

Modeling Local Dependence Using Bifactor Models

Xin Xin

Summer Intern of American Institution of Certified Public Accountants

University of North Texas

Jonathan D. Rubright

American Institution of Certified Public Accountants

Abstract

Unidimensional item response theory assumes that the observed item response residual variances are perfectly uncorrelated. When several items form a testlet, the residuals of these items might be correlated due to the shared content, thus violating the local item independence assumption. Such violation could lead to bias in parameter recovery, underestimates of standard errors, and overestimates of reliability. To model any possible violation of local item independence, the present project utilized bifactor and correlated factors to distinguish between the general trait and any secondary factors. In bifactor models, each item loads on a primary factor that represents the general trait and one of the secondary factors that represents the testlet effect. Using the AUD, FAR, and REG sections' task-based simulation data, possible testlet effects are studied. It appears that the testlet format does not cause the violation of local item independence; however, when simulations come from the same content area, such shared content areas form extra dimensions over and above the general trait. In this study, operational proficiency scores are compared with bifactor general factor scores, factor loadings from bifactor and alternative models are compared, and the reliabilities of primary and secondary factor scores are provided.

Modeling Local Dependence Using Bifactor Models

Item response theory (IRT) has been widely used to analyze categorical data arising from the application of tests. Traditionally, unidimensional IRT models assume that there is only one latent trait, that the observed item response residual variances are uncorrelated, and that there is a monotonic non-decreasing relationship between the latent trait scores and the possibility of answering an item correctly (Embretson & Reise, 2000). However, in reality, these assumptions rarely perfectly hold. For example, if several questions follow a single simulation, then the residuals left after extracting the major latent trait might be correlated due to the shared content. Researchers usually call such subsets of items a “testlet” (Wainer, 1995; Wainer & Kiely, 1987), and such correlated residuals can result in a violation of the local item independence assumption (Embretson & Reise, 2000).

Literature Review

Research has repeatedly shown that the violation of local item independence leads to bias in item parameter estimates, underestimates of standard error of proficiency scores, and overestimation of reliability and information, which in turn leads to bias in proficiency estimates (Wainer, Bradlow, & Wang, 2007; Yen, 1993; Ip, 2000, 2001, 2010a). A handful of statistics have been developed to detect whether this violation appears in a dataset (Chen & Thissen, 1997; Rosenbaum, 1984; Yen, 1984, 1993). However, more recently there has been an increased focus on the extent to which, and even how, this assumption is violated. With new developments in the psychometric literature, the possible violation of local item independence can be captured by more advanced modeling techniques, revealing more detail on how residuals in a particular content are correlated. As DeMars (2012) summarized, the two main approaches to model local item independence are governed by how a violation is defined, either 1) when the covariances

among item responses are non-zero over and above the general trait, or, 2) when the probability of one item's observed response becomes a function of another item's observed response (Andrich & Kreiner, 2010). The former approach models local item dependence in terms of dimensionality, whereas the latter adopts regression on the observed responses. The present project adopts the multidimensionality approach and uses bifactor models to analyze a violation of the local item independence assumption.

The CPA Exam includes both multiple choice and testlet format items. All items in the same testlet share the same simulation stimulus, raising concerns over whether such a testlet format might introduce correlations among the residuals over and above the general trait. Also, some testlets belong to the same content coded area; such shared content might also introduce a violation of local item independence. In the literature, several models have been used to model such possible violations of local item independence, and they will next be discussed.

The Testlet Model. Multidimensional IRT models have been developed to unveil the complex reality in which the traditional IRT assumptions rarely flawlessly hold. As Ip (2010b) pointed out, item responses could be unidimensional while item responses display local dependency over and above the general trait. Moreover, he also proved that with Taylor expansion the observed response covariance matrix yielded by locally dependent IRT models was mathematically identical to that yielded by multidimensional IRT models. The present project adapts the multidimensional approach to model the violation of local item independence for task-based simulation (TBS) data. Studies have repeatedly shown that such testlet formats are susceptible to violation of local item independence (Sireci, Wainer, & Thissen, 1991; Wainer & Thissen, 1996).

Bradlow, Wainer, and Wang (1999) first proposed to model residual correlations with a testlet, dubbed the testlet effect, by adding a random testlet term to the item response function.

$$t_{ij} = \alpha_i(\theta_j - b_i + \gamma_{d(i)j}) + \varepsilon_{ij} \quad (1)$$

As shown in (1), t_{ij} is the latent score of candidate i on item j , ε_{ij} is a unit normal variate for error, θ_j stands for proficiency, α_i and b_i are the discrimination and difficulty of the i th item respectively. $\gamma_{d(i)j}$ stands for the person specific effect of testlet d that contains the i th item, and the variance of $\gamma_{d(i)j}$ is allowed to vary across different testlets as an indicator of local item dependence. Notice the general trait and testlet effect share the same discrimination α_i , which is a very strict assumption that item j discriminates the same way for both the general and testlet traits. Li, Bolt, and Wu (2006) modified this restriction to allow the testlet effect discrimination α_{i2} to be proportional of the general trait discrimination α_{i1} . Yet the modified testlet model is still so restrictive that the general and testlet discriminations are not separately estimated. In contrast to these testlet models, bifactor IRT models allow testlet effects to be freely estimated within the same testlet. As Rijman (2010) proved, testlet models are a restricted special case of bifactor models; as shown in his empirical data analysis, the testlet response model was too restrictive, while a bifactor structure modeled the reality better in that all loadings were freely estimated without any proportionality constrains. Because of these limitations of testlet-type models, they are not used in the present project.

The Bifactor Model. Gibbons and Hedeker (1992) developed bifactor IRT for dichotomous scores. Here, the probability of responses are conditioned on both general trait g and k group factors,

$$P(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^J P(y_{j(k)}|\theta_g, \theta_k) \quad (2)$$

where $y_{j(k)}$ denotes the binary response on the j th item, $j = 1, \dots, J$, within k factors, $k = 1, \dots, K$. \mathbf{y} denotes the vector of all responses. $\boldsymbol{\theta} = (\theta_g, \theta_{k1}, \dots, \theta_{kK})$, θ_g is the general trait, and θ_k is the testlet trait. Furthermore, through the logit link function,

$$g(\pi_j) = \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j \quad (3)$$

where β_j is the intercept for item j , and α_{jk} are loadings of item j on the general and specific traits, respectively. In the present analysis, the α matrix will be a J by $(K+1)$ matrix,

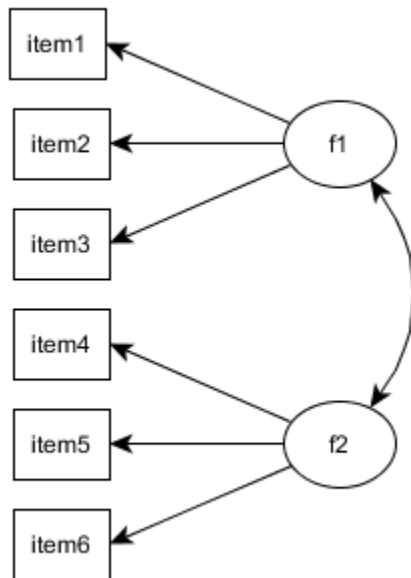
$$\alpha = \begin{bmatrix} \alpha_{11} & 0 & 0 & \alpha_{1g} \\ 0 & \alpha_{jk} & 0 & \alpha_{jg} \\ & & \dots & \\ 0 & 0 & \alpha_{jk} & 0 & \alpha_{Jg} \end{bmatrix} \quad (4)$$

Studies have shown the numerous advantages of bifactor models. For example, Chen, West and Sousa (2006) argued that bifactor models offered better interpretation than higher-order models. Reise and colleagues have published years of studies promulgating the application of bifactor models (Reise, Morizot, & Hays, 2007; Reise, Moore, & Haviland, 2010; Reise, Moore, & Maydeu-Olivares, 2011). As Reise (2012) summarized, bifactor models partition item variance into the general trait and the testlet effect; therefore, unidimensionality and multidimensionality became two extremes of a continuum and bifactor models illustrate the extent to which data are unidimensional. Bifactor models use secondary factors to represent any local item dependencies, provide model fit indices, variance-accounted-for type of effect sizes, and allow researchers to compare the primary factor loadings with secondary factor loadings and to compute reliabilities of the primary and secondary factor scores.

Alternative Models. Good model fit does not guarantee a true model. The fact that a bifactor model fits better than a competing model cannot guarantee a true bifactor structure. Therefore, alternative models are also tested. Both model fit indices and model comparisons are considered to determine whether a bifactor structure appears. Although the true model remains unknown, such evidence can help researchers made more educated decisions. Here, since a unidimensional 3PL IRT model is used operationally on the CPA Exam, bifactor models will be compared with unidimensional models for model selection and scoring.

Correlated models are also rival models because they are more commonly used than bifactor models in multidimensional studies. In a typical correlated factor model as shown in the following Figure 1, the variance of item 2 can be directly attributable to f1, yet can also indirectly contribute to f2 through the correlation between f1 and f2. Although the correlated factors models can provide the total variance explained by all factors, it cannot tell the exact contribution of which factor(s) explain the observed response variances.

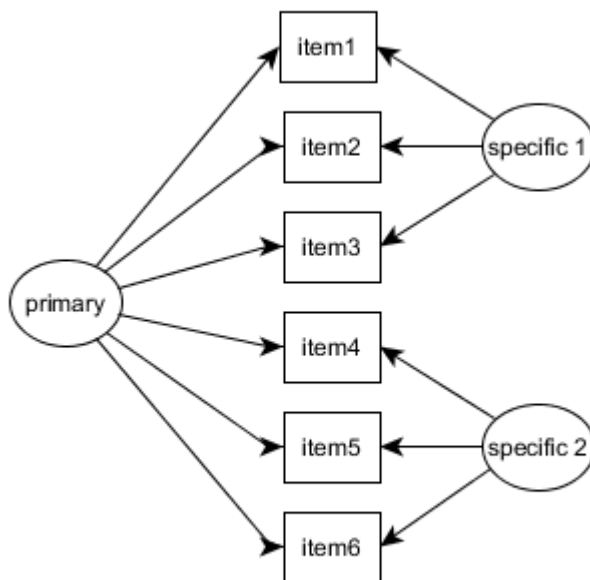
Figure 1 *Correlated Factors Model*



In contrast, bifactor models acknowledge the correlation between factors, yet are also able to isolate the proportion attributable to each factor. For example, looking at Figure 2 below, the primary factor explains the shared proportion of the two subdomains, and specific 1 explains the unique contribution to item 2 above and beyond the primary factor. Therefore, a bifactor model is able to tell researchers which factors explain the total variance of item 2.

As Reise (2012) pointed out, since primary and specific effects are confounded in a correlated factor models, model fit indices might indicate bad model fit, yet not be able to tell the source of any misspecification. By fitting unidimensional models, correlated factor models, and bifactor models, the model fit indices and effect sizes can collectively help us gain insights on what might be the true generating model.

Figure 2 *Bifactor Model*



Therefore, bifactor models are chosen to model the violation of local item independence, and both the unidimensional 3PL IRT model and the correlated factors model are also tested.

Method

Data Description

The present project used CPA Exam operational data from the third window of 2013. In the operational data, the several measurement opportunities (MOs) within each task-based simulation (TBS) were used to form each testlet. Three sections of the Exam, AUD, FAR, and REG, contain the TBS format. Besides a structure based on MOs nested within a TBS, a shared content area might also be a source of local item dependence. For example, if 2 TBSs share the same content area (but different group) codings in the item bank, all of the MOs within both TBSs may form a secondary factor. Therefore, we also model MOs based on their content area codings. The secondary factor structure based on TBS and based on content area codes are shown in Table 1.

Table 1 *Secondary Factor Structures*

AUD sample size = 745	FAR sample size = 807	REG sample size = 810
By TBS format		
TBS1517: MO1-MO5	TBS552: MO1-MO7	TBS631: MO1-MO8
TBS1699: MO6-MO13	TBS610: MO8-MO15	TBS567: MO9-MO15
TBS2748: MO14-MO19	TBS2702: MO16-MO24	TBS2384: MO16-MO21
TBS3339: MO20-MO24	TBS3564: MO25-MO32	TBS4911: MO23-MO29
TBS3834: MO25-MO34	TBS540: MO34-MO41	
By content area		
AREA4: MO1-MO5, MO14-MO19	AREA2: MO1-MO7, MO25-MO32, MO34-MO41	AREA6: MO1-MO8, MO9-MO15
AREA2: MO6-MO13, MO20-MO24	AREA3: MO8-MO15, MO16-MO24	AREA5: MO16-MO21 MO23-MO29
AREA3: MO25-MO34		

Note: When there was only one MO in a given TBS, such item only loaded onto the primary factor in bifactor models; such items had 0 loadings onto any factor in correlated factors model. Such items are: MO35 in AUD, MO33 in FAR, MO22 in REG.

Software Used

Mplus 7 (Muthén & Muthén, 2012), TESTFACT (Bock et al, 2003), and flexMIRT 2.0 (Cai, 2014) were all used, as they provide different model fit indices and R-square type of effect sizes. Given that many argue that model fit indices should be interpreted collectively and no firm cut-off points should be blindly followed (Kline, 2009), all three softwares are used to provide insight on the models by collectively interpreting different model fit indices and effect sizes. More details on the differences between the softwares will be examined shortly.

