Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*

A Comparison of the Added Value of Subscores Across Two Subscore Augmentation Methods

Mohammad Afsharrad[1], Reza Pishghadam[2]*, Purya Baghaei[3]

Abstract
Testing organizations are faced with increasing demand to provide subscores in addition to the total test score. However, psychometricians argue that most subscores do not have added value to be worth reporting. To have added value, subscores need to meet a number of criteria: they should be reliable, distinctive, and distinct from each other and from the total score. In this study, the quality of subscores from two subscore augmentation models (Wainer and Yen) was compared in terms of distinctness and variability. The reliabilities of the Wainer-augmented subscores were also examined. The models were applied to a high-stakes English language proficiency test in Iran. The results of the study showed that Yen better-satisfied subscore distinctness while Wainer best-preserved variability and had high-reliability subscores. In other words, Yen-augmented subscores had lower correlations while Wainer-augmented subscores better discriminated examinees with different ability levels. Thus, none of the examined models of subscoring satisfied all criteria. The results of the study are discussed and suggestions for future research are provided.

Keywords: subscore augmentation; subscore distinctness; subscore variability; Wainer; Yen

1. Introduction

Educational and psychological tests often have subsections related to different content categories. Language proficiency tests are typically composed of different sections corresponding to language skills (listening, reading, speaking, and writing) and subskills (e.g. grammar and vocabulary). Separate scores for each of these modules are referred to as subscores.

Subscores are appealing to policymakers, admissions officers, teachers, and examinees, and thus testing programs face considerable demand for subscore reporting (Lim & Lee, 2020; Monaghan, 2006). Unsuccessful candidates can better plan for future remedial work by knowing their subscores, and consequently their strengths and weaknesses, on different sections of the test they have taken. In addition, teachers can draw on subscores to adjust their future instruction to address learners' weaknesses. Moreover, some universities and colleges require performance profiles of their graduates for better evaluation of their training as well as

[1] Department of English Language and Literature, Faculty of Letters and Humanities, Ferdowsi University of Mashhad, Mashhad, Iran; Email: mohammad.afshar@mail.um.ac.ir
[2] Department of English Language and Literature, Faculty of Letters and Humanities, Ferdowsi University of Mashhad, Mashhad, Iran; Email: pishghadam@um.ac.ir
[3] Department of English Language, Islamic Azad University, Mashhad Branch, Mashhad; Email: pbaghaei@mshdiau.ac.ir

remediation decisions (Haladyna & Kramer, 2004). Test users' demand for fine-grained section scores seems logical given that tests may be inherently multidimensional (Ackerman et al., 2003) and usually measure more than one ability simultaneously (Reckase, 2009; Yao, 2010)

Despite different stakeholders' desire to have subscores (Brennan, 2012; Haberman, 2008) and official requirement for providing diagnostic information about examinees' performance (Pellegrino et al., 2001, United States Congress, 2002), developers of assessments need scientific justifications for reporting separate scores for each section of the test. Diagnostic information can be provided in the form of subscores (Goodman & Hambleton, 2004) and some testing programs are already providing subscores (the College Board, 2017; ACT, 2016). Standard 1.14 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 27) reads that "when interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided." Moreover, it states that "when a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated, and the interrelationship of those scores should be shown to be consistent with the construct(s) being measured" (p. 27). However, many subscores provided by educational tests fail to meet the reliability and distinctiveness criteria (see Haberman, 2008; Puhan et al., 2010; Sinharay et al., 2007).

Reporting subscores is useful when they provide information over and above the total score. Wainer et al. (2001) and Tate (2004) believe that not only the total score but also the subscores measuring specific content must be reliable. Sinharay (2010) also highlights the requirements of added value and psychometric soundness for subscores to be reported (20 items per section, low correlation between subscores, and highly reliable subscores). Moreover, subtests deserve separate scores when they are distinct from each other (Sinharay et al., 2007) and when tests "support measurement of differential performance on the subtests" (Lee et al., 2017, p. 2). Hence, test developers have to establish an equilibrium between the qualifications needed for subscores to warrant reporting and score users' desire to have detailed information about examinee performance. This is where subscore augmentation comes in to help.

2. Review of Literature and Theoretical framework

2.1. Wainer's augmentation (Wainer et al., 2000 and 2001)

Wainer et al. (2001) used raw scores from both Classical Test Theory (CTT) and Item Response Theory (IRT) in their model. In their augmentation procedure, they drew on the regression model proposed by Kelley (1947) for estimating true scores. The difference between the two is that in Kelley's model scores are regressed on the mean score, while in Wainer's (2001) method the observed scores of all subtests are used to predict the true subscore (called augmented subscore) of interest. In other words, strength is borrowed from the more reliable subscores to improve the measurement accuracy of the less reliable ones. Wainer et al. (2001) drew on empirical Bayes theory and employed mathematical procedures to use information from other subtests in order to stabilize subscores.

In this model, when tests comprise subtests, but are unidimensional in nature, the insufficiently reliable subscores are replaced with the more reliable total score. According to

Wainer, Sheehan, and Wang (2000), if reliabilities of subscores and the total score are equal, then the test is unidimensional and subscores do not have enough additional information to be worth reporting.

