Evaluation of Vertical Linking Items in a Learning Progression Context

Unhee Ju, Jing Jiang, Han Yi Kim

Introduction

There is an increasing use of learning progressions (LPs) in developing vertical linking (VL) items and in interpreting the subsequent vertical scale (VS) (Briggs & Dadey, 2015). Unlike the traditional approach to construct a VS, LPs are intended to help overcome construct shift across grade levels by postulating learning as a sequence of increasing complexity and sophistication in thinking, understanding, and skills within a well-defined content domain (Briggs & Peck, 2015; Wilson & Bertenthal, 2005). However, little research has evaluated VL items from an LP perspective. Since each LP is composed of relatively lower number of VL items and inferences about students' progress mainly relies on the VL items across adjacent grades, a detailed study is needed to explore how best to use LPs in conjunction with VS, so that we get reliable and valid inferences.

Thus, the present study evaluates an LP-based VS – identifying problematic VL items and investigating their causes, especially as they relate to the LPs. Specifically, it is investigated whether the VL items behave differently across different LPs compared to the entire data, as well as whether on-grade VL items perform differently from off-grade. This helps us look for LP interaction effects. For these purposes, this study employed new procedures to evaluate the parameter stability of VL items and the efficacy of the Stocking-Lord (S-L) transformations.

Method

Data

Data were collected from the Measured Progress eMPower Assessment program, a new college and career readiness assessment solution consisting of ELA and mathematics tests in grades 3 to 8. The grades were vertically linked via common items between adjacent grades. Test lengths ranged from 64 to 66 points for ELA, and 48 to 54 points for mathematics. The number of linking points for a pair of adjacent grades was 82 for ELA, and 50 for mathematics. Details about the number of items and the range of points per LP-based linking set for both subject areas can be found in Table 1 and Table 2.

Each LP that was used to link a pair of grades consisted of items from both grades that were administered to students in both grades. Thus, in any one grade, students took both on-grade items and off-grade items. The typical sample size within a certain grade level was 12,500 for an on-grade item, and 1,000 for an off-grade item.

In conducting our investigation, because of time considerations we limited ourselves to studying three pairs of grades: one pair in elementary school (3/4), one pair linking elementary and middle school (5/6), and one pair in middle school (7/8).

Insert Table 1 and Table 2 about here

Procedure

VSs can be established for each of LPs respectively, but depending on the number of items for each LP, two or more LPs were combined to get the sufficient number of VL items to evaluate the linking behavior of the items. The combinations were carefully chosen, so that LPs measuring similar content strands were combined. For instance, one of combined LPs of interest, “Key Ideas and Details” in reading assessment covers two content strands of “Reading Literature Key Ideas and Details” and “Reading for Informational Text Key Ideas and Details.” Table 3 and Table 4 show the LP combinations for reading and mathematics, respectively, used in our study.

Insert Table 3 and Table 4 about here

After calibrating item parameter estimates, to detect problematic VL items, this study used four screening procedures: IRT a-a plot, IRT b-b plot, delta plot, and an iterative procedure using beta index which provides a more detailed evaluation. The three-plot procedures were used to select stable VL items before the S-L transformation, whereas a one-item-at-a-time DIF-like analysis was conducted to identify drifting items after the transformation. Then, the consistency of the SL transformation was evaluated using a new index based on the Haebara-based index (HBI) criterion. The three plots (IRT a-a plot, IRT b-b plot, and delta plot) are standard operating procedures used by Measured Progress. Recently we developed new statistics that give additional information about the performance of the VL items. We found that the union of all these give use a very thorough and complete analysis of the linking items. Using such methods, this study analyzed all VL items together as well as the separate LPs in order to see if the separate LPs behave differently than the combination of all VL items with respect to: (a) flagged items, (b) the S-L linking coefficients, and (c) HBI.

