UCLA | CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

CALIBRATION, SCORING, AND REPORTING AT MULTIPLE LEVELS OF GRANULARITY: APPLICATION TO ALTERNATE ENGLISH LANGUAGE PROFICIENCY ASSESSMENT

CRESST Report 875

Yun-Kyung Kim & Li Cai

JUNE 2024

The work reported herein was supported by grant number Sponsor Award #S368A190007 from the U.S. Department of Education with funding to the Iowa Department of Education and subaward to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST)).

To cite from this report, please use the following as your APA 7th edition reference: Kim, Y., & Cai, L. (2024). *Calibration, Scoring, and Reporting at Multiple Levels of Granularity: Application to Alternate English Language Proficiency Assessment* (CRESST Report 875). UCLA/CRESST.

Table of Contents

Calibration, Scoring, and Reporting at Multiple Levels of Granularity: Application to Alternate English Language Proficiency Assessment

Yun-Kyung Kim & Li Cai

CRESST/University of California, Los Angeles

Executive Summary

This report describes a calibration, scoring, and reporting method applicable to an assessment that requires the results to be generated at multiple levels of aggregation. The method is demonstrated using ELPA21's Alternate English language proficiency assessment (Alt ELPA) as an example. Alt ELPA is an English language proficiency assessment designed to evaluate English language proficiency of English learners with the most significant cognitive disabilities. Alt ELPA is required to produce scores at three different levels of granularity: domain-level scores (Listening, Reading, Speaking, and Writing scores), modality-level scores (Receptive and Productive modality scores), and overall scores. In addition, due to the specificity of the target population (e.g., frequent domain exemptions), the sample size is relatively small, and the degree of missingness and incompleteness is relatively high, which poses an additional challenge to calibration.

A method proposed in this report involves using three different item factor analysis models. First, a two-dimensional item response theory model is used as the primary calibration and scoring model to calibrate item parameters and generate modality-level scores. The item response probabilities are modeled using the multidimensional graded response model, where the item slope parameters are constrained to equality within a modality to limit the number of freely estimated parameters in recognition of the small population of students eligible for the Alt ELPA. Second, using the calibrated item parameters, a four-dimensional item response theory model was used to generate augmented subscores, i.e., domain-level scores. Third, an item bifactor model is used to generate overall scores as a projection of modalities, using the item parameter estimates from the primary calibration model.

The proposed method has two primary benefits. First, it improves the precision of modality- and domain-level scores by taking advantage of the substantial correlations between modalities and domains, respectively, as represented by the multidimensional models. Second, all the scores produced are placed on the same scale, thereby facilitating interpretation and the use of standard setting results.

Calibration, Scoring, and Reporting at Multiple Levels of Granularity: Application to Alternate English Language Proficiency Assessment[1]

Yun-Kyung Kim & Li Cai

CRESST/University of California, Los Angeles

Abstract: This report describes a calibration, scoring, and reporting method applicable to an assessment requiring aggregation at multiple levels. We demonstrate the method using an example of an assessment that targets a small and diverse population, an English language proficiency assessment for English learners with the most significant cognitive disabilities.

Introduction

In the fall of 2020, more than 5.0 million English learners were enrolled in public elementary and secondary schools, representing 10.3% of the K-12 student enrollment (National Center for Education Statistics [NCES], 2023). In service of this population, an English language proficiency assessment is federally mandated to evaluate their English language proficiency (ELP) in four domains of language—listening, reading, speaking, and writing—as defined by states' ELP standards. An ELP assessment aims to ensure opportunities for English learners to achieve challenging content achievement standards as their non-English learner peers (National Center on Educational Outcomes [NCEO], n.d.). An ELP assessment is pivotal in identifying students needing English language development services and those who should be exited from their English learner status. Moreover, under the Every Student Succeeds Act (ESSA), the percentage of English learners reaching proficiency in an ELP assessment is part of states' accountability systems. As of 2024, eight states thus administer the general summative ELP assessment as part of the English language proficiency assessment for the 21st Century (ELPA21) consortium to meet federal requirements.