Methods Used

Confirm testlet effect. DeMars (2012) argued that although ignoring a testlet effect led to estimation bias, blindly modeling a testlet effect led to over-interpretation of results and a waste of time in looking for a fictional testlet effect. Reise (2012) also discussed the indicators of bifactor structure. Unfortunately, there is no clear cut-off point for any statistics or model comparisons to definitively confirm a bifactor structure. In the present study, both the confirmatory factor analysis (CFA) approach and the IRT approach were adopted.

CFA is a limited information method that focuses on a test level analysis, whereas IRT is a full information method that focuses more on an item level analysis. As a first step in looking for a testlet effect, a unidimensional confirmatory factor model, correlated factor model, and bifactor model are tested in Mplus 7 using weighted least square estimation with mean and variance correction (WLSMV). According to Brown (2006), the WLSMV estimator is superior to others within a factor analysis framework for categorical data. And at least until 2012, Mplus 7 is the only software available for WLSMV (Byrne, 2012). Because WLSMV does not maximize the loglikelihood, model fit indices like $-2LL(-2\log\text{-likelihood})$, and those based on the maximum loglikelihood value such as Akaike's Information Criteria (AIC; Akaike, 1987),

Bayesian Information Criteria (BIC; Schwarz, 1978), and the sample-size adjusted BIC (SSA-BIC; Sclove, 1987), are not available.

However, flexMIRT 2.0 can provide -2LL, AIC, and BIC. flexMIRT 2.0 is also flexible in that both Bock-Aikkin Expectation-Maximization (BAEM) algorithm and Metropolis-Hastings Robbins-Monro (MHRM) algorithm are available. Additionally, flexMIRT 2.0 can estimate a lower asymptote (pseudo-guessing) parameter in correlated factors models and bifactor models, which neither Mplus 7 nor TESTFACT can do. Therefore, flexMIRT 2.0 was used to fit the same models in order to see the loglikelihood-based model fit indices. Although Mplus 7 provides the proportion of observed variance explained for each MO, TESTFACT provides the proportion of total variance explained by each factor. Also, TESTFACT provides an empirical histogram of general factor scores. If the latent variable diverts from the expected normal bell curve, flexMIRT 2.0 can empirically estimate the latent distribution rather than assuming the normal distribution by default.

Mplus 7, TESTFACT and flexMIRT 2.0 will be used collectively to provide a coherent profile of the true latent structure. Notice it is always the researcher's decision whether the evidence is strong enough to concur with a bifactor structure. Also, considering that both the simulation format and shared content areas might cause violation of the local independence, both the simulation format and shared content areas will be modeled.

CFA and IRT. Besides the different model fit indices and effect sizes provided by Mplus 7, TESTFACT, and flexMIRT 2.0, another reason to use all three is due to the nature of the analysis. We use Mplus 7 to implement a CFA measurement model, while using TESTFACT and flexMIRT 2.0 for an item level analysis. There are abundant studies on the similarities and differences between factor analysis and IRT. McLeod, Swygert, and Thissen (2001) pointed out

that factor analysis was developed for continuous data, although with new developments it could analyze categorical data. These authors noted that the parameterization usually had problems, such as the tetrachoric correlation matrix were “often not positive definite” (p.197). In contrast, IRT was designed for categorical data, and therefore has less parameterization problems. Forero and Maydeu-Olivares (2009) also argued that although IRT and CFA belonged to the larger family of latent trait models, IRT was designed to handle non-linear relationships whereas factor analysis was mainly designed for linear relationships. Takane and de Leeuw (1987) summarized that dichotomization was straight forward in IRT, but the marginalization might be time-consuming; whereas marginalization became a trivial issue in CFA but dichotomization was difficult. For such different concentrations, many measurement invariance studies adopt both CFA and IRT together as a full treatment. In the present study, CFA conducted in Mplus 7 emphasizes the test level analysis, whereas item level analyses conducted in TESTFACT and flexMIRT 2.0 recover item level statistics. Model fit indices from all softwares fitting the same model will be summarized. This also means that, since Mplus 7 and TESTFACT are unable to estimate a lower asymptote parameter, the factor structure discussion utilizes the 2PL implementation of each model. Parameter recovery and scoring comparisons use 3PL implementations in flexMIRT 2.0.

Parameter Recovery. After the factor structure is fully analyzed as discussed above, flexMIRT 2.0 will be used to recover the parameter estimates and to compute the proficiency scores since it is the newest and most versatile IRT software. The factor loadings of unidimensional model and bifactor models will be compared. The operational proficiency scores utilizing a unidimensional 3PL IRT model and general factor scores from bifactor models will

also be compared. To help determine the merit of specific factor scores, reliability of both general and specific factor scores will also be provided.

Notice that although Mplus 7 can generate factor scores, bifactor models should not be scored in Mplus 7, because the factor scores will only stay uncorrelated as intended when factor determinacies equal 1 (Skrondal & Laake, 2001). TESTFACT has not been updated since 2004, and only calibrates the general factor scores for bifactor models. In contrast, flexMIRT 2.0 can produce both general and specific factor scores.

Results

Modeling Local Dependence Caused by TBS-Based Testlet Effect

Unfortunately, computational issues arose while estimating models based on the TBS structure of the Exam. In most cases, Mplus 7 had a hard time achieving convergence. Residual matrices were often non-positive definite. Although modification indices suggested eliminating a few problematic MOs, convergence results did not improve after hand-altering each model.

TESTFACT also ran into errors before convergence was achieved. Although in the long run flexMIRT 2.0 could satisfy the convergence criteria, the correlated factors models took hours to converge, while bifactor models converged within 10 minutes. In sum, none of the correlated factors models or bifactor models consistently converged across the three software packages for AUD, FAR, or REG.

These initial results led to a suspicion that the TBS format might not be the source of secondary factors. When convergence was achieved in Mplus 7 for the REG section, the bifactor model had better model fit indices than the correlated factors model. Yet for AUD and FAR using Mplus 7, the correlated factors models had less computational errors than the bifactor models. This indicated that multidimensionality existed, yet possibly not aligning with the TBS

format. Considering that some of the TBSs were coded to the same content area, it might be problematic to force all of the TBS-based secondary factors to be perfectly uncorrelated with one another. Therefore, modeling the secondary factors based on the content codes of the TBSs was also considered.

Modeling Local Dependence Caused by Shared Content Area

The Latent Structure and Model Selection. Once modeling the testlet effects focused on the content codings, all software implementations (Mplus 7, TESTFACT, and flexMIRT 2.0) encountered little computational problems. Using WLSMV in Mplus 7, the output provided the chi-square test of model fit, root mean square error of approximation (RMSEA), comparative fit index (CFI), chi-square test of model fit for the baseline model, and weighted root mean square Residual (WRMR). Chi-square tests of model fit can be compared by using DIFFTEST command in Mplus 7 for nested models, yet the correlated factors model is not nested within the bifactor model. Chi-square tests of the baseline models are almost always statistical significant regardless of the model, yet the p values across the different models cannot be compared.

Therefore, only the RMSEA and CFI are listed below in Table 2.