On the other hand, in multidimensional tests, strength is borrowed from other subtests. In other words, the less reliable subscores are regressed on, or shrunk towards, the more reliable ones to become more stable. The size of shrinkage depends on the reliabilities of the subscales and the collateral information, and correlations between the subtests. The less reliable a subscore is, the more it shrinks and the higher the correlations are, the more precise the empirical Bayes (regressed) estimates will become.

Wainer et al. (2001) applied this score augmentation procedure to both CTT observed raw scores and IRT scale scores. However, in the process of augmenting IRT subscale scores some features of CTT are applied. For example, error of measurement is considered to be constant (the way it is in CTT) while in IRT standard errors vary from one value to another. In this study, Wainer's method is applied to CTT-based subscores.

2.2. Objective performance index (Yen, 1987)

In her Objective Performance Index approach (OPI), Yen (1987), employing Bayesian and unidimensional IRT procedures, estimated subscale scores (called objectives) by drawing on prior information, which could be examinees' performance in school or score from another test, but mostly was the rest of the test which included the subscale. The score resulting from this procedure is labeled as OPI. For subscores to have added value, each section of a test comprised of subsections must measure a different trait or a different aspect of an attribute. Since OPI is based on UIRT, it might not properly describe multidimensional data, and consequently not provide accurate measurement (Sinharay, Haberman, & Puhan, 2007).

Yen's (1987) OPI is different from Wainer's method in two ways. First, the two methods are computationally different. While Yen used binomial distributions, Wainer et al. used normal distribution theory to estimate subscores. Moreover, for estimating each subscore Yen (1987) considered the rest of the items, except those in the subtest of interest, as one unit while Wainer et al. (2001) recognized other subtests as separate units. Based on Skorupski and Carvajal (2010, p. 370) "the regression approaches [Wainer's model] increase reliability by making every examinee's score profile look more like the overall group's score profile. The BIRT method [Yen's model] increases reliability by making every examinee's subscore look more like his or her total score."

Subscore augmentation has been widely used to examine the added value of subscores and to improve their reliability (e.g. Choi & Papageorgiou, 2020; Papageorgiou & Choi, 2018; Sawaki & Sinharay, 2013, 2018; Skorupski & Carvajal, 2010). In most cases, subscores were found not to have added value. For example, Sinharay (2010) evaluated the quality of subscores from 25 operational tests. The number of subscores in the tests ranged from two to seven and there were 92 subscores altogether, out of which eventually only 16 subscores had added value.

For example, Skorupski and Carvajal (2010) compared Wainer et al. and Yen models to augment subscores. Two regression techniques were used in the study: one with CTT-based raw scores and the other with IRT-based raw scores. They examined data from a test with four subsections. The reliability values of the original raw scores were not in the acceptable range,

which rendered reporting subscores inappropriate. The Wainer augmented scores from CTT- and IRT-based regressions were similar to the baseline raw scores while the reliability of the subscores was improved. Estimates from Yen's model, however, changed scores as compared to baseline raw scores, though it improved reliability in a similar fashion to the other two models. All the three models made the values of the subscores and standard deviations similar, with Yen's estimates making them identical. Skorupski and Carvajal (2010) argue that though the reliability of subscores are increased, they become similar to each other (especially via Yen's approach) or similar to the group's pattern (Wainer et al.'s approach) and hence, augmented scores might not be appropriate for the purpose for which they went through the process of augmentation, providing additional information for diagnostic purposes. That is, though one of the conditions for a valuable subscore is satisfied (reliability), the very process of improving reliability undermines distinctness criteria; it makes subscores similar and such subscores are not distinctive. Skorupski and Carvajal (2010) did not use MIRT for subscore estimation on the grounds that Luecht (2003) found that this procedure is not as good as the other ones for subscore improvement.

2.3. The current study

Since the most important problem with subscores is low reliability, the primary concern of augmentation procedures has been improving precision and reliability of subscores. Hence, applying subscore augmentation methods usually improves subscore reliability and eliminates the low reliability problem of subscores. However, the other conditions for added value of subscores, low correlation between subscores and the total score, might not be satisfied and even deteriorated by such augmentation procedures. When the correlation between the total score and a subscore is high, there is not much information left to be given by the subscore which is not already provided by the total score. This makes reporting subscores useless.

Many studies have examined subscore accuracy across different subscoring methods (Erdemir & Atar, 2020; de la Torre et al., 2011; de la Torre & Patz, 2005; Wang et al., 2004; Yao & Boughton, 2007). However, such studies are limited in the area of foreign language testing (e.g., Longabach & Peyton, 2018). To the best of our knowledge, the subscore quality of none of the high-stakes language proficiency tests in Iran has been investigated. Moreover, unlike some studies (de la Torre et al., 2011; de la Torre & Patz, 2005; Edwards & Vevea, 2006; Yao & Boughton, 2007) which used simulated data, in this study real data are used. As mentioned by Langabach and Peyton (2018), irregularities inherent in real data are not available in simulated data. In studies with simulated data, generated data meet the assumptions of the model, which does not often happen with real data (de la Torre & Song, 2009). The majority of subscore correlation studies in the literature have been conducted on simulated data to figure out how subscore correlations (along with other criteria such as sample size, test length, etc.) affect the quality of subscores (e.g. Lee et al, 2017).