While an ELP assessment should serve all English learners, English learners with the most significant cognitive disabilities have additional considerations for access and have thus been particularly underserved. Though small in population size, English learners with the most significant cognitive disabilities are highly diverse in the types of disabilities they have (Christensen et al., 2018; Liu et al., 2020; 2021), and they are one of the subgroups that are

[1] This work was originally presented at the annual meeting of the National Council of Measurement in Education (NCME) held in Pennsylvania, PA, United States. We thank the discussant, Dr. Laurie Davis, for her comments and suggestions.

known to have lower reclassification rates (e.g., Shin, 2020; Thompson, 2017). The intersection of the special education subgroup and the English learner subgroup raises unique challenges for ELP assessments (Karvonen et al., 2021; Liu, Thurlow, & Quenemoen, 2015; Sullivan, 2011; Wagner, Francis, & Morris, 2005). One of the challenges is that these students are frequently exempted from one or more language domains as delineated by their Individualized Education Plans (IEPs) (e.g., a non-verbal student may be exempted from the speaking domain; a visually impaired student may be exempted from reading and writing domains, and so forth). The frequent exemptions are attributable to the fact that the ways in which English learners with the most significant cognitive disabilities receive and produce language may not be the same as the way students in the general population receive and produce language. For instance, English learners with the most significant cognitive disabilities may write by speaking, speak by writing, read by listening, or listen by reading, using a speech-to-text or text-to-speech device.

Under the mission to increase accessibility and fairness for English learners with the most significant cognitive disabilities, states committed to the development of alternate assessments of ELP. As such, the Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP; https://elpa21.org/alt-elpa) was a grant-funded project with ten participating states tasked with designing a fair and reliable assessment for this population. The resulting assessment, the Alternate ELPA by CAAELP (hereinafter: Alt ELPA), which is part of ELPA21's suite of ELP assessments, completed its first operational field test year in 2022-23. The Alt ELPA is a standards-based assessment designed for eligible EL students who are unable to participate in the general ELPA even with accommodations. The Alt ELPA measures ELP in two modalities—receptive and productive modalities—instead of the more narrowly defined four domains: reading and listening are part of the receptive modality, and speaking and writing are part of the productive modality. The modality-based assessment design allows the assessment to be delivered to English learners with the most significant cognitive disabilities with the same accessibility features or accommodations used for their instruction. Moreover, the Alt ELPA features items with lower Lexile levels, color-free artwork, fewer response options, and other means of reducing unnecessary cognitive load while still targeting alternate ELP standards. The Alt ELPA thus provides eligible students the opportunity to demonstrate their English language proficiency on an assessment based on alternate performance expectations for English language development (ELPA21, 2024a).

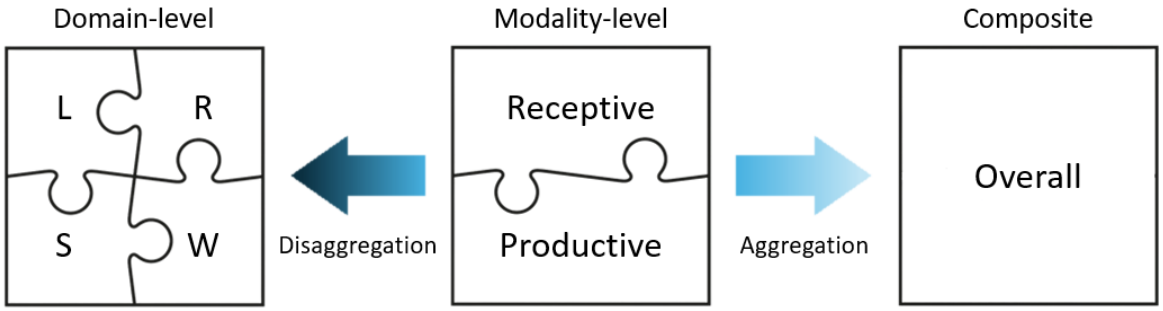
This paper recounts the calibration, scoring, and reporting of scores in the Alt ELPA, addressing two significant challenges in the process. First, in the Alt ELPA, scores need to be reported at multiple levels of aggregation. Although CAAELP participating states made a consensus-based decision, based on extensive research (Thurlow, Christensen, & Shyyan, 2016) and discussion among key stakeholders, to determine proficiency based on modality-level scores, federal legislation still requires reporting of domain-level scores and composite scores (i.e., overall scores). Hence, as shown in Figure 1, modality-level scores had to be disaggregated to domain-level scores and simultaneously aggregated to composite scores. Second, calibration

and scoring had to be done under a relatively small sample size and a high degree of data missingness due to frequent exemptions. The small sample size, attributable to the specificity of the target population, sets limits to the complexity of psychometric models that can be employed.

In the following, we first briefly review the existing score disaggregation and aggregation methods. Then, we introduce the proposed approach in a general manner. Finally, we demonstrate an implementation of the proposed approach in the context of Alt ELPA.

Figure 1

Summary of Alt ELPA Score Disaggregation and Aggregation



Existing methods

Producing subscores and overall or composite scores are often referred to as disaggregation and aggregation of scores, respectively. This section briefly introduces literature that offers various ways to conduct disaggregation, aggregation, or both simultaneously.

Disaggregation

There has been increasing interest in subscore generation due to its potential diagnostic values (e.g., Gorney & Sinharay, 2024; Haberman, 2008; Haberman et al., 2024; Sinharay, 2010; 2019). Existing methods to generate subscores include subscore augmentation (Wainer et al., 2001), objective performance index (OPI; Yen, 1987), the use of multidimensional item response theory (MIRT; Reckase, 2009) (de la Torre & Patz, 2005; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007), the use of item bifactor model (Gibbons et al., 2007; Cai, Yang, & Hansen, 2011) (Liu, Li, & Liu, 2019; Dueber & Toland, 2023), and the use of higher-order item response theory (higher-order IRT) model (de la Torre & Song, 2009). Among the methods, MIRT-based subscores are known to be optimal in that they borrow strengths across dimensions to compensate for smaller numbers of items in each dimension, which is why MIRT-based subscores are referred to as *augmented* subscores (de la Torre, Song, & Hong, 2011; Haberman & Sinharay, 2010; Yao, 2010).

Due to advances in computational algorithms, such as the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a; 2010b; 2010c), MIRT-based subscores have become more computationally feasible than ever. Especially when item parameters are fixed, the computational burden is minimal. In circumstances where item parameters need to be estimated in more than two dimensions, a shortcut proposed by Thissen (2013) can be undertaken to approximate MIRT-based subscores. By virtue of the dimension reduction technique activated by an item bifactor model, the shortcut requires an integration over only two dimensions. To simplify the process even further, prior to Thissen's (2013) procedure, one can calibrate item parameters by fitting unidimensional IRT models *within each dimension* and fixing the item parameters to the calibrated values (Hansen & Leon, 2017).

Aggregation

As the assumption of unidimensionality is useful for certain measurement practices, test practitioners often desire to compute a single score that summarizes the multidimensional traits known or presumed to be present in the data. In other words, a unidimensional approximation of a multidimensional space is used to generate overall or composite scores that are useful in practice.

A widely used method to generate composite scores is a linear projection using item factor models such as higher-order IRT model (de la Torre & Song, 2009), MIRT model (Yao, 2010), and item bifactor model (Liu et al., 2019). Yao (2010) and Liu et al. (2019) compared performances of different weights, and the weights they found to be optimal very closely approximated the weights used for *reference composite* as implied by the linear composite conjecture (LCC; Wang, 1986). The weights of the reference composite are based on the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}'\mathbf{A}$, where $\mathbf{A}$ is a matrix of item slopes on all dimensions (DeMars, 2021). The plausibility of LCC was confirmed for multidimensional test setting such as a two-dimensional compensatory IRT model with dimensions forming a positive manifold (Strachan et al., 2022). Caution is required because the reference composite can be nonlinear (i.e., LLC can be violated) under certain conditions, such as when some dimensions contain more difficult items than others (see Liao, Bolt, & Kim, 2024). Under such conditions, a nonlinear approximation (e.g., Ip et al., 2013) should be used.

An alternative approach to generating composite scores is to employ projective IRT. The projective IRT is grounded in a theory that MIRT and locally dependent unidimensional IRT models are empirically indistinguishable (Ip, 2010). In the projective IRT framework, the MIRT model can be projected onto the corresponding unidimensional IRT model with local dependencies (Ip & Chen, 2012).