Table 2 *Mplus 7 Model Fit Results*

	unidimensional model	correlated factor model	bifactor model
AUD	RMSEA .058 CFI .769	RMSEA .048 CFI .854	RMSEA .036 CFI .916
FAR	RMSEA .091 CFI .789	RMSEA .081 CFI .845	RMSEA .063 CFI .906
REG	RMSEA .049 CFI .717	RMSEA .047 CFI .757	RMSEA .027 CFI .923

As shown in Table 2, for all three sections bifactor models had better model fit than the rival models, with the lowest RMSEA and the highest CFI, consistently above .90. Moreover, bifactor models can explain more observed variance than correlated factor models as shown in

Table 3. For example, in the REG section, the R-square output from Mplus 7 is shown as follows. On average, 16% of the observed MO variances can be explained by the unidimensional model, 19.8% can be explained by the correlated factors model, and 25.8% can be explained by the bifactor model.

Table 3 *Comparing Variance-Accounted-For*

REG	unidimensional model		correlated factor model		bifactor model	
	R-square	residual variance	R-square	residual variance	R-square	residual variance
MO1	0.292	0.708	0.315	0.685	0.318	0.682
MO2	0.332	0.668	0.341	0.659	0.455	0.545
MO3	0.133	0.867	0.135	0.865	0.158	0.842
MO4	0.234	0.766	0.256	0.744	0.277	0.723
MO5	0.261	0.739	0.276	0.724	0.348	0.652
MO6	0.142	0.858	0.147	0.853	0.162	0.838
MO7	0.088	0.912	0.089	0.911	0.091	0.909
MO8	0.192	0.808	0.209	0.791	0.219	0.781
MO9	0.424	0.576	0.479	0.521	0.925	0.075
MO10	0.402	0.598	0.423	0.577	0.444	0.556
MO11	0.393	0.607	0.448	0.552	0.688	0.312
MO12	0.048	0.952	0.046	0.954	0.052	0.948
MO13	0.08	0.92	0.081	0.919	0.081	0.919
MO14	0.02	0.98	0.025	0.975	0.04	0.96
MO15	0.153	0.847	0.18	0.82	0.351	0.649
MO16	0.106	0.894	0.201	0.799	0.715	0.285
MO17	0.13	0.87	0.229	0.771	0.633	0.367
MO18	0.141	0.859	0.202	0.798	0.217	0.783
MO19	0.062	0.938	0.073	0.927	0.068	0.932
MO20	0.016	0.984	0.028	0.972	0.027	0.973
MO21	0.111	0.889	0.143	0.857	0.136	0.864
MO22	0.053	0.947	NA	NA	0.058	0.942
MO23	0.12	0.88	0.168	0.832	0.146	0.854
MO24	0.113	0.887	0.169	0.831	0.142	0.858
MO25	0.052	0.948	0.086	0.914	0.06	0.94
MO26	0.054	0.946	0.093	0.907	0.068	0.932
MO27	0.123	0.877	0.182	0.818	0.148	0.852
MO28	0.092	0.908	0.107	0.893	0.116	0.884
MO29	0.283	0.717	0.424	0.576	0.33	0.67
average	0.160345	0.839655	0.198393	0.801607	0.25769	0.74231

Note. MO22 did not load on any factor in the correlated factor model, but it was part of the dataset. R-square is the observed MO variances explained by the models, and residual variances are the observed variances not explained by any factor in the models.

As shown above, the variance explained in correlated factors models can be attributed to the fact that MOs loaded onto the primary factor directly, and it can also contribute to the other factors that impact MOs through the correlations among factors. Although Mplus 7 does not distinguish between the variance explained by the primary or specific factors, bifactor models are able to provide such information. For example, TESTFACT provides the total variance explained by primary factor and specific factors as shown in Table 4.

Table 4 TESTFACT Variance-Accounted-For Results

	AUD	FAR	REG
TESTFACT	GENERAL 14.4820	GENERAL 30.3027	GENERAL 14.0363
(Bock et al,	AREA4 6.1079	AREA2 6.1137	AREA6 6.6787
2003)	AREA2 3.7196	AREA3 9.7724	AREA5 4.9569
	AREA3 4.0122	Residual 53.8112	Residual 74.3281
	Residual 71.6783		

Note: All analyses reached at least .001 changes were considered converged in TESTFACT (Bock et al, 2003).

In flexMIRT 2.0, unidimensional 2PL IRT models, 2PL correlated factors models, and 2PL bifactor models are compared by AIC, BIC, and SSA-BIC, where $AIC = -2LL + 2p$ (Akaike, 1987) and p is the number of parameter estimated, $BIC = -2LL + p(\ln(N))$ (Schwarz, 1978), and $SSA-BIC = -2LL + p(\ln((N+2)/24))$ (Sclove, 1987). Note that flexMIRT 2.0 provides -2LL, here -2LL were used to compute SSA-BIC; considering -2LL does not penalize on number of parameters estimated or simple size, -2LL would be more helpful to compare nested models. Therefore, -2LL was not compared in Table 4 in that the correlated factors models were not nested within bifactor models. AIC, BIC and SSA-BIC are shown in Table 5.

Table 5 flexMIRT 2.0 Model Fit Results

flexMIRT 2.0 (Cai, 2014)	unidimensional model	correlated factor model	bifactor model
AUD	AIC 30603.71	AIC 30295.06	AIC 29929.75
	BIC 30926.64	BIC 30627.23	BIC 30409.54
	SSA-BIC 30704.37	SSA-BIC 30398.60	SSA-BIC 30079.30
FAR	AIC 33963.78	AIC 33285.73	AIC 32092.98
	BIC 34348.63	BIC 33670.58	BIC 32665.56
	SSA-BIC 34088.24	SSA-BIC 33410.19	SSA-BIC 32278.14
REG	AIC 27136.30	AIC 27096.65	AIC 26766.36
	BIC 27408.73	BIC 27369.08	BIC 27170.30
	SSA-BIC 27224.54	SSA-BIC 27184.89	SSA-BIC 26897.20

Note: flexMIRT 2.0 provided AIC, and BIC. SSA-BIC was manually computed.

In sum, bifactor models had the lowest AIC, BIC and SSA-BIC. For example, in the FAR section, AICs of the unidimensional model, correlated factors model, and bifactor model are 33963, 33285, and 32093, respectively; BIC and SSA-BIC demonstrated the same trend. Therefore, bifactor models provided better fit than both unidimensional and correlated factor models. Only after we have much evidence that support the bifactor structure, we shall use bifactor models for parameter recovery and scoring.

