NOVEMBER 2025

Preliminary Validity Evidence for the IXL LevelUp™ Diagnostic for Math: Construct Validity and Criterion Validity Using NWEA MAP® Growth™

Xiaozhu An, Ph.D.

Table of Contents

_____

Executive Summary

_____

IXL is an end-to-end teaching and learning solution that engages learners in Pre-K through 12th grade with a comprehensive curriculum, first-of-its-kind assessment suite, and personalized recommendations for meeting learning goals. A core component of the IXL Assessment Suite is the IXL LevelUp™ Diagnostic, an innovative, just-in-time assessment solution that uniquely connects assessment, insights, and instruction. The LevelUp Diagnostic in Benchmark mode is designed to confidently identify students for intervention, analyze what is and isn't working within and across schools, predict performance on state assessments, and plan ahead for the next school year without waiting on official state test results. The LevelUp Diagnostic can also be utilized between Benchmarks in Real-Time mode to get up-to-the-minute insights into students' performance and easily update and refine their scores and recommendations. See the IXL LevelUp Diagnostic technical manual for more details.

The goal of the present study was to assess the validity of the IXL LevelUp Diagnostic for math (LevelUp Math), including internal structure, multi-group measurement invariance, and criterion validity using the NWEA MAP® Growth™ assessment (MAP Growth)[1]. We found that:

- LevelUp Math demonstrated a coherent, one-factor structure as a measure of math achievement, supporting the use of a unidimensional Rasch model in scoring.
- LevelUp Math demonstrated measurement invariance across key student subgroups, including gender, race, ethnicity, English learner status, and special education status.
- There were strong positive correlations between LevelUp Math and MAP Growth performance in Kindergarten through 8th grade.

---

[1] _Note: MAP®, MAP® Growth™, and NWEA are trademarks owned by HMH Education Company, and their use here is for identification purposes only. No claim of affiliation or endorsement is made._

Preliminary Validity Evidence for the IXL LevelUp™ Diagnostic for Math: Construct Validity and Criterion Validity Using NWEA MAP® Growth™

Background

_____

IXL is an end-to-end teaching and learning solution that engages learners in Pre-K through 12th grade with a comprehensive curriculum, first-of-its-kind assessment suite, and personalized recommendations for meeting learning goals. It covers five main subject areas: mathematics, English language arts (ELA), science, social studies, and Spanish. As of this writing, IXL is used by 30% of students in the U.S. and more than 17 million students worldwide. IXL is deeply rooted in learning sciences research (see Bashkov et al., 2021) and engages each student in a personalized learning experience tailored to their working level. As a result, students work through problems that are neither too easy nor too difficult, which in turn supports their self-efficacy and motivation for continued learning (An & Schonberg, 2024).

A core component of the IXL Assessment Suite is the IXL LevelUp Diagnostic, an innovative, just-in-time assessment solution that uniquely connects assessment, insights, and instruction for students in Pre-K through 12th grade. The IXL LevelUp Diagnostic for math (LevelUp Math) is designed to assess the math skills that are critical for success in grade-level math content. LevelUp Math was designed by a team of educators and subject matter experts, and it is a computerized adaptive test (CAT) comprised of questions that adjust to each student's ability. Difficulty increases as students respond correctly and decreases as they respond incorrectly until a stable estimate of ability is obtained. By adapting to each student's ability, a CAT can substantially reduce testing time, increase precision, minimize test-taker fatigue, enhance the testing experience, support repeated testing, and improve test security, among other benefits.

The goal of the present study was to assess the validity of LevelUp Math, including construct validity and measurement invariance across student subgroups , and criterion validity using the NWEA MAP Growth assessment (MAP Growth). We aimed to answer the following research questions:

1. **Internal structure.** Does LevelUp Math exhibit a coherent internal structure as a measure of math achievement?
2. **Measurement invariance.** Does LevelUp Math demonstrate measurement invariance across student subgroups, including gender, race, ethnicity, English learner status, and special education status?
3. **Criterion validity.** Are there strong positive correlations between LevelUp Math and MAP Growth math performance?

Study Design and Methodology

_____

DATA SOURCES

IXL LevelUp Diagnostic for Math

We obtained 2024-25 beginning-of-year (August 2024 to November 2024) and middle-of-year (December 2024 and February 2025) LevelUp Math data from IXL's database. The math content on LevelUp Math is organized into strands that align with domains used in many states' standards across grades, as shown in Figure 1.