Simultaneous Disaggregation and Aggregation

As an approach to generating subscores and composite scores simultaneously, the use of a higher-order IRT model was proposed (de la Torre & Song, 2009; Rijmen et al., 2014). de la

Torre and Song (2009) proposed a second-order IRT model, and Rijmen et al. (2014) extended the model to a third-order IRT model. The second-order and third-order IRT models are constrained versions of bifactor and trifactor models, respectively (Rijmen, 2010; Rijmen et al., 2014). In a second-order IRT model, overall scores and domain scores are directly obtained from *expected a posteriori* (EAP) predictions for the second- and first-order dimensions, respectively. Likewise, in a third-order IRT model, overall scores, domain scores, and sub-domain scores are directly obtained from EAP predictions for the third-, second-, and first-order dimensions, respectively.

Proposed approach

The proposed approach can be described in three steps. Separate models are specified for each aggregation level in lieu of trying to obtain scores of different aggregation levels from a single model. The separate modeling is advantageous in that it obviates the need to apply weights, does not assume any parametric relation between the scores of different aggregation levels, and keeps each model parsimonious to be stably estimated with a small sample size. At the same time, the scale of the scores is preserved by fixing the item parameter estimates to be consistent across all the models.

Step 1. A primary calibration and scoring model is specified based on the assessment design. The model is a d-dimensional IRT model where dimension(s) represent key unit(s) of the assessment. From this model, we obtain item parameter estimates (e.g., item slopes and intercepts, in the case of the two-parameter logistic IRT model), group parameter estimates (correlations between the dimensions), and EAP predictions for examinees' proficiency on d factor(s). Item parameters are calibrated using this model because it operates at the level of aggregation that is most suitable for the target population and assessment design. For instance, in an English language proficiency assessment, if the test is designed based on the theory of four intercorrelated language domains (i.e., Listening, Reading, Speaking, and Writing), these domains should be represented by correlated latent dimensions in the primary model.

Step 2. An augmented subscore model is specified based on the more fine-grained unit of assessment. The fine-grained unit is typically determined by the needs of the end-users (De Mars, 2013) or the intended structure of the test (Yao, 2010). The model is a d^+-dimensional compensatory MIRT model where item parameters are *fixed* to the values calibrated in the primary calibration model. Fixing the parameters automatically places the resulting subscores on the same scale as the scores from the primary calibration model without necessitating any linking procedure. From this model, we obtain group parameter estimates (variance-covariances among dimensions) and EAP predictions for examinees' proficiency on d^+ factors.

Step 3. A composite model is specified to obtain overall scores corresponding to the coarsest and broadest unit of assessment. The model is an item bifactor model (Gibbons et al.,

2007; Cai, Yang, & Hansen, 2011) with one general factor and d specific factors that represent the key unit of assessment. Item parameters are again *fixed* to the values calibrated in the primary calibration model. Specifically, each item's slopes on the general and specific factors are both constrained to the value calibrated in the primary calibration model. Given the constraints, the model becomes formally equivalent to a testlet response model (Wainer, Bradlow, & Wang, 2007) with a proportionality constant of one, which in turn is isomorphic to a second-order item factor analysis model (e.g., Rijmen, 2010). From this model, we obtain group parameter estimates (variance-covariances among dimensions) and EAP predictions for examinees' proficiency on the general factor and specific factors. The computation of overall scores using general and specific factor scores depends on the intended meaning of the overall scores. If one's intention is to generate overall scores with maximum information (i.e., minimum error), overall scores should be derived by reference composite (De Mars, 2013; 2021) or its approximation per Liu et al.'s (2019) Method 4. If one intends to have overall scores represent all dimensions equally, general factor scores can be employed as overall scores.

Application to Alt ELPA

This section demonstrates the implementation of the proposed approach to Alt ELPA, focusing on the psychometric models employed to generate modality-level, domain-level, and composite scores, given a small sample size and large missingness. Calibration and scoring were conducted independently for each of the six grade levels or grade bands (KG, 1, 2-3, 4-5, 6-8, and 9-12), given that Alt ELPA does not assume a vertical scale. It is assumed that the construct being measured (i.e., English language proficiency) is different between the grade levels or grade bands (for discussion on a non-vertical scale, see Thissen, 2012).

Data

In the 2022-2023 school year, an operational field test of Alt ELPA was conducted across nine states, with 3,797 students participating altogether (ELPA21, 2024b). There were 80 items (40 items in each test form) in each grade level or grade band, and the sample size ranged from 395 to 828. The 80 items are comprised of 40 receptive items (20 listening and 20 reading items) and 40 productive items (20 speaking and 20 writing items). All items in the receptive modality were dichotomously scored, while items in the productive modality included polytomous items. For instance, among items in grade band 2-3, 68 items were dichotomously scored, and 12 items—all in productive modality—were scored with four score categories. All analyses were performed using flexMIRT® version 3.64 (Cai, 2021).

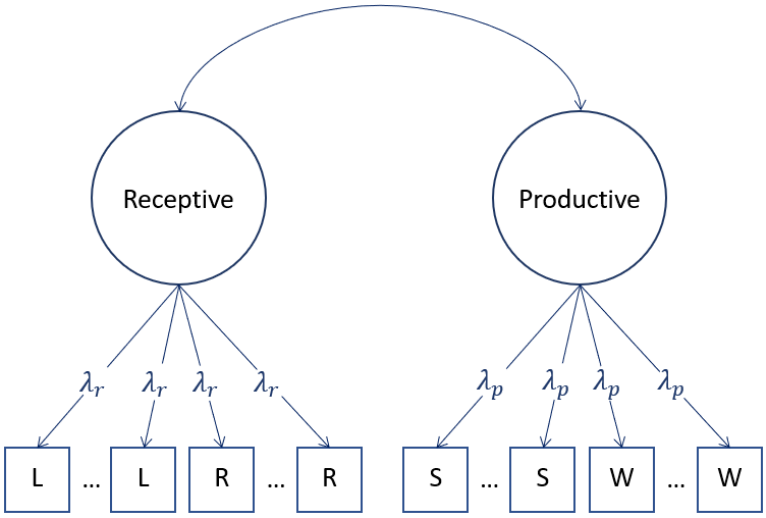
Calibration and Scoring

Primary calibration and scoring model (modality-level)

Item parameter estimates and modality-level scores were obtained by applying a MIRT model with two dimensions corresponding to two modalities, receptive and productive modalities, respectively. The model had an independent cluster factor pattern (McDonald, 2000). Each item is loaded on one of the dimensions: listening and reading items load on the receptive modality, and speaking and writing items load on the productive modality. The means and variances of the two dimensions were fixed to zeros and ones, respectively, to achieve identifiability, and the correlation between the two dimensions was freely estimated. The item response probabilities were modeled using a logistic graded response model (Samejima, 1969). The item slope parameters were constrained to equality within each modality to limit the number of free parameters in recognition of the small sample size. Thus, within each grade band, only two slope parameters (λ_r and λ_p) were freely estimated. The path diagram of this modality model is presented in Figure 2.

Figure 2

Path diagram of modality-level calibration and scoring model



Note. Squares indicate observed item responses: L, R, S, and W represent listening, reading, speaking, and writing items, respectively. Circles indicate unobserved latent dimensions.

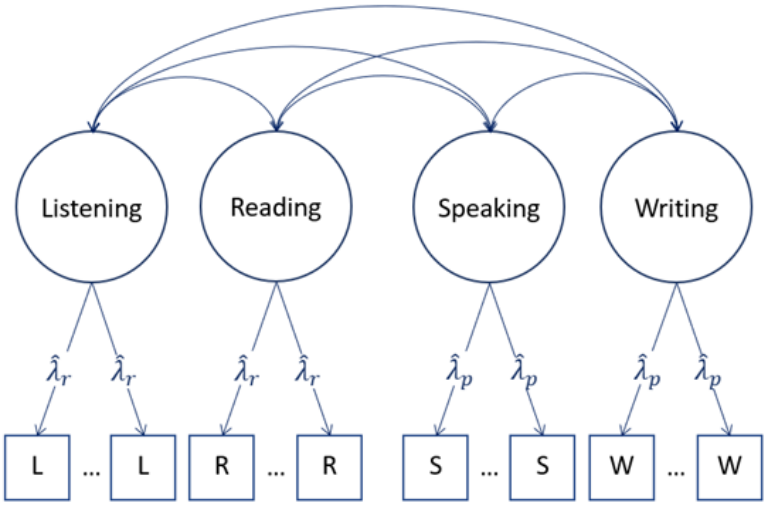
Augmented subscore model (domain-level)

Domain-level scores were obtained by applying a MIRT model with four dimensions corresponding to four language domains (i.e., listening, reading, speaking, and writing). Item parameters were fixed to the values estimated in the modality-level model. Only the means and covariances of the four dimensions were freely calibrated. The item response probabilities were

modeled using a logistic graded response model (Samejima, 1969). The path diagram of this domain model is presented in Figure 3. The resulting domain-level scores are augmented in that information from one domain is shared by all other domains via correlations between the domains.