In this study, the psychometric properties of the subscores from a nationwide English language proficiency test, the National University Entrance Exam (NUEE, known as Konkoor in Iran), are examined. More precisely, the subscores given by two subscore augmentation models (namely Wainer, and Yen) are examined in terms of subscore distinctness and subscore variability. Subscore variability (distinctiveness) refers to how each person’s subscores are

different from each other (within-person variability) and how different examinees' scores on the same test section (between-person variability) are different. This research attempts to answer the following questions:

1. Does Wainer-augmentation improve the reliability of the subscores to an acceptable level?
2. How distinct are the Yen-augmented and Wainer-augmented subscores and how does distinctness change as a function of subscoring method? In other words, to what extent do the subscores from these models correlate with each other and how do these correlations compare across different modeling strategies?
3. In which subscore augmentation model are the subscores more distinct from the total test score?
4. In which subscore augmentation model do the subscores have more within-person and between-person variability?

3. Method

3.1. Participants and materials

The measure used in this study, NUEE, was administered to the candidates of English major in 2011. NUEE is a multiple choice four-choice test and for each item, examinees are required to select the correct option. This test is held once a year for different majors including math and physics, science, humanities, art, and foreign languages. For the purpose of this study, only the data obtained from the candidates for foreign languages were considered. This test includes six sections. Subscale length and descriptive statistics for the test are given in Table 1.

Table 1

Descriptive Statistics for the Raw Subscores and the Total Scores of NUEE

Section	Items	N	Mean	Std. Deviation	Std. Error	Reliability
Grammar	10	3175	3.41	1.91	0.03	0.47
Vocabulary	15	3175	3.69	2.34	0.04	0.61
Structure	5	3175	1.53	1.30	0.02	0.49
Functions	10	3175	4.92	1.94	0.03	0.54
Cloze	10	3175	5.01	2.23	0.04	0.64
Reading	20	3175	5.51	4.10	0.07	0.82
Total	70	3175	24.07	10.01	0.18	0.88

The participants of the study included those high school students who wanted to enter university and continue their studies in English teaching, English translation, or English literature majors. They were both males and females and their age range was 18-20.

3.2. Procedure

The statistical analyses of the study were conducted using R package 'subscore'. First, the Wainer-augmented and Yen-augmented subscores were computed. Then, to examine the preciseness of the Wainer-augmented subscores, the reliability of the subscores obtained from the Wainer model was computed. Then, to examine subscore distinctness, correlations between

the subscores as well as the correlation between each subscore and the total score were calculated. Finally, the variability of the subscores was examined. The procedure related to distinctness and variability was carried out for the two augmentation methods and the results were compared.

4. Results

4.1. Subscore precision

The reliability of the raw Wainer-augmented subscores are presented in Table 2. The results revealed that Wainer's augmentation improved the reliability of the subscores and the reliability of all subscores were acceptable.

Table 2

Reliability of the Raw and the Wainer Subscores

Section	Raw	Wainer
Grammar	0.47	0.83
Vocabulary	0.61	0.82
Structure	0.49	0.76
Functions	0.54	0.78
Cloze	0.64	0.84
Reading	0.82	0.87

4.2. Distinctness

To examine how distinct the subscores are from each other and from the total score, correlations between the subscores as well as the correlation between each subscore and the total score were computed. The results are reported in the following sections.

4.2.1. Distinctness of subscores from each other

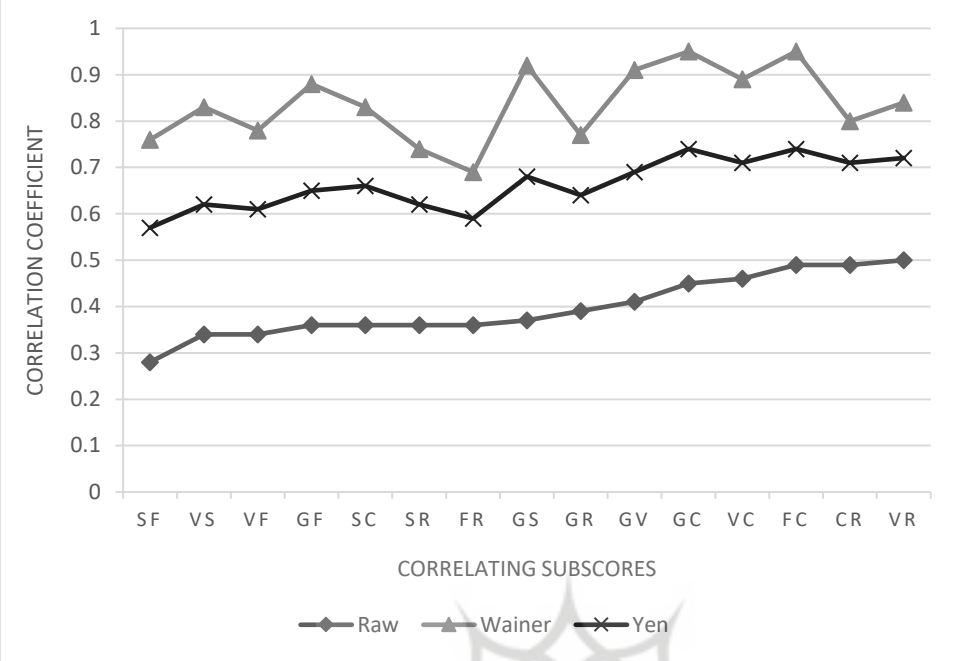
There were 15 possible correlations between the subscores of NUEE. Correlations between the subscores under different subscoring methods are provided in Table A1 in the appendix.

First, the influence of two subscore estimation models on subscore distinctness was examined. Second, the order of correlations in the augmented-subscores was compared against the original order in raw subscores to see which method better maintained the correlation observed in baseline raw scores.

The pattern of correlations among the subscores from the two models, demonstrated in Figure 1, were similar to each other but different from the pattern observed in the raw subscores. Nevertheless, Yen model was partly better than Wainer in preserving the correlations found in the original raw subscores. That is, the order of correlations in Yen model was the closest to that of the raw subscores (in both the raw and Yen subscores the ranks of four correlations were the same, six were different, and five were reversed). Wainer augmentation, however, changed the original order of correlations observed in the raw subscores more.

Figure 1

Correlations Between Subscores of NUEE Across Four Estimation Methods



Note: S= Structure; V= Vocabulary; F= Functions; G= Grammar; C= Cloze; R= Reading;

The highest and lowest correlating subtests and their coefficients across all methods are presented in Table 3. Structure-Functions had the lowest correlation in both the raw and Yen subscores, while in Wainer model the lowest correlation was different from that of the raw subscores. In this model, Functions-Reading had the lowest correlation. In none of the augmentation methods, however, the highest correlation was between the same subtests observed in the baseline raw subscores. The highest correlation was between similar subscores of NUEE (Grammar-Cloze and Functions-Cloze) in Wainer and Yen models while in the baseline raw subscores Vocabulary-Reading had the highest correlation.

Table 3

Highest and Lowest Correlations Between Subscores of NUEE

Correlation	method	NUEE
lowest	Raw	SF (.28)
	Wainer	FR (.69)
	Yen	SF (.57)
highest	Raw	VR (.50)
	Wainer	GC FC (.95)
	Yen	GC FC (.74)

Note. S= Structure; F= Functions; R= Reading; G= Grammar; C= Cloze; S1= Structure1; V= Vocabulary

In the next step, correlations between subscores were compared with a criterion correlation. This criterion value, against which correlations were compared, comes from the

simulation study of Lee et al. (2017), in which they determined optimal psychometric characteristics of subscores which are worth reporting. In their study, they found that for a test with five sections and 10 items within each section (the closest condition to that of NUEE), subscores are worth reporting when the correlation between them is .38 or below, the subscore reliability is .61 or above, and the total score reliability is .84 or above.

Half of the raw subscore correlations (GS, GF, VS, VF, SF, SC, SR, FR) and the total raw score reliability met the eligibility criteria. In addition, three subtests (Vocabulary, Reading, and Cloze) had acceptable reliability of .61 or above. Although the reliability of all Wainer-augmented subscores were acceptable (beyond the .61 criterion), the correlation was not satisfactory for any of the Wainer-augmented subscores. The minimum correlation from the augmented subscores was .57 (Structure-Functions) in Yen model. This lowest correlation is much higher than the criterion value (.38), though.

In order to have one correlation value for each estimation model to compare with the criterion correlation, the mean of the correlations between subtests under each estimation method was computed. The mean correlation of the raw, Wainer augmented, and Yen-augmented subscores were .40, .84, and .66, respectively. Yen model gave better results than Wainer model. Though the correlations of the Yen-augmented subscores were more acceptable than those of Wainer, they were still above the criterion value and, hence, not satisfactory. In short, considering the distinctness of subscores from each other, the Yen model performs better than Wainer.

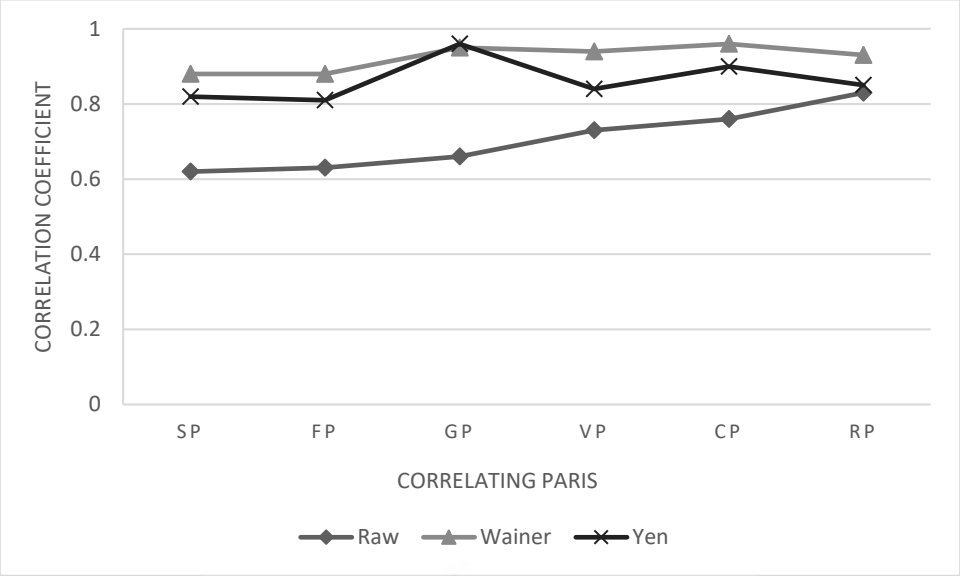
4.2.2. Distinctness of subscores from the total score