Grade	PK	K	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
Number & Operations in Base Ten*	NBT	NBT	NBT	NBT	NBT	NBT	NBT							
Operations & Algebraic Thinking	OAT	OAT	OAT	OAT	OAT	OAT	OAT							
Measurement & Data	MD	MD	MD	MD	MD	MD	MD							
Geometry	G	G	G	G	G	G	G	G	G	G	G	G	G	G
Numbers & Operations - Fractions					NF	NF	NF							
The Number System								NS	NS	NS				
Expressions & Equations								EE	EE	EE				
Ratios & Proportional Relationships								RP	RP					
Statistics & Probability								SP	SP	SP	S	S	S	S
Functions										F	F	F	F	F
Algebra											A	A	A	A
Number & Quantity											N	N	N	N

*Number and Operations in Base Ten also includes content from Kindergarten Counting and Cardinality.

Figure 1. Grade-level strand groupings

LevelUp Math is a variable-length CAT that selects questions based on a student's responses to generate a precise estimate of ability and ends the session once a sufficiently precise estimate is achieved. Upon completion, teachers and students receive an easily interpretable score that is strongly tied to expected performance on grade-level content. For example, a score of 400 indicates readiness for beginning-of-4th-grade content, while a score of 450 indicates readiness for middle-of-4th-grade content. While these scores provide an intuitive interpretation for educators, the present study relied on item response theory (IRT) theta scores, which offer a continuous and psychometrically robust measure for psychometric analysis.

Criterion Assessment and Demographic Data

To validate LevelUp Math with MAP Growth, we obtained 2024–25 middle-of-year (December 2024 to February 2025) MAP Growth scores and students' demographic data from IXL's database. On IXL, educators and students have the option to upload scores from third-party assessments, such as MAP Growth, into the StudyWise tool to create personalized study plans with an exact list of skills that will help students improve their scores in each subject. Educators can monitor student progress between uploads using the IXL LevelUp Diagnostic in Real-Time mode.

ANALYTIC APPROACH

Internal Structure

We performed confirmatory factor analyses (CFAs) to assess the use of a unidimensional scale to measure math ability. CFA is a statistical method for determining the appropriateness of a theoretical model specifying the relationships between observed variables (e.g., math strands) and a latent, unobserved variable (e.g., general math ability; Brown, 2015; Kline, 2016). Based on the groupings of shared math strands across certain grade levels (see Figure 1), grades were grouped into four bands (i.e., grades K–2, 3–5, 6–7, and 8) with CFAs conducted separately for each grade band. The analysis used strand theta values collected during the early portion of the 2024-25 academic year (August to November). In total, we analyzed data from 383,715 students in Kindergarten through 8th grade.

We examined several indices of global model-data fit as well as local misfit (i.e., residuals) to evaluate how well the theoretical unidimensional CFA model represents the data. First, we examined χ^2, which compares the difference between a specified model and the variances and covariances inherent to the data. Ideally, we aim for a statistically nonsignificant χ^2 value, indicating no significant difference between the theoretical model and the characteristics of the data. However, χ^2 is notoriously sensitive to sample size, where larger samples almost inevitably result in a statistically significant χ^2. Next, we examined the comparative fit index (CFI), where values less than 0.90 are considered unacceptable, values greater than 0.90 but less than 0.95 are considered acceptable, and

values greater than 0.95 are considered excellent (Hu & Bentler, 1999). We also examined the root mean squared error of approximation (RMSEA), where values greater than 0.08 are considered unacceptable, values less than 0.08 but greater than 0.06 are acceptable, and values less than 0.06 are considered excellent (Hu & Bentler, 1999). Then we examined the standardized root mean square residual (SRMR), where values greater than 0.10 are unacceptable, values less than 0.10 but greater than 0.08 are considered acceptable, and values less than 0.08 are considered excellent (Hu & Bentler, 1999). Lastly, we examined the standardized covariance residuals to detect areas of misfit, where correlation coefficients produced by the model are discrepant from the correlation coefficients observed in the data. Correlation residuals greater than |0.10| typically indicate areas of local misfit. (Brown, 2015). For internal consistency, we calculated McDonald's omega.

Measurement Invariance

For LevelUp Math, we evaluated measurement invariance across several student subgroups: gender, race, ethnicity, English learner status, and special education status. Specifically, for each demographic variable, we built the unidimensional CFA model for each student subgroup and then tested the combined model in three steps: 1) configural invariance, