion of the use of subscores representing these subfactors can be continued. It may also be useful to partition items into two groups for creating subscores. Nevertheless, the establishment of statistical evidence of multidimensionality in student responses to CAOS is an important prerequisite for such efforts.

5.2. Limitations

While the identification of multidimensionality is an important result, the lack of similar findings in studies with smaller sample sizes indicates that the power of statistical tests employed in this study may be of questionable quality, despite the use of robust model comparison measures. Therefore, a power analysis or replication of these methods must be conducted in order to verify that the result is not a false positive.

Although the scope of this study is by definition limited to students completing CAOS, it is unknown whether the population of institutions, instructors, and students utilizing CAOS is in any way representative of the larger community of statistics educators and students. Furthermore, it is possible that the characteristics of CAOS users has changed over time, thus limiting the generalizability of these results.

Similarly, CAOS was designed to conform to courses following the consensus curriculum. However, many new curricula have since been developed and adopted which focus on simulation-based methods (Garfield, delMas, & Zieffler, 2012; Lock et al., 2013; Tintle et al., 2015). It may be that there is no longer full alignment between CAOS and classroom instruction for students completing CAOS. The adoption of these curricula has occurred in recent years, and is not reflected in previous analyses of CAOS to the extent which it may be reflected in the present analysis. Thus, improvements in statistics education as a result of these new curricula may have led to increased reliability of student responses to CAOS in recent years. While this is a boon to analyses of student performance on CAOS, it implies that there may be significant *differential item functioning* (DIF; Holland & Thayer, 1988) for students instructed with different curricula and across time, which may have biased analyses of the dimensionality of student responses to CAOS. Therefore, further research is required to ensure the homogeneity of the full CAOS sample and the appropriateness of utilizing the full dataset in any such analysis.

5.3. FUTURE ANALYSES

While the distinctness of subscores is an important attribute, the reliability of subscores must also be evaluated in order to fully establish validity evidence that meets current standards. Such analyses utilizing procedures prescribed by Haberman and Sinharay (2010) and Biancarosa et al. (2019) are an important next step to fully justify the utilization of CAOS subscores.

Although this investigation focused on the use of subscores, results from delMas (2014) identifying non-equivalent discrimination parameters in a 2PL model for dichotomous items, also replicated in the present study, indicate that the raw total score may not be a reliable indication of students' understanding. Similarly, Haberman and Sinharay (2010) note that raw subscores are less reliable than weighted subscores generated by MIRT models. Therefore, a thorough reexamination of the validity of total score interpretation using advanced IRT models in addition to the 2PL and GRM is required to fully understand the implications of analyzing raw total scores from student responses to CAOS in addition to raw subscores.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Allen, K. (2006). *The statistics concept inventory: Development and analysis of a cognitive assessment instrument in statistics*. (Doctoral dissertation, University of Oklahoma, 2006). Retrieved from SSRN.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and psychological measurement*, *79*(1), 65-84.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184-186.

Brown, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230-258.

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*(3), 318-338.

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1).

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of outcome measurement*, *3*(1), 1-20.

de Ayala, R.J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

delMas, R. (2014). Trends in students' conceptual understanding of statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute.

delMas, R., Garfield, J., & Chance, B. (2003). The web-based ARTIST: An online resource for the assessment of instructional outcomes. Paper presented at the *Joint Statistical Meetings,* San Francisco, August, 2003.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy (on CD).* Auckland, New Zealand: University of Auckland.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice*, *36*(1), 5-13.

Garfield, J., delMas, R., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project*. NSF CCLI grant ASA- 0206571.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883-898.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Haberman, S.J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling, 14*, 202–226.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41-54.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*, 391-406.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136-153.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59-81.

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment. *Journal of Applied Testing Technology*, *13*(2).

Lock, R. H., Lock, P.F., Lock Morgan, K., Lock, E.F., & Lock, D.F. (2013). *Statistics: Unlocking the power of data*. Hoboken, NJ: Wiley.

Lyren, P. (2009). Reporting subscores for college admissions tests. Practical Assessment, Research, and Evaluation, 14, 1-10.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333-356.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Meijer, R. R., Boevé, A. J., Tendeiro, J. N., Bosker, R. J., & Albers, C. J. (2017). The use of subscores in higher education: When is this useful?. *Frontiers in psychology*, *8*, 305.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289-337.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research*, *47*(5), 667-696.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Sabbag, A. G. (2016). *Examining the relationship between statistical literacy and statistical reasoning*. (Doctoral dissertation, University of Minnesota, 2016). Retrieved from the International Association for Statistical Education.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150-174.

Sinharay, S., Haberman, S.J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21-28.

Spearman, C. (1904). " General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*(5), 406-427.

Tintle, N.L., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2015). *Introduction to statistical investigations.* New York: Wiley.

Tintle, N.L., Rogers, A., Chance, B., Cobb, G., Rossman, A. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.

Tintle, N. L., Topliff, K., VanderStoep, J., Holmes, V. L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

Tintle, N. L., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1).

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods, 17*(2), 228-243.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Waller, N. G. (2018). Direct Schmid–Leiman Transformations and Rank-Deficient Loadings Matrices. *Psychometrika*, *83*(4), 858-870.