Figure 3

Path diagram of domain-level augmented subscore model



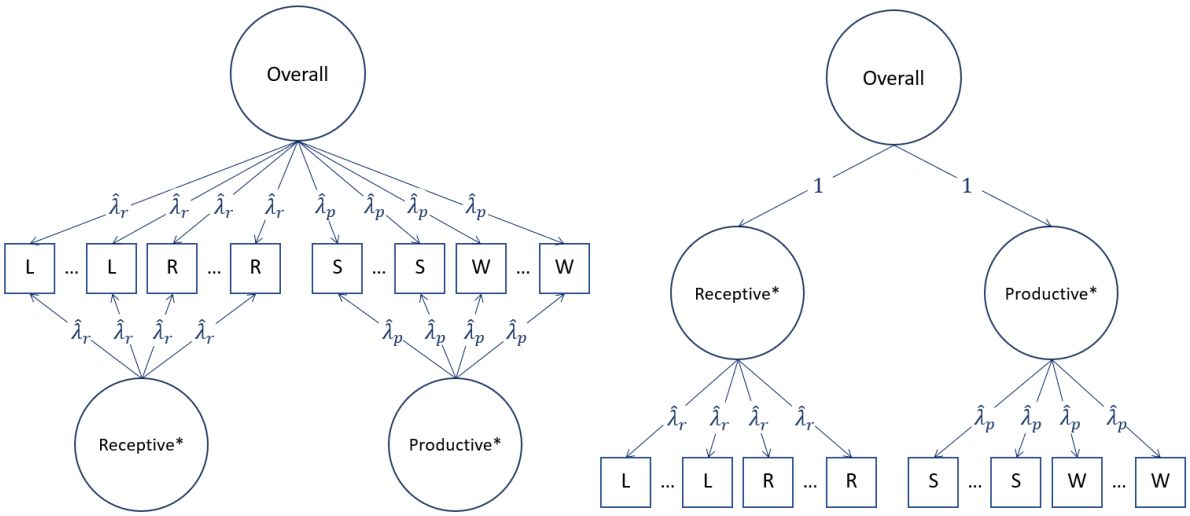
Composite model (overall)

Composite scores were obtained by applying an item bifactor model (Gibbons et al., 2007; Cai et al., 2011). The composite model projects multiple dimensions (i.e., modality-specific ELPs) onto a single dimension of interest (i.e., overall ELP). The primary/general factor corresponds to the overall ELP, while the specific factors correspond to the two modalities. Item parameters are again fixed to the values estimated in the modality model. Moreover, the variances of the specific factors are constrained to equality. The constraint is in place to ensure that the general factor is not dominated by one of the specific factors. By imposing such constraint, the general factor equally weighs items in the receptive and productive modalities (Hansen & Leon, 2017). The EAP predictions on the general dimension are adopted as overall scores.

Figure 4 illustrates how the composite model can be depicted as a constrained bifactor model (or testlet response model) or as a higher-order IRT model with second-order slopes fixed to ones. Note that the Receptive* and Productive* dimensions in the composite model bear different interpretations from the Receptive and Productive dimensions in the modality model. The Receptive* and Productive* dimensions account for the residual covariances left unexplained by the overall dimension.

Figure 4

Path diagram of overall score projection model represented as a constrained bifactor model (or testlet model; left) or as a second-order model (right)



Reporting

Following the calibration and scoring, the modality-level cut scores were established by embedded standard setting procedure (Lewis & Cook, 2020; Creative Measurement Solutions, 2024). For each modality and grade band, three cut scores were generated to categorize students into four performance levels (PLs). Moreover, the modality-level cut scores were automatically transferrable as domain-level cut scores by virtue of the modality-level scores and domain-level scores being on the same scale.