Parameter Recovery and Scoring

As flexMIRT 2.0 is the newest, most flexible and most versatile software for IRT calibration and has encountered the least computational issues with these data, only flexMIRT 2.0 was used for parameter recovery and scoring. To ease the computation load, quadrature points were set to 21 between -5 to 5, maximum cycle was set to 5000, with convergence criterion set to .001, and the convergence criterion for iterative M-steps was set to .0001. Given that the operational proficiency scores were calibrated using a unidimensional 3PL IRT model in BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 2003), for this analysis flexMIRT 2.0 used the bifactor model with a lower asymptote to recover the parameter estimates and calibrate the general and specific scores. In both softwares, *Expected-a-posteriori* (EAP) scores were calculated. The prior for logit-guessing was set to a normal distribution with mean of -1.39 and

standard deviation of 0.5, which is equivalent to set guessing to a mean of .20 with three standard deviations below the mean to .05 and three standard deviation above the mean to .53.

The primary and content-area factor loadings of the AUD section are shown in Table 6. It shows that most MOs load more highly onto the primary trait, but a few MOs might be more content specific than the others. Looking specifically at MO16 and MO17, their content specific loadings are .80 and .85, respectively, whereas other MOs in the same content area have much smaller loadings. MO25 and MO28 demonstrate a similar pattern. MO20 to MO23 load closer onto the content area factor, but the MOs in the same testlet load similarly onto the primary trait. Although MO6 to MO13 are supposed to belong to the same content area as MO20 to MO24, the two simulations illustrate very different relationships on the content area factor. This may indicate that the groups under the same content area might not function similarly. Compared with the factor loadings from the unidimensional 3PL model, the bifactor model primary factor loadings are mostly close. However, only the bifactor model can distinguish the source of the loadings; for example, MO25 has a strong loading according to unidimensional model, yet in the bifactor model, the strong loading comes from the secondary factor instead of the primary factor. This indicates MO25 may be more content specific and less related to the general trait.

Table 6 *Factor Loadings from Unidimensional and Bifactor Model in the AUD section*

AUD	unidimensional 3PL model	Primary	Area4	Area2	Area3
MO1	0.08	0.05	-0.05		
MO2	0.42	0.47	-0.02		
MO3	0.14	0.19	-0.01		
MO4	0.10	0.14	-0.06		
MO5	0.18	0.20	-0.13		
MO6	0.38	0.37		0.12	
MO7	0.18	0.19		0.05	
MO8	0.29	0.29		0.11	
MO9	0.33	0.33		0.12	
MO10	0.28	0.25		0.15	

MO11	0.21	0.23	0.0	
MO12	0.23	0.19	0.17	
MO13	0.22	0.24	0.06	
MO14	0.73	0.98	0.0	
MO15	0.84	0.93	0.21	
MO16	0.89	0.58	0.81	
MO17	0.86	0.50	0.86	
MO18	0.77	0.77	0.14	
MO19	0.46	0.49	0.05	
MO20	0.58	0.46		0.65
MO21	0.67	0.45		0.83
MO22	0.77	0.46		0.68
MO23	0.67	0.39		0.85
MO24	0.51	0.42		0.40
MO25	0.86	0.18		0.88
MO26	0.39	0.39		0.13
MO27	0.41	0.61		0.78
MO28	0.26	0.26		0.37
MO29	0.39	0.45		-0.17
MO30	0.69	0.57		0.57
MO31	0.66	0.64		0.21
MO32	0.55	0.59		0.16
MO33	0.67	0.4		0.62
MO34	0.24	0.17		0.36
MO35	0.37	0.37		

Note: The latent variances were constrained to be 1, the loadings constrained to be 0 were left as blank in the table. Average standard error for unidimensional 3PL factor loadings is .14. Average standard error for bifactor model primary factor loadings .14; average standard error for content area specific factor loadings .14.

The factor loadings for the FAR section are shown in Table 7. Most MOs load more strongly onto the primary factor, yet MO1 through MO7 behaved differently from the other testlets. These MOs load more strongly onto the content specific factor. MO8 to MO15 also behaved different from MO16 to MO24. Notice that MO8 and MO9 are highly correlated, the standard error for the content area loadings are .75 and 1.64, which are much higher than the average standard error of .10 for the rest of content area loadings, and average standard error of .10 for primary factor loadings. Comparing unidimensional factor loadings and bifactor

primary factor loadings, MO1 through MO7 demonstrate strong loadings in the unidimensional model, but the bifactor model reveals that such strong loadings generally come from the secondary factor; they are more content specific than testing the primary trait.

Table 7 *Factor Loadings from Unidimensional and Bifactor Model in the FAR section*

FAR	Unidimensional 3PL model	Primary	Area2	Area3
MO1	0.80	0.54	0.60	
MO2	0.82	0.61	0.79	
MO3	0.87	0.62	0.60	
MO4	0.84	0.63	0.55	
MO5	0.94	0.51	0.86	
MO6	0.91	0.54	0.82	
MO7	0.97	0.58	0.78	
MO8	0.64	0.99		-0.13
MO9	0.67	0.99		-0.13
MO10	0.09	0.12		-0.03
MO11	0.41	0.45		0.0
MO12	0.53	0.54		0.01
MO13	0.47	0.55		-0.02
MO14	0.74	0.73		0.1
MO15	0.58	0.66		-0.01
MO16	0.57	0.44		0.53
MO17	0.62	0.48		0.75
MO18	0.64	0.4		0.86
MO19	0.67	0.45		0.73
MO20	0.59	0.37		0.75
MO21	0.63	0.47		0.74
MO22	0.58	0.46		0.49
MO23	0.61	0.45		0.69
MO24	0.52	0.39		0.38
MO25	0.54	0.62	0.03	
MO26	0.56	0.63	0.01	
MO27	0.65	0.71	0.06	
MO28	0.62	0.55	0.22	
MO29	1.00	0.98	0.2	
MO30	0.99	0.98	0.18	
MO31	0.61	0.55	0.23	
MO32	0.43	0.41	0.15	
MO33	0.55	0.54		

MO34	0.43	0.51	0.08
MO35	0.35	0.35	0.0
MO36	0.46	0.59	-0.08
MO37	0.70	0.61	0.22
MO38	0.39	0.47	-0.07
MO39	0.48	0.57	-0.07
MO40	0.41	0.48	-0.05
MO41	0.40	0.54	-0.12

Note: The latent variances were constrained to be 1, the loadings constrained to be 0 were left as blank in the table. Average standard error for unidimensional factor loadings is .10. Average standard error for bifactor model primary factor loadings .10; average standard error for content area specific factor loadings .13.

The factor loadings of the REG section are shown in Table 8. MO9, MO11, and MO15 belong to the same content area, yet they load much more strongly onto the area factor than the primary trait. In contrast, under the unidimensional model, MO9, MO11 and MO15 have moderate loadings and one cannot discern the content-specific structure. Meanwhile, the other MOs in the same content area load more strongly onto the primary trait. MO16 and MO17 illustrate similar behavior. MO20 has high standard error in both models, which indicates MO20 might need review.

Table 8 *Factor Loadings from Unidimensional and Bifactor Model in the REG section*

REG	Unidimensional 3PL model	Primary	Area6	Area5
MO1	0.86	0.79	0.16	
MO2	0.68	0.75	-0.12	
MO3	0.43	0.42	-0.01	
MO4	0.66	0.64	0.01	
MO5	0.68	0.77	-0.08	
MO6	0.45	0.45	0.04	
MO7	0.30	0.27	0.11	
MO8	0.63	0.67	0.0	
MO9	0.54	0.35	0.93	
MO10	0.69	0.61	0.44	
MO11	0.57	0.35	0.85	
MO12	0.27	0.23	0.16	
MO13	0.44	0.38	0.11	

MO14	0.44	0.21	0.29	
MO15	0.55	0.29	0.70	
MO16	0.27	0.16		0.88
MO17	0.33	0.23		0.83
MO18	0.58	0.51		0.65
MO19	0.59	0.58		0.27
MO20	0.05	0.57		0.71
MO21	0.46	0.45		0.15
MO22	0.26	0.25		
MO23	0.38	0.39		-0.03
MO24	0.37	0.37		0.13
MO25	0.29	0.3		-0.06
MO26	0.25	0.27		0.01
MO27	0.37	0.39		0.01
MO28	0.6	0.59		-0.14
MO29	0.59	0.6		0.0

Note: The latent variances were constrained to be 1, the loadings constrained to be 0 were left as blank in the table. In unidimensional model, MO14 and MO20 has standard error of .67 and 9.13 respectively, average standard error of the rest of factor loadings is .15. Average standard error of primary loadings is .16; MO16 through MO18 has standard error of .55, .53, and .67 respectively, average standard error of the rest of content area loadings is .16.

Scoring. The primary factor scores from flexMIRT 2.0 compared with the operational theta scores. The Pearson correlations between the operational EAP scores and bifactor model general factor EAP scores are shown in Table 9.

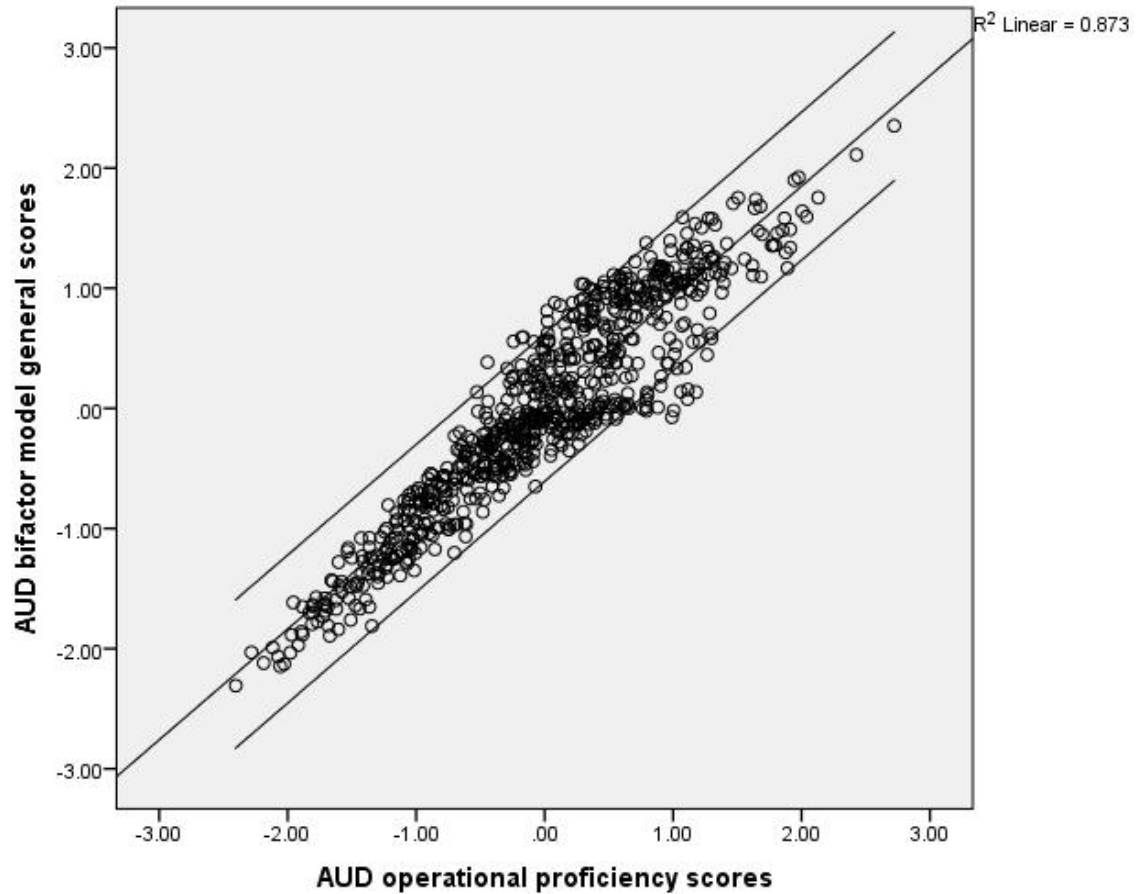
Table 9 *Pearson Correlations between the Operational Theta and the General Factor Score*

	AUD bifactor model general factor scores	FAR bifactor model general factor scores	REG bifactor model general factor scores
operational EAP proficiency scores	.935 **	.888 **	.968 **

Note: ** means correlation is statistically significant at .001 level (2-tailed).