Step 1. Item calibration. Item parameters were separately estimated via marginal maximum likelihood estimation (MMLE) using PARSCALE 4.1 (Muraki & Bock, 2003) with a three-parameter logistic (3PL) IRT model as follow:

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left(-Da_i(\theta_j - b_i)\right)}$$

where $P_i(\theta_j)$ is the probability that an examinee *j* responds to item *i* correctly, a_i is the item discrimination, b_i is the item difficulty, c_i is the pseudo-guessing parameter, and *D* is a scaling constant of 1.7.

Step 2. Pre-evaluation of vertical linking items. Before the Stocking-Lord scale transformation, the three types of plots – IRT b-b plot, IRT a-a plot, and delta plot – were used to identify the misbehaving VL items.

IRT b-b plot and IRT a-a plot. The IRT *b-b* plot and IRT *a-a* plot methods (Huff & Hambleton, 2001) are based on item difficulty estimates (i.e., *b*-parameter) and discrimination estimates (i.e., *a*-parameter) for the VL items between two adjacent grades. That is, the IRT b-b plot method plots *b*-parameter estimates of the VL items plotted against each other, while the IRT a-a plot method does using *a*-parameter estimates in order to find outliers. The outliers denote items that are behaving differently for the adjacent groups with respect to the magnitude of *a*- or *b*-parameter. To determine which item is an outlier, a best-fitting line was drawn that minimizes the perpendicular distance through the estimates. Any item that lays more than three standard deviations of the distances away from this line is a candidate for exclusion from the set of VL items. The general recommendation is usually not to remove any problematic VL items detected only from these plots unless there is a compelling reason for the differential performance in the adjacent grade scales. However, as Lam (2009) suggested in practice, our study did linking in both ways, excluding misbehaving VL items and including them in the linking process, and compared results obtained from each grade.

Delta plot. The delta-plot method (Angoff, 1972; Michaelides, 2003) is based on the *p*-values, the proportion of correct response of the VL items, for the VL items between the two adjacent grades. It is conducted in a very similar way to the IRT b-b/a-a plots. Since the delta plot method does not require IRT calibration, it is widely used in practice as a simple and comprehensible graphical method to obtain evidence at first sight regarding large changes in the *p*-values across adjacent grades. In the delta-plot method, the *p*-values are transformed to the delta metric through linear transformation of the inverse normal cumulative function (Dorans and Holland, 1993):

$$\Delta = 13 - 4\{\Phi^{-1}(p)\}$$

The delta metric has a mean of 13 and a standard deviation of 4, and higher values represent more difficulty items. Like the IRT a-a/b-b plots, delta values of the VL items obtained from two adjacent grades are then plotted against each other. To look for outliers, a best-fitting line is drawn through the values and a VL item is flagged for taking a caution for investigation if the item lays more than three standard deviations from this line. The flagged items are candidates for exclusion from the VL item set. Note that the delta plot method can be easily incorporated to detecting misbehaving polytomous items by rescaling the item responses to a scale of zero to one and plotting the plots along with the dichotomous items (Sukin, 2010).

Step 3. Linking (scale transformation). By the property of scale invariance to linear transformation, a single linear function can make two separate IRT parameter estimates put on the same scale. STUIRT (Kim & Kolen, 2004) was used to find linking coefficients – A (slope) and B (intercept) of the linear function – using the Stocking-Lord method (Stocking & Lord, 1983). The Stocking-Lord method is to find the linking coefficients that minimize the summation of the squared difference between the test characteristic curves (TCCs) of a set of equating items over examinees – criterion function:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}\left(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}\right) - \sum_{j:V} p_{ij}\left(\theta_{Ji}; \frac{\hat{a}_{Jj}}{A}, A\hat{b}_{Jj} + B, \hat{c}_{Jj}\right)\right]^2$$

$$SLcrit = \sum_{i} SLdiff(\theta_i)$$

Using the linking coefficients, the item parameter estimates on the lower grade scale (labeled Year 1) were then linearly transformed to the upper grade scale (labeled Year 2) by using