Given the uncertainty of modality-level scores, students' probabilities of being categorized to each PL were reported along with their assigned PLs. The probabilities of a student being categorized to each PL are obtained by integrating over the regions in the student's posterior distribution that are associated with the PL. The posterior distribution is approximated by a multivariate normal distribution with mean equal to the score estimates $\hat{\theta}$ and variance equal to the associated error $\sigma(\hat{\theta})$ (Rudner, 2001; 2005). For example, for $m = 1, \ldots, L$, the probability of student i being classified to PL of m defined by score range $[s_{m-1}, s_m)$ is

$$p_{im} = \Phi\left(s_m; \hat{\theta}_i, \sigma(\hat{\theta}_i)\right) - \Phi\left(s_{m-1}; \hat{\theta}_i, \sigma(\hat{\theta}_i)\right)$$

where $\Phi(q; \mu, \sigma)$ is a cumulative distribution function of a Gaussian density at q with mean μ and standard deviation σ. For $m = 1$, $s_0 = -\infty$ and thus $\Phi(s_0) = 0$. Likewise, for $m = L$, $s_L = \infty$ and thus $\Phi(s_L) = 1$. Based on the probabilities of modality-level PL assignment, we further compute the probabilities of a student being categorized to the overall PL of Emerging, Progressing, or Proficient.

Conclusion

This paper presents an approach that streamlines calibration, scoring, and reporting at multiple levels of granularity, given a small sample size with large data missingness. The approach is demonstrated using Alt ELPA, which strongly motivated the very approach. In summary, based on the modality-level scores generated by the two-dimensional MIRT model, modality-level scores are disaggregated to domain-level scores via the four-dimensional MIRT model and simultaneously aggregated to composite/overall scores via the constrained item bifactor model. All three models used fall under the umbrella of the item factor analysis (IFA) model (Bock, Gibbons, & Muraki, 1988; Cai, 2010a, 2010b, 2010c); MIRT is the IFA model without specific dimensions, and item bifactor model is a restricted hierarchical IFA model.

The key benefits of the illustrated approach are twofold. First, the precision of modality-level scores and domain-level scores is improved by virtue of the substantial correlation between dimensions. In the modality model, the correlation between the two modalities was very strong, ranging from .858 (in grade band 2-3) to .925 (in grade band 9-12). In the domain model, correlations among four domains were also strong, ranging from .788 (between listening and speaking in grade band 2-3) to .968 (between listening and reading in grade band 9-12). Second, all scores are placed on the same scale, thereby facilitating the interpretation and use of the cut scores. Hence, this approach is useful in providing scores at multiple aggregation levels (e.g., modality-level, domain-level, and overall) within the heavily constrained conditions of Alt ELPA data, such as small sample size and large data missingness.

References

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

Cai, L. (2021). flexMIRT® version 3.64: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33-57.

Cai, L. (2010c). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307-335.

Cai, L., Yang, J. S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248.

Christensen, L. L., Mitchell, J. D., Shyyan, V. V., & Ryan, S. (2018, September). *Characteristics of English learners with significant cognitive disabilities: Findings from the Individual Characteristics Questionnaire*. Madison, WI: University of Wisconsin–Madison, Alternate English Language Learning Assessment (ALTELLA).

Chung, S., & Cai, L. (2021). Cross-classified random effects modeling for moderated item calibration. *Journal of Educational and Behavioral Statistics, 46*(6), 651-681.

Creative Measurement Solutions (2024). Alt ELPA Standard Setting Technical Report [Final deliverable submitted to ELPA21 at CRESST]. Author

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: a practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311.

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.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

DeMars, C. E. (2021). A Note on the Relation Between the Angle of the Reference Composite and Liu, Li, and Liu's Method 4 for Domain Scores. *Applied Psychological Measurement, 45*(2), 130-133.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International journal of testing*, *13*(4), 354-378.

Dueber, D. M., & Toland, M. D. (2023). A bifactor approach to subscore assessment. *Psychological Methods, 28*(1), 222.

English Language Proficiency for the 21st Century. (2024a). *The Alt ELPA Technical Manual.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

English Language Proficiency for the 21st Century. (2024b). *The Alt ELPA 2022-23 Field Test Technical Report.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Gholson, M. L., & Guzman-Orth, D. (2019). *Developing an alternate English language proficiency assessment system: A theory of action*. Educational Testing Service Research Report RR–19-25.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4-19.