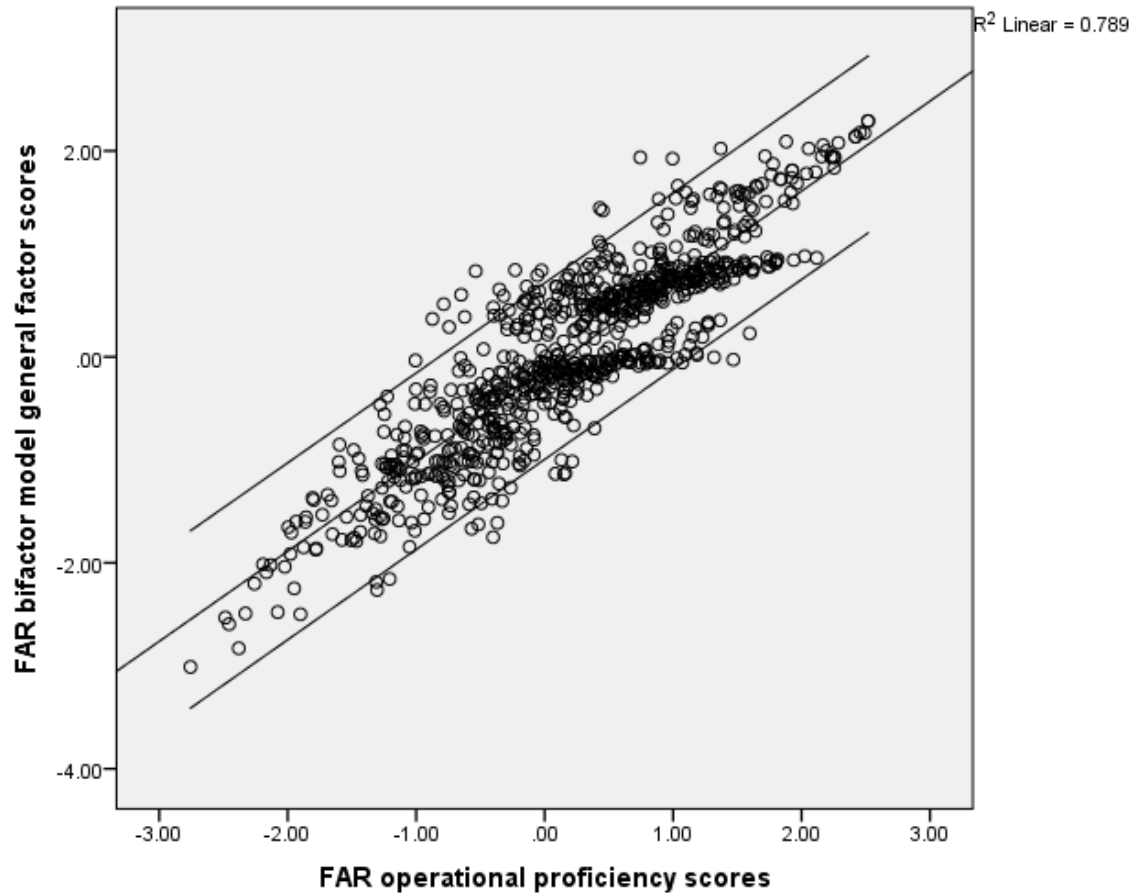
The scatterplots between the proficiency estimates from the unidimensional and the bifactor models for each exam section are shown in Figure 3 through Figure 5, with the regression line and 95% confidence intervals of individual scores.

Figure 3 *Scatterplot of the Operational Theta and the General Factor Score in AUD*



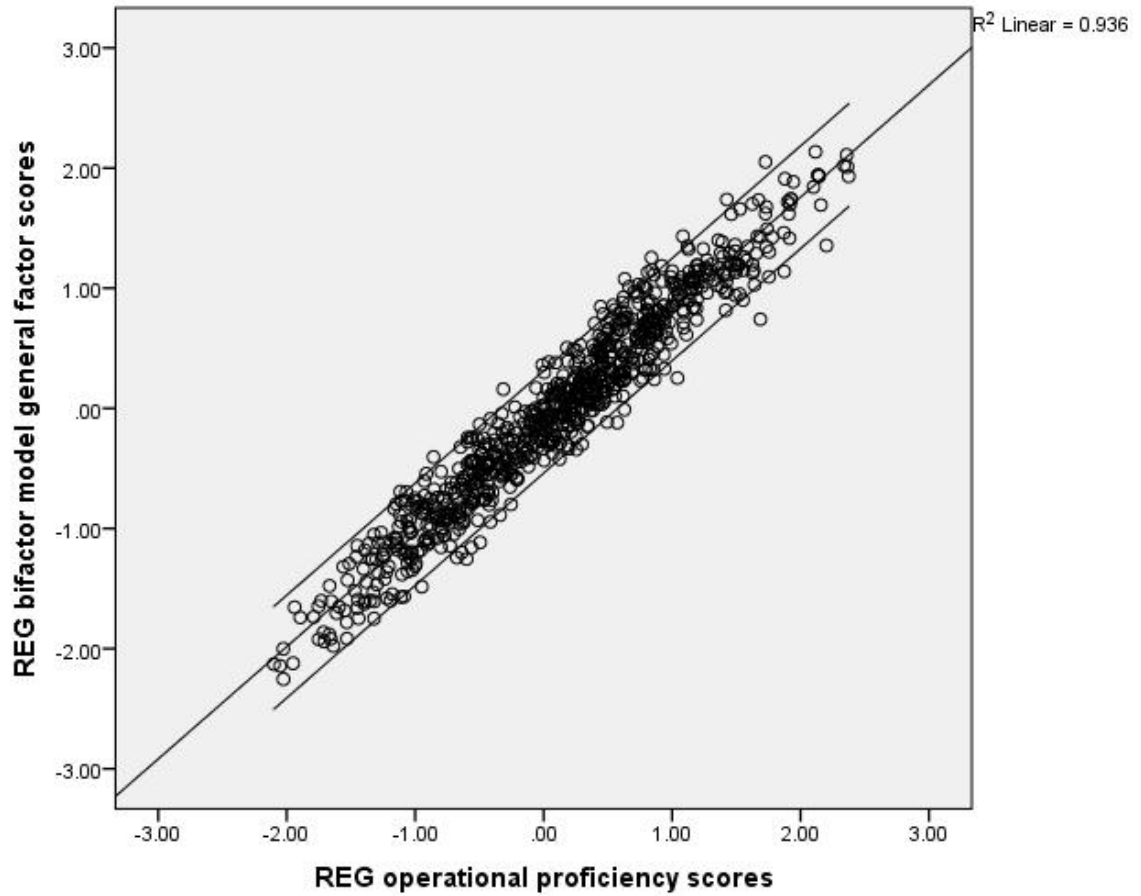
In the FAR section, some candidates at the same proficiency level of the bifactor model general factor score were considered to have different proficiency scores from the 3PL unidimensional model.

Figure 4 *Scatterplot of the Operational Theta and the General Factor Score in FAR*



In REG section, operational proficiency scores appear to match well with the general factor scores from bifactor model.

Figure 5 Scatterplot of the Operational Theta and the General Factor Score in REG



Reliability. Due to the number of MOs, specific factor scores are usually less reliable than the general factor scores; therefore, the reliability of general and specific scores should be considered before deciding on how to compute or use the subscale scores. Reliability was computed as $1 - \frac{S_e^2}{S_\theta^2}$ (Wainer, Bradlow, & Wang, 2007, p.76), where S_e^2 was the mean of squared standard errors and S_θ^2 was the true score variance (set to 1 in estimation). As shown in Table 10, in AUD section, content specific factor scores have much lower reliability than the primary factor scores.

Table 10 Reliabilities for the AUD section

AUD	SD of $\hat{\theta}$	Mean SE* of $\hat{\theta}$	Reliability
Primary factor	.8780	.4895	.7603
Area4	.7219	.6949	.5171

Area2	.6985	.7196	.4821
Area3	.6796	.7376	.4559

Note: * Mean SE was the square root of the mean squared standard errors

In the FAR section, although content secondary factor scores still have lower reliability than the primary factor scores, they are higher than those from AUD.