$$a_{i,Year1\ on\ Year2} = \frac{a_{i,Year1}}{A_i}$$

$$b_{i,Year1\ on\ Year2} = A_i b_{i,Year1} + B_i$$

and

$$c_{i,Year1} = c_{i,Year2}.$$

Step 4. Post-evaluation of vertical linking items using beta. This iterative procedure evaluates the stability of VL items one by one. Item parameter drift (IPD) is estimated by calculating the average weighted difference between the Year 2 and the transformed Year 1 item response functions (Shealy & Stout, 1993):

$$\beta = \int \left(P_{year2}(\theta) - P_{year1\_SL}(\theta)\right) f(\theta) d\theta$$

where $P_{year2}(\theta)$ and $P_{year1\_SL}(\theta)$ indicate the probability of correct response on an item in the Year 2 and the transformed Year 1 groups respectively, $f(\theta)$ is the density function for θ, and β is a weighted expected score difference between Year 1 and Year 2 examinees of the same ability on the item. If the absolute value of β is greater than a reasonable criterion (e.g., 0.05 or 0.10), the item is flagged and removed. The scale transformation and beta analysis are repeated until no more items are flagged, or until the remaining VL items represent 20 percent of all items, and the resulting items are the final VL set. This set of items is used to generate the scale transformation based on the S-L linking coefficients of slope and intercept, and all parameters estimates of Year 1 items are adjusted and checked for IPD.

Step 5. Evaluate the efficacy of the transformation. The consistency of the Stocking-Lord transformation was evaluated, using a new index based on the sum of the root mean squared differences between the item response functions (IRFs) on the upper grade and linked lower grade scales. This index is named the *Haebara-based index* (HBI) because of its similarity to the criterion of the Haebara linking method (Haebara, 1980). For the dichotomous items, the HBI can be defined as

$$HBI = \frac{\sum_{j=1}^{n} \sqrt{\frac{\sum_{i=1}^{m}\left[p_{i,Year2}(\theta_j) - p^*_{i,Year1}(\theta_j)\right]^2}{m}}\, g(\theta_j)}{\sum_{j=1}^{n} g(\theta_j)}$$

where $p_{i,Year2}(\theta_j)$ and $p^*_{i,Year1}(\theta_j)$ are the IRFs of item *i* (*i* = 1, … , *m* for the VL items) on the Year 2 scale and the transformed Year 1 scale for a given examinee *j* (*j* = 1, … , *n*), and $g(\theta_j)$ is weightsand. Note that HBI is on a 0 to 1 scale with an interpretation similar to the standardization DIF statistic (Dorans & Kulick, 1986). The critical value of HBI with at least sample size of 1000 examinees was .043 at a significance level of .05.

As our data include both dichotomous items and polytomous items, we extended the above HBI equation to incorporate the polytomous items as follows:

$$HBI_{poly} = E_\theta[HBI|\theta] = \frac{\sum_{j=1}^{n} \sqrt{\frac{\sum_{i=1}^{m}\left[E(Y_{i,Year2}|\theta_j) - E^*(Y_{i,Year1}|\theta_j)\right]^2}{m}}\, g(\theta_j)}{\left(\frac{\sum_{i=1}^{m} k_i}{m}\right) \sum_{j=1}^{n} g(\theta_j)}$$

where $E(Y_{i,Year2}|\theta_j)$ and $E^*(Y_{i,Year1}|\theta_j)$ denote the expected score point of item *i* on the Year 2 scale and the transformed Year 1 scale for given an examinee *j*, and k_i is the number of points for item *i*. As the HBI value is close to zero, it indicates that the resulting transformation adequately adheres to the underlying theoretical assumption that the same single linear transformation satisfactorily transforms each of the set of VL items.

Results

Qualitative

First results we would like to comment on are the content of the VS in terms of its LP components. Note that for reading, there were three LP-based linking sets for each grade, and the same combinations of LPs were used for each linking set across the grades (see Table 3). In other words, the LP-based linking sets for reading were consistent from elementary through middle school. On the other hand, for mathematics, the number of LP-based linking sets differed for every pair of grades, and the combinations of LPs that composed the linking sets were all different from each other (see Table 4). For reading, one can make a fairly strong argument that a VS can be based on LPs in such a way that the VS has a consistent meaning as you go from grade to grade. However, for mathematics, the meaning of a VS based on LPs seems to vary substantially as one moves across the grades from elementary school to middle school.