Gorney, K., & Sinharay, S. (2024). Added value of subscores for tests with polytomous items. *Educational and Psychological Measurement*, 00131644241268128.

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

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

Haberman, S., Sinharay, S., Feinberg, R. A., & Wainer, H. (2024). *Subscores: A Practical Guide to Their Production and Consumption*. Cambridge University Press.

Hansen, M., Leon, S. (2017). *2015-2016 English Language Proficiency Assessment for the 21st Century*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-415.

Ip, E. H. S., & Chen, S. H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement*, *36*(7), 581-601.

Ip, E. H., Molenberghs, G., Chen, S. H., Goegebeur, Y., & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*(4), 534-562.

Karvonen, M., Clark, A. K., Carlson, C., Wells Moreaux, S., & Burnes, J. (2021). Approaches to identification and instruction for students with significant cognitive disabilities who are

English learners. *Research and Practice for Persons with Severe Disabilities, 46*(4), 223-239.

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39*(1), 8-21.

Liao, X., Bolt, D. M., & Kim, J. S. (2024). Curvilinearity in the Reference Composite and Practical Implications for Measurement. *Journal of Educational Measurement*.

Liu, K. K., Thurlow, M. L., & Quenemoen, R. F. (2015). *Instructing and assessing English learners with significant cognitive disabilities*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Liu, K. K., Thurlow, M. L., Lazarus, S. S., & Dosedel, M. (2020). *A literature review of evidence-based literacy assessment and instruction practices for English learners with significant cognitive disabilities* (NCEO Report 422). National Center on Educational Outcomes.

Liu, K. K., Wolforth, S., Thurlow, M. L., Jacques, C., Lazarus, S. S., & August, D. (2021). *A framework for making decisions about participation in a state's alternate ELP assessment* (NCEO Report 426). National Center on Educational Outcomes.

Liu, Y., Li, Z., & Liu, H. (2019). Reporting valid and reliable overall scores and domain scores using bi-factor model. *Applied Psychological Measurement, 43*(7), 562-576.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114.

National Center on Educational Outcomes (NCEO). (n.d.). *English Language Proficiency (ELP) Assessments*. https://nceo.info/Assessments/elp_assessment

National Center for Education Statistics (NCES). (2023). *English Learners in Public Schools. Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved October 3, 2023, from https://nces.ed.gov/programs/coe/indicator/cgf

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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*(3), 361-372.

Rijmen, F., Jeon, M., Von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 39*(4), 235-256.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14), 1-5.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13), 1-4.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monographs No. 17). Richmond, VA: Psychometric Society.

Shin, N. (2020). Stuck in the middle: Examination of long-term English learners. *International Multilingual Research Journal, 14*(3), 181-205.

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

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2018). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.) *Score reporting research and applications* (pp. 35-49). Routledge.

Sinharay, S. (2019). Added value of subscores and hypothesis testing. *Journal of Educational and Behavioral Statistics, 44*(1), 25-44.

Strachan, T., Cho, U. H., Ackerman, T., Chen, S. H., de la Torre, J., & Ip, E. H. (2022). Evaluation of the linear composite conjecture for unidimensional IRT scale for multidimensional responses. *Applied Psychological Measurement, 46*(5), 347-360.

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children, 77*, 317–334.

Thissen, D. (2012). *Validity Issues Involved in Cross-Grade Statements about NAEP Results*. American Institutes for Research.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (pp. 29–40). New York: Springer.

Thompson, K. D. (2017). English learners' time to reclassification: An analysis. *Educational Policy, 31*(3), 330-363.

Thurlow, M.L., Christensen, L.L., and Shyyan, V.V. (2016). White Paper on English Language Learners with Significant Cognitive Disabilities. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, English Language Proficiency Assessment for the 21st Century.

Wagner, R. K., Francis, D. J., & Morris, R. D. (2005). Identifying English language learners with learning disabilities: Key challenges and possible approaches. *Learning Disabilities Research & Practice, 20*(1), 6-15.

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

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

Wang, M. M. (1986). *Fitting a unidimensional model to multidimensional item response data: The effect of latent trait misspecification on the application of IRT*. (Research Report MW: 6-24-85). Iowa City, IA: University of Iowa.

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.

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.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360.

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

UCLA

CRESST

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532
www.cresst.org