To examine how distinct the subscores are from the total score and how each augmentation method changes subscale-total distinctness, correlations between each subscore and the total score of Yen and Wainer models were computed (see Table A2 in the appendix for correlations).

As shown in Figure 2, in both Wainer and Yen models, the correlations were higher than those of the baseline raw subscores. The correlation between each subscore and the total score in the Wainer model was higher than its counterpart correlation in the Yen model. That is, the Yen model yielded lower correlation values and the subscores were more distinct from the total score in this model. The mean subscore-total correlation values for Wainer and Yen models are .92 and .86. However, Yen was worse than Wainer in maintaining the original arrangement of the subscore-total correlations observed in the raw subscores.

Figure 2

Subscore-Total Correlations for NUEE Across the Augmentation Models



Note: S= Structure; P= Proficiency; F= Functions; G= Grammar; V= Vocabulary; C= Cloze; R= Reading

Structure had the lowest correlation with the total score in the Wainer model (.88). In the raw subscores, Structure had the lowest correlation with the total score, too (see Table 4). Functions also had the lowest correlation with proficiency in the Wainer model. In the Yen model, Functions had the lowest correlation (.81) with proficiency.

Table 4

Highest and Lowest Subscore-Total Correlations Across Different Methods

Correlation	Method	Correlations
lowest	Raw	SP
	Haberman	SP
	Wainer	SP FP
	Yen	FP
highest	Raw	RP
	Haberman	GP
	Wainer	CP
	Yen	GP

Note. P= proficiency; S= Structure; F= Functions; R= Reading; G= Grammar; C= Cloze; S1= Structure1; S2 Structure2; V= Vocabulary

4.3. Subscore variability

To examine how each augmentation method influences subscore variability (both within-person and between-person) of examinees at different ability levels, subscores of 6 examinees (indicated by S1 to S6 in Figures 3 to 6) were analyzed. The examinees included the highest-scoring examinee, the lowest-scoring one, one randomly chosen examinee whose

total score was very close to the mean of the sample, and three examinees selected randomly. Each examinee's subscore variability (within-person variability), as compared to the variability in their raw subscores, was reduced in both Wainer and Yen models (see Figures 3 to 5). In other words, augmenting procedures made scores of each examinee on different subscales close to each other. Yen model reduced within-person variability the most.

The variability of each subscore across examinees (between-person variability) was also decreased as a result of augmentation. Wainer had a less noticeable effect than Yen did. In other words, the discrimination power of the test was reduced by applying subscore augmentation methods. In the case of Wainer, between-person variability was least influenced when the variance of scores was high and differences between examinees were big (e.g. Reading subscore).

Figure 3

Within-Person and Between-Person Variability for NUEE Raw Subscores

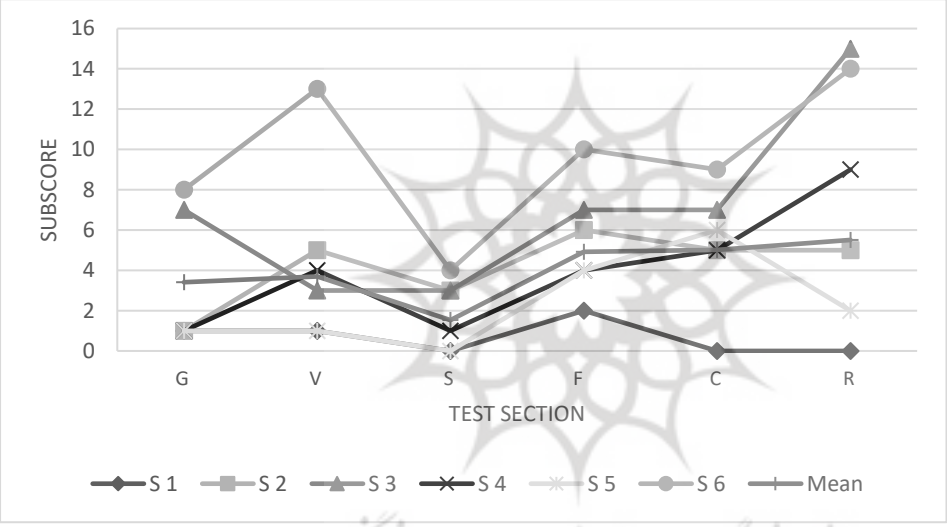


Figure 4

Within-Person and Between-Person Variability for NUEE Wainer Subscores

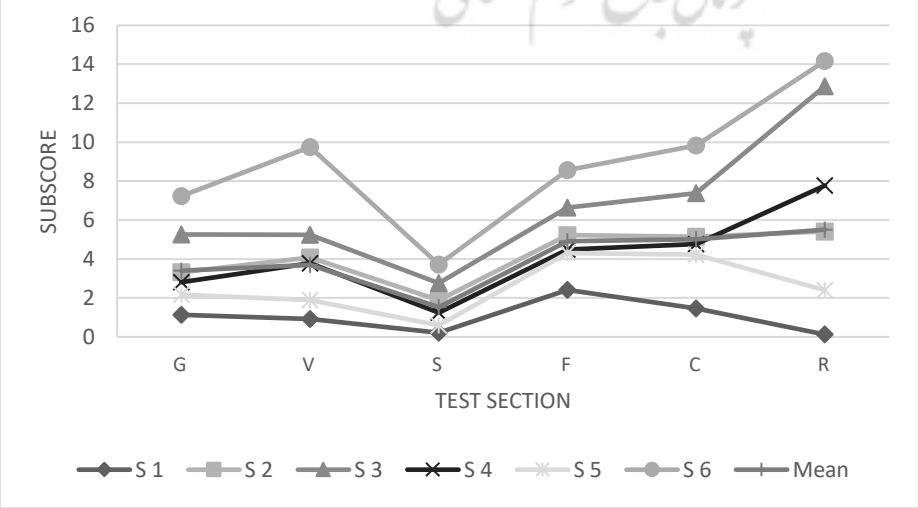
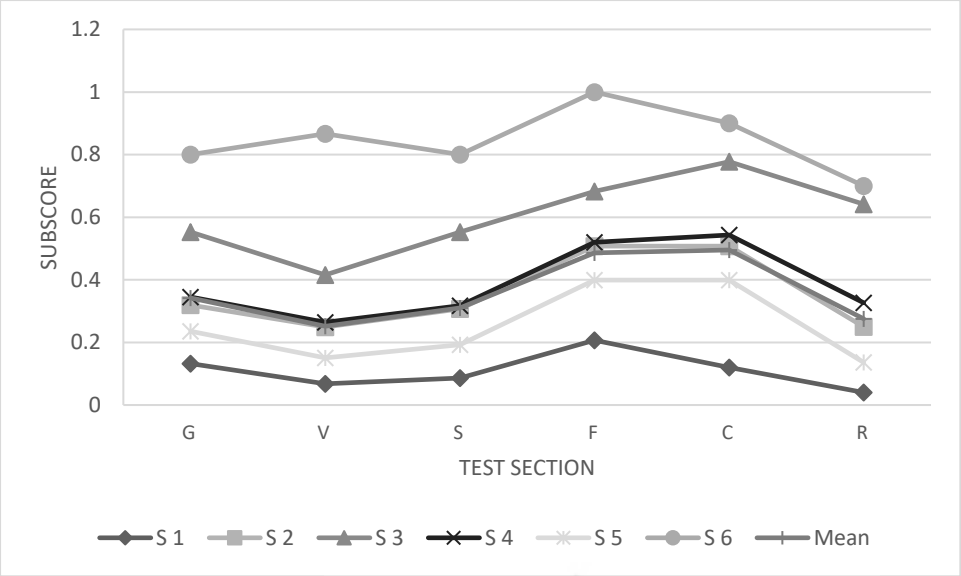


Figure 5

Within-Person and Between-Person Variability for NUEE Yen Subscores



In the raw subscores, between-person variability was different from one subscore to another. For example, Grammar subscores were not as varied as Reading subscores. This feature of variability was preserved by the Wainer model to some extent. In the Yen model, however, this change in variability became slight. That is, in the Yen model discrimination between persons was very similar across all subscores.

5. Discussion

The results of the study revealed that Wainer, and Yen augmentation models behave similarly in terms of the order of correlations among subscores. It means that if all possible correlations between subscores in each method are computed and arranged in a(n) ascending/descending order, the pattern of correlations is comparable across these models. This pattern, however, is different from the order of correlations in the baseline raw subscores. The order of correlations obtained from the Yen model is closest to that of the raw subscores and that from the Wainer model has the biggest difference. Hence, the Yen model best preserves the order of correlations existing among the raw subscores. The same pair of subscores (Structure-Functions) had the lowest correlation in both the raw and Yen subscores. The lowest correlation in the Wainer model, however, was between Functions and Reading, which is different from those of the raw and Yen models. The highest correlation for NUEE subscores was between the same subscores (Grammar-Cloze) in both Wainer and Yen models. In the raw subscores, however, the highest correlation was between Vocabulary and Reading. Considering the mean correlation of the subscores, the Yen model gives the best-quality subscores (lowest correlation). Although Yen gives lower correlations than Wainer, the correlations in this model are too high for the subscores to have added value. In sum, as indicated by Figure 1, the Yen model acts more like the raw subscores than does the Wainer model. The similarity lies in both the pattern of correlations (their order) and correlation values. In other words, the Yen model better preserves the nature of the raw subscore correlations.