Quantitative

Presented in Table 5 and Table 6 are the results of our analyses. Overall, the results indicated that we could successfully link pairs of grades using either all the items or using the LP combinations separately. Also, the VL items sometimes behaved differently depending on the LP combinations for both reading and mathematics.

Insert Table 5 and Table 6 about here

Grade-level composition of VL items across LPs. For reading, an initial set of VL items consisted of roughly equal proportion of the upper and lower grade-level VL items for most LPs and grade pairs, except for LP2 in Grade 3/4 and Grade 5/6, and LP3 in Grade 3/4. The

Grade 3/4 LP2 had more lower-grade VL items, while Grade 3/4 LP3 and Grade 5/6 LP2 had a larger amount of higher-grade VL items. According to the screening results, we can see that for Grade 3/4, there were in total eight items detected as drifting items across all three LP combinations, and five out of them were from the lower grade level (Grade 3). However, for Grade 5/6, there were also eight drifting items totally, and all of them were from the higher grade level (Grade 6). For Grade 7/8, the misbehaving items were equally from lower and upper grade levels.

Meanwhile, for mathematics, the initial set of the VL items also consisted of roughly equal proportion of the grade-level VL items across most LPs and grade pairs, except for one LP combination in each of Grade 5/6 and Grade 7/8. In both of these cases, one of the LP combinations had a higher proportion of lower grade VL items than upper grade VL items. Through our screening procedures, drifting items were detected across all grade levels and LP combinations. Unlike reading, misbehaving items for most LPs were almost equally from lower and upper grades, but for LP1 within Grade 5/6 most drifting items – 7 out of 9 – were from the lower grade level (Grade 5). Within the Grade 7/8 pair, final sets of VL items for LP2 and LP4 were composed of only one grade-level items either from Grade 7 or from Grade 8. Additionally, in the LP-based analyses, fewer items were dropped overall for Grade 3/4, but more items were dropped overall for Grade 7/8 in terms of percentage of points being dropped. In Grade 5/6, similar numbers of items were removed from the linking set overall.

Among the three LP-based linking sets for reading, LP2 seems to be the most problematic since a large portion of items were flagged, especially in Grade 5/6 and Grade 7/8. A possible reason might be that LP2 has less items than other LPs, but it is highly recommended to take a close look at the items within this LP-based linking set. With regard to mathematics, among all LPs across grade pairs, one LP-based linking set in Grade 5/6 and one in Grade 7/8 have a larger proportion of items that behave anomalously across the two grades. Thus, it is necessary to scrutinize the VL items in these LPs with respect to specific content strands covered in the relevant grade levels.

Linking coefficients. In almost all cases, except for mathematics Grade 5/6, at least one instance occurred where two LPs had notably (or sometimes huge) differences in their linking coefficients, indicating multidimensionality across the two grades in terms of their LPs.

Specifically, in terms of reading, for all three pairs of grades, at least one LP-based linking set had notable differences in their intercept linking coefficient. In addition, for mathematics, in Grade 3/4, one LP-based linking set had its intercept linking coefficient substantially different from the overall intercept. Within Grade 7/8, a similar result occurred except that the difference in intercepts was quite large. However, within Grade 5/6, the linking coefficients for all LP combinations were similar to each other as well as similar to the overall linking coefficients.

Haebara-based index (HBI). Generally, for both reading and mathematics, the HBIs were usually larger for the entire set of VL items than LP-based linking set. It is understandable since the entire set has more items, and it is more probable to have drifting items, which might result in higher HBIs.

For reading, when taking a look at the initial set of items, we found that the HBI showed that the relative performance of LP-based linking sets within each grade pair varied across grade pairs. For example, in Grade 3/4, the corresponding HBIs were all less than the critical value, however, in Grade 5/6, one LP combination performed much worse than the other two, while in Grade 7/8, one LP combination performed slightly worse than the other two.

As expected, after removing flagged problematic items, nearly all the final sets of VL items displayed HBI values under the critical value of 0.043, indicating that our screening procedures were successful in identifying problematic linking items.

Specifically the results showed that the traditionally used IRT b-b and delta plots might be problematic in certain cases, since after excluding the drifting items detected from these two procedures, the HBIs sometimes increased (e.g, Grade 7/8 LP1), indicating that non-drifting items were detected incorrectly based on the b-b and delta plots. Additionally, even after the IRT b-b and delta plot procedures, the HBI values for certain set of items still did not meet the criterion (e.g., much greater than 0.043). In other words, by using the IRT a-a plot and the newly developed beta method improved performance was obtained since most of the HBI values of remaining items from IRT a-a plot and beta methods were smaller than the critical value.

For mathematics, for all the initial sets of LP-based VL items, except LP4 in Grade 7/8, the HBI values indicate that at least some VL items had different psychometric models across the two grades. After excluding misbehaving items, the final LP-based sets of VL items did function well in the linking process across the two grades for all grade pairs, showing lower HBIs than the critical value. All the final sets of VL items across LPs and grade pairs had desirable HBI values that were lower than the suggested critical value.

Based on the HBI values, sometimes, items detected from standard operational screening procedures such as IRT a-a, b-b, or delta plots did not contribute to improved Stocking-Lord performance across the two grades, producing rather higher HBI values. For instance, the IRT a-a plot identified one drifting item from the LP4-based linking set within the Grade 7/8 pair, but the item was not really an outlier as it was not far away from the best-fit line in the plot. So removing this item resulted in increasing HBI, implying worse behavior. Thus, we decided not to exclude this item from the final linking set. Hence, it would be recommended not to automatically drop items detected only from such plot procedures before implementing the Stocking-Lord transformation. Based on HBI values computed from different screening procedures, the combined use of IRT a-a plot and the iterative beta procedure may work better than the IRT b-b and delta plots in order to evaluate the performance of the VL items in the linking process.

Discussion

The current study evaluates the performance of VL items from the LP perspective, and examines the causes of misbehaving items that relate to the LPs. For this purpose, we used standard operational screening procedures (i.e., IRT a-a, IRT b-b, and delta plots) before the Stocking-Lord transformation as well as newly developed statistics (i.e., iterative procedure using beta index and Haebara-based index).

First, we noted that the qualitative nature of the LP combinations used to link pairs of grades displayed major differences between math and reading. The reading LP combinations displayed consistent content across all three grade pairs that were investigated, whereas the content of the math LP combinations varied greatly across the three grade pairs. These results have important obvious implications for interpretation of the VS in terms of the LPs.

Second, for both reading and mathematics, the initial set of VL items displayed HBIs showing that at least some of the VL items are behaving differently across the two grade levels. As expected, after conducting our screening procedures including b-b plots, delta plots, a-a plots, and beta analysis, the final set of VL items showed much improved HBIs. The composition of the initial set of VL items consisted of roughly the same proportion of the grade-level VL items across most LPs. After removing items based on our screening procedures, the composition sometimes changed. For some LP combinations, the removed items were disproportionately from the lower grade while for other LP combinations they were from the upper grade. On the other hand, some LP combinations maintained their initial composition. More research is needed to better understand why these differences occurred.

Third, we also observed non-negligible differences in linking coefficients across the LP combinations. While the overall results for a grade pair seemed to indicate that a unidimensional VS could be created across the two grades, the differences in S-L coefficients across the LP combinations indicate that separate LP combinations sometimes behave as distinct vertical scales. This apparent multidimensionality across grade levels in terms of their LPs is worth being further investigated, e.g., using MIRT.