Table 11 *Reliabilities for the FAR section*

FAR	SD of $\hat{\theta}$	Mean SE* of $\hat{\theta}$	Reliability
Primary factor	0.9481	.3792	.8562
Area2	0.8432	.5330	.7159
Area3	0.8646	.5577	.6889

Note: * Mean SE was the square root of the mean squared standard errors

In the REG section, the reliability of the primary factor scores is lower than those from AUD and FAR; partly this is due to the fact that FAR has less MOs and is therefore a shorter test. Content specific factor scores have lower reliability, and any composite of both general and specific factor scores should be computed with caution.

Table 12 *Reliabilities for the REG section*

REG	SD of $\hat{\theta}$	Mean SE* of $\hat{\theta}$	Reliability
Primary factor	.8569	.5295	.7196
Area6	.8004	.6027	.6367
Area5	.7313	.6854	.5302

Note: * Mean SE was the square root of the mean squared standard errors

Discussion

The present project analyzed the operational data of TBS items from the AUD, FAR and REG sections of the Uniform CPA Exam. The purpose was to model possible violations of the local item independence assumption using bifactor models. Based on fit indices of unidimensional as compared to bifactor models, there is some evidence of the violation of the local item independence assumption. However, note that there are no established thresholds to

determine what constitutes a significant improvement in fit when comparing such models. Importantly, as evidenced by non-convergence of models specified based on the TBS structure, the bifactor structure was better modeled using content area codes as secondary factors. Future work on testlet structures utilizing the CPA Exam data may be informed by dimensionality assessments to guide where secondary factors should be placed.

Since secondary factor scores are generally less reliable than the primary factor scores, the operational proficiency scores were only compared with the general factor scores. Comparing operational, unidimensional ability scores to those from the bifactor primary trait, the REG section shows the closest alignment between the two, as the scatterplot reveals a tighter clustering around the regression line and the scores show a higher R^2 value than the AUD and FAR sections. One possible reason for these observed scoring differences is that the general and specific factor scores were confounded in the unidimensional model, while the bifactor model allows us to separate out the less reliable secondary factor scores.

Mplus 7, TESTFACT and flexMIRT 2.0 all made a unique contribution to the analysis. Overall, Mplus 7 and flexMIRT 2.0 were easier to use, more updated, and more flexible computationally. Mplus 7 is still easy to use when fitting alternative models and detecting the latent factor structure. flexMIRT 2.0 is computationally versatile and allows both BAEM algorithm and MHRM algorithm. When BAEM is time consuming under certain conditions, MHRM can be applied.

There are limitations to the present project. First, other tests are available to confirm the presence of secondary factors, which were not utilized here. For example, DeMars (2012) proposed to use TESTFACT to test the complete bifactor model against an incomplete bifactor model, where one of the secondary factors is omitted. Reise (2012) also suggested statistics using

squared general and secondary factor loadings to describe the relative strength of the factor loadings. Second, although using content area codes to guide the creation of secondary factors led to appropriate model convergence, it appears that TBSs did not behave the same way within the same content area. For example, in the FAR section, MO1 through MO7 behaved very differently from MO25 through MO32 and MO34 through MO41. This might indicate that groups of MOs within the same area function differently, and that some groups are closer to one another than others. Still, these questions are best addressed by a combination of content experts and statistical interpretation. Again, future work on bifactor structure may be informed by prior dimensionality assessment to guide the placement of the secondary factors. Additionally, in flexMIRT 2.0, when lower asymptote parameters were estimated, the same normal priors were used to keep the guessing parameters within a reasonable range. Different priors were not tested; in other words, a sensitivity test on priors is lacking.

Overall, modeling a simulation format or shared content area is much more complicated than using a unidimensional model, yet it can recover more information to understand what might be the true generating model. Bifactor models hold an advantage over rival multidimensional models of having more interpretable parameter estimates and flexible options of scoring. However, bifactor models are not immune to any of the widely discussed and studied but not yet settled issues in factor analysis or structural equation modeling, such as the validity of model fit indices and the lack of cut-off points for decision making between models. As operational tests become more structurally complicated, the application of unidimensional models needs to be justified. Bifactor models are capable of providing estimates to help researchers decide whether data are unidimensional enough (Reise, Morizot, & Hays, 2007).

Additionally, bifactor models can reveal more detail on how testlets or subdomains behave over and above the primary trait.

Future research steps may include testing testlet effects using complete bifactor structure and all-but-one bifactor structure as DeMars (2012) suggested, or adopting more statistics to help determine the true factor structure. These results might lead to working with content experts and modifying content codings so that the secondary factors could better model the true latent structure. Also, subscores can be computed using the primary and secondary factor scores, and the reliabilities could be considered to provide more detailed feedback to candidates. DeMars (2013) provided several ways to compute subscores, and each has different considerations. Bifactor models allow such scoring flexibility. In sum, more research is needed to determine the factor structure of CPA Exam data, and bifactor models may provide an opportunity to confirm this structure and provide information about subscale performance. Still, utilizing such models operationally carries practical challenges.

References

- Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181-192.
- Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). Testfact (Version 4.0) [Computer software and manual]. Lincolnwood, IL: Scientific Software International.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.
- Cai, L. (2014). flexMIRT™ version 2.0: A numerical engine for multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group.

- Chen, F. West, S., & Sousa, K. (2006). A comparison of bifactor and second-order models of quality-of-life. *Multivariate Behavioral Research, 41*, 189–225.
- Chen, W. -H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.
Retrieved from <http://www.jstor.org/stable/1165285>
- DeMars, C.E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*, 145–168.
- DeMars, C. E. (2012). Confirming Testlet Effects. *Applied Psychological Measurement, 36*, 104-121. doi:10.1177/0146621612437403
- DeMars, C. E.(2013) A tutorial on interpreting bifactor model scores, *International Journal of Testing, 13*, 354-378, doi: 10.1080/15305058.2013.799067
- Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.
- Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika, 65*, 73-91.
- Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*, 109-132.
- Ip, E. H. (2010a). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*, 467-482. doi:10.1177/0146621610364975
- Ip, E. H. (2010b). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-416.

- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21.
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement, 73*, 254-274, doi:10.1177/0013164412453841
- Muthén, L. K., & Muthén, B. O. (2012). Mplus version 7: Base Program and Combination Add-On (64 bit) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Reise, S. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696.
- Reise, S., 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*, 544-559, doi:10.1080/00223891.2010.496477.
- Reise, S., Moore, T. M. & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71*, 684-711, doi: 10.1177/0013164410378690
- Reise, S., Morizot, J. & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.
- Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361-372.
- Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

- Sireci, S. G., Thissen, D. & Wainer, H. (1991), On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28: 237–247. doi: 10.1111/j.1745-3984.1991.tb00356.x
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, 8, 157-186.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201. Retrieved from <http://www.jstor.org/stable/1434630>
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22–29.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, 30(3, Performance Assessment), 187-213. Retrieved from <http://www.jstor.org/stable/1435043>
- Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2003). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.