Considering Lee et al.'s (2017) reliability criterion, three out of the six raw subscores of NUEE were acceptable and with regard to the correlation criterion, eight out of the 15 correlations among the raw subscores had added value. Wainer augmentation improved the reliability of all subscores to the acceptable level. This reliability improvement is the purpose for which augmentation is implemented and the goal is achieved. Reliability improvement as the result of augmentation, however, the enhanced correlation between the subscores so drastically that none of them were acceptable. With regard to the acceptability of correlations (based on Lee et al.'s criterion), more favorable results (lower correlations among subscores) are obtained from the Yen model.

Although Yen was better than Wainer in preserving the nature of subscore relationships (the order of correlations among the subscores) and had the lowest correlations, it was the worst in maintaining the order of subscore-total correlations. The order of subscore-total correlations in the Wainer model was more similar to that of the raw subscores. Yen model yielded lower (which means better in this context) subscore-total correlations, and Wainer gave higher correlations. The lowest subscore-total correlation in NUEE was between Structure and Proficiency in the raw and Wainer models while in the Yen model it was between Functions and Proficiency.

All subscore augmentation models reduced both within-person and between-person variability. That is, the augmentation process made the scores of each individual on different subtests (within-person variability) similar to each other. It also pulled the scores of different examinees on the same subscale (between-person variability) closer to each other. This is expected since augmentation borrows strength from other parts of the test to improve subscore reliability. However, Wainer better maintains both within-person and between-person variability observed in the baseline raw. Unlike Wainer, Yen model fundamentally changes variability at both levels. This is in line with the findings of Skorupski and Carvajal (2010), who found that subscores become similar to each other in the Yen method but similar to the group's pattern in the Wainer model. Therefore, although Yen is the best model in preserving the original order of subscore correlations and keeping them low, it is the worst in maintaining subscore variability. Since differences between examinees on the same subtests as well as differences between the abilities of an examinee on different subtests are important, augmentation, especially Yen augmentation, is unfavorable. This model reduces the discrimination power of the test and makes examinees' strengths and weaknesses less clear. Thus, with the criteria of subscore reliability, distinctness of subscores from each other and from the total score, and subscore variability, there is no one best method of subscoring.

In sum, for NUEE to have subscores with higher added value, the number of items in different sections of the test should be increased and interrelationships between them decreased. Based on Sinharay (2010, p.168) "the subscores have to consist of several items (at least 20) and be sufficiently distinct from each other (with disattenuated correlations less than .85) to have any hope of having added value".

6. Limitations and future research

The first limitation of this study is that it did not have the reliabilities of Yen subscores so that reliabilities could be compared. This limitation results from the unavailability of

appropriate computer software, which gives comprehensive output related to this model. R package 'subscore', which was used in this study, gives detailed output for the Haberman model and sufficient information related to Wainer while for the Yen model it provides researchers with only OPI subscores. Another rarely-used computer software which has been employed to conduct Yen analyses is 'SUBSKOR' (Skorupski, 2008, 2010), which was not available (it might have been developed for personal use). It is true that the reliabilities of CTT- and IRT-based models are conceptually different and making direct comparisons might not be feasible but CTT-based reliability formulas could be used to compute IRT-based subscores only for the purpose of comparison, as Longabach and Peyton (2018) did.

Another limitation of the study is that more recent multidimensional IRT models were not used for subscore measurement in this study. Including such models as correlated factor multidimensional IRT, higher-order model, and bifactor model is beyond the scope of one single study.

Although there is evidence that in designing NUEE more emphasis has been given to the internal consistency of the test as a whole and compartmentalization of test sections does not reflect the multidimensionality of language construct, the dimensionality of these tests needs to be investigated by future research. Since the Yen model is a unidimensional IRT model, better results of Yen could be considered as evidence for the unidimensionality of NUEE. Yen OPI subscores better preserve the original relationship of raw subscores and this model is a unidimensional IRT-based model. However, whether NUEE is unidimensional or multidimensional, and if the latter is the case, which multidimensional model better fits the data need to be addressed by future research.

7. Conclusion

In this research two subscore measurement models (Wainer and Yen) were used to investigate the subscore quality of a high-stakes test (NUEE). Compared to Wainer, Yen better maintained the original relationships observed among the raw subscores and that between each subscore and the total score. It also kept subscores more distinct from each other. However, this method reduced both within-person and between-person subscore variability. That is, Wainer subscores better discriminated the examinees and better showed the examinees' strengths and weaknesses. Altogether, no single model satisfies all criteria for subscore reporting.

Subscore augmentation eliminates the problem of low reliability for tests with short subtests but increases the correlations between subtests and makes subtests not have added value.

Declaration of Conflicting Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The corresponding author appreciates the research was supported by a grant-in-aid from Ferdowsi University of Mashhad (n=52210) in 2021

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

ACT. (2016). Interpretive guide for ACT Aspire® summative reports. Retrieved from http://actaspire.avocet.pearson.com/actaspire/Home#8439

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.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment. https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-33.pdf

Choi, I., & Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: A case of reading and listening subscores for young EFL learners. *Language Testing*, 37(2), 254-279. https://doi.org/10.1177/0265532219879654

College Board (2017). *SAT® suite of assessments technical manual: Characteristics of the SAT.* Retrieved from https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295–311. https://doi.org/10.3102/10769986030003295

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620–639. https://doi.org/10.1177/0146621608326423

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296–316. https://doi.org/10.1177/0146621610378653

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241–259. https://doi.org/10.3102/10769986031003241

Erdemir, A., & Atar, H. (2020). Simultaneous Estimation of Overall Score and Subscores Using MIRT, HO-IRT and Bi-factor Model on TIMSS Data. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 61-75. https://doi.org/10.21031/epod.645478

Fu, J. & Qu, Y. (2018). A Review of Subscore Estimation Methods. *ETS Research Report Series, 2018*(1), 1-15. https://doi.org/10.1002/ets2.12203

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220. https://doi.org/10.1207/s15324818ame1702_3

Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioural Statistics, 33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. British *Journal of Mathematical and Statistical Psychology, 62*(1), 79–95. https://doi.org/10.1348/000711007X248875

Haladyna, S.J., & Kramer, G.A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 24*(7), 349–368. https://doi.org/10.1177/0163278704270010

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Lee, M. k., Sweeney, K., & Melican, G. J. (2017). Test assembly implications for providing reliable and valid subscores. *Educational Assessment, 22* (4), 205-219. https://doi.org/10.1080/10627197.2017.1381552

Lee, M. k., Sweeney, K., & Melican, G. J. (2017). Test assembly implications for providing reliable and valid subscores. *Educational Assessment, 22*(4), 205-219. https://doi.org/10.1080/10627197.2017.1381552

Lim, E. & Lee, W. C. (2020). Subscore equating and profile reporting. *Applied Measurement in Education, 33*(2), 95-112. https://doi.org/10.1080/08957347.2020.1732381

Longabach, T., Peyton, V. (2018). A comparison of reliability and precision of subscore reporting methods for a state English language proficiency assessment. *Language Testing, 35*(2), 297-317. https://doi.org/10.1177/0265532217689949

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics, 15*(2), 91–112. https://doi.org/10.3102/10769986015002091

Luecht, R. M. (2003, April). *Applications of multidimensional diagnostic scoring for certification and licensure tests* [paper presentation]. The meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Monaghan, W. (2006). The facts about subscores (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Papageorgiou, S. & Choi, I. (2018). Adding value to second-language listening and reading subscores: Using a score augmentation approach. *International Journal of Testing, 18*(3), 207-230. https://doi.org/10.1080/15305058.2017.1407766

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment.* Washington, DC: National Academies Press.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education, 23*(3), 266–285. https://doi.org/10.1080/08957347.2010.486287

Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences).* New York: Springer.

Sawaki, Y. & Sinharay, S. (2013). Investigating the value of section scores for the TOEFL IBT® test. *ETS Research Report Series, 2013*(2), i-113. https://doi.org/10.1002/j.2333-8504.2013.tb02342.x

Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing, 35*(4), 529-556. https://doi.org/10.1177/0265532217716731

Sinharay, S. (2010). How Often Do Subscores Have Added Value? Results from Operational and Simulated Data. *Journal of Educational Measurement, 47*(2), 150-174. https://doi.org/10.1111/j.1745-3984.2010.00106.x

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey (ETS RM-08-18)*. Educational Testing Service, Princeton, NJ, United States. https://www.ets.org/Media/Research/pdf/RM-08-18.pdf

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. https://doi.org/10.1111/j.1745-3992.2007.00105.x

Skorupski, W. P., & Carvajal, J. A. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, 70(3), 357-375. https://doi.org/10.1177/0013164409355694

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*(2), 89–112. https://doi.org/10.1207/s15324818ame1702_1

United States Congress. (2002). No child left behind act of 2001. *Public Law 107-110*. 107th Cong., Washington, DC. 1425–2095.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K., Thissen, D. (2001). Augmented scores – 'borrowing strength' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Erlbaum.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116-136. https://doi.org/10.1037/1082-989X.9.1.116

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360. https://doi.org/10.1111/j.1745-3984.2010.00117.x

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83–105. https://doi.org/10.1177/0146621606291559

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance* [paper presentation]. The annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

Appendix

Table A1

Correlations Between the Subscores of NUEE in Different Subscore Estimation Methods

Subscores	Raw	Wainer	Yen
GV	0.41	0.91	0.69
GS	0.37	0.92	0.68
GF	0.36	0.88	0.65
GC	0.45	0.95	0.74
GR	0.39	0.77	0.64
VS	0.34	0.83	0.62
VF	0.34	0.78	0.61
VC	0.46	0.89	0.71
VR	0.50	0.84	0.72
SF	0.28	0.76	0.57
SC	0.36	0.83	0.66
SR	0.36	0.74	0.62
FC	0.49	0.95	0.74
FR	0.36	0.69	0.59
CR	0.49	0.8	0.71
Mean correlation	0.40	0.84	0.66
Correlation range	0.22	0.26	0.17

Table A2

Correlation Between Each Subscore and the Total Score of NUEE in Different Subscore Estimation Methods

Subscores	Raw	Wainer	Yen
GP	0.66	0.95	0.96
VP	0.73	0.94	0.84
SP	0.62	0.88	0.82
FP	0.63	0.88	0.81
CP	0.76	0.96	0.90
RP	0.83	0.93	0.85
Mean correlation	0.71	0.92	0.86
Correlation range	0.21	0.08	0.15