Overall, this study shows a promising initial step in the direction of evaluating a LP-based set of VL items. Further research with deeper investigation to better explain the observed differences should be conducted. In particular, exploring MIRT models may shed important light on the results observed here. Finally, replication of these results with other datasets should be established.

References

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper Presented at the Annual Meeting of the American Psychological Association, Honolulu (ERIC Document Reproduction Service No. ED 069686).

Briggs, D. C. & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment, 20*(1), 1-22.

Briggs, D. C., & Peck, F. A. (2015a). Using Learning Progressions to Design Vertical Scales that Support Coherent Inferences about Student Growth, *Measurement: Interdisciplinary Research and Perspectives*, 13:2, 75-99, DOI: 10.1080/15366367.2015.1042814

Dorans, N. J., and Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Function.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355-368.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Huff, K. L., & Hambleton, R. K. (2001). *The detection and exclusion of differentially functioning anchor items*. (Center for Educational Assessment Research Report No. 415). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer software]. Iowa City, IA: Iowa Testing Programs, University of Iowa. Available from http://www.education.uiowa.edu/casma

Lam, W. (2009). Linking current and future score scales for the AICPA uniform CPA exam.

Technical Report W0902. Ewing, NJ: American Institute of Certified Public Accountants.

Michaelides, M. P. (2003, April). *Sensitivity of IRT equating to the behavior of test equating items*. Paper presented at the AERA annual meeting in Chicago, IL.

Muraki, E., & Bock, R. D. (2003). PARSCALE (Version 4.1): IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201-210.

Sukin, T. M. (2010). *Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees*. Doctoral dissertation, University of Massachusetts Amhers.

Wilson, M., & Bertenthal, M. (2005). *Systems for state science assessment*. Washington, DC: National Academy Press.

Tables

Table 1.

Number of Vertical Linking (VL) Items and Points by Learning Progression for Reading

Grade Pairs	LP-based Linking Set	Number of VL Items	Number of Points
Grade 3/4	LP1: Craft and Structure	9	12
	LP2: Integration of Knowledge and Ideas	8	13
	LP3: Key Idea and Details	13	17
Grade 5/6	LP1: Craft and Structure	10	12
	LP2: Integration of Knowledge and Ideas	7	9
	LP3: Key Idea and Details	13	22
Grade 7/8	LP1: Craft and Structure	10	13
	LP2: Integration of Knowledge and Ideas	4	9
	LP3: Key Idea and Details	16	22

Table 2.

Number of Vertical Linking (VL) Items and Points by Learning Progression for Mathematics

Grade Pairs	LP-based Linking Set	Number of VL Items	Number of Points
Grade 3/4	LP 1: Algebraic Thinking + Operations + Number and Number Systems	14	16
	LP 2: Measurement + Geometric Properties	15	24
Grade 5/6	LP 1: Functions + Number and Number Systems + Geometric Properties	19	29
	LP 2: Fractions + Measurement + Statistics and Data	7	7
	LP 3: Algebraic Thinking	10	12
Grade 7/8	LP 1: Functions + Fractions	12	13
	LP 2: Probability + Statistics and Data	10	15
	LP 3: Algebraic Thinking + Operations	11	16
	LP 4: Geometric Properties + Spatial geometry	4	5

Table 3.

Combinations of Learning Progressions for Reading

Grade Pair	LP Combinations	Learning Progression (LP)
Grade 3/4 Grade 5/6 Grade 7/8	LP1: Craft and Structure	Reading for Informational Text Craft and Structure
		Reading Literature Craft and Structure
	LP2: Integration of Knowledge and Ideas	Reading for Information Text Integration of Knowledge and Ideas
		Reading Literature Integration of Knowledge and Ideas
	LP3: Key Ideas and Details	Reading for Informational Text Key Ideas and Details
		Reading Literature Key Ideas and Details

Table 4.

Combinations of Learning Progressions for Mathematics

Grade Pair	LP Combinations	Learning Progression (LP)
Grade 3/4	LP1	Algebraic Thinking
		Operations
		Number and Number Systems
	LP2	Measurement
		Geometric Properties
Grade 5/6	LP1	Functions
		Number and Number Systems
		Geometric Properties
	LP2	Fractions
		Measurement
		Statistics and Data
	LP3	Algebraic Thinking
Grade 7/8	LP1	Functions
		Fractions
	LP2	Probability
		Statistics and Data
	LP3	Algebraic Thinking
		Operations
	LP4	Geometric Properties
		Spatial geometry

Table 5.

Results of Evaluation of Vertical Linking Items for Reading

Reading		Initial set				Final set			
		N. of items (points)	S-L linking coefficients		HBI	*N.* of items (points)	S-L linking coefficients		HBI
			A	B			A	B	
Grade 3/4	All	64(82)	0.9710	-0.4653	0.0441	48 (64)	0.9712	-0.4994	0.0351
	LP1	9(12)	0.9190	-0.4445	0.0328	7 (10)	0.9131	-0.4174	0.0215
	LP2	8(13)	0.9322	-0.5958	0.0312	6 (11)	0.9166	-0.6690	0.0172
	LP3	13(17)	1.0268	-0.5027	0.0361	9 (12)	1.0312	-0.4902	0.0322
Grade 5/6	All	64(83)	0.9482	-0.2326	0.0555	48 (64)	0.9520	-0.2686	0.0320
	LP1	10(12)	0.9884	-0.3090	0.0282	8 (10)	0.9672	-0.3175	0.0246
	LP2	7(9)	1.0023	-0.1028	0.0595	4 (5)	1.0049	-0.0682	0.0371
	LP3	13(22)	0.9450	-0.2890	0.0284	10 (19)	0.9351	-0.2926	0.0267
Grade 7/8	All	63(82)	0.9915	-0.2230	0.0811	54 (69)	0.9982	-0.2660	0.0335
	LP1	10(13)	1.0188	-0.3834	0.0471	8 (11)	0.9954	-0.3965	0.0452
	LP2	4(9)	0.9351	-0.1934	0.0298	3 (7)	0.9504	-0.1141	0.0091
	LP3	16(22)	1.0541	-0.2438	0.0364	14 (20)	1.0201	-0.2615	0.0315

Table 6.

Results of Evaluation of Vertical Linking Items for Mathematics

Math		Initial set				Final set			
		N. of items (points)	S-L linking coefficients		HBI	*N*. of items (points)	S-L linking coefficients		HBI
			A	B			A	B	
Grade 3/4	All	38 (50)	0.9507	-0.5171	0.0730	23 (34)	1.0063	-0.5510	0.0367
	LP1	14 (16)	1.0736	-0.7661	0.0375	9 (11)	1.1091	-0.8841	0.0301
	LP2	15 (24)	0.9156	-0.4463	0.0479	10 (19)	0.9966	-0.4414	0.0262
Grade 5/6	All	39 (51)	1.0221	-0.4650	0.0687	21 (31)	0.9966	-0.3857	0.0369
	LP1	19 (29)	1.0068	-0.4178	0.0664	9 (18)	1.0120	-0.4311	0.0344
	LP2	7 (7)	1.0553	-0.4225	0.0442	5 (5)	1.0077	-0.3029	0.0307
	LP3	10 (12)	0.9741	-0.4210	0.0655	6 (7)	0.9865	-0.4551	0.0342
Grade 7/8	All	37 (49)	0.9898	-0.2613	0.0571	26 (38)	1.0126	-0.2695	0.0427
	LP1	12 (13)	1.0284	-0.2487	0.0553	6 (7)	0.9780	-0.3379	0.0315
	LP2	10 (15)	1.0645	-0.2887	0.0625	4 (8)	0.9644	-0.0359	0.0176
	LP3	11 (16)	0.8954	-0.2587	0.0466	7 (11)	0.9491	-0.3639	0.0298
	LP4	4 (5)	0.9885	-0.2065	0.0232	4 (5)	0.9885	-0.2065	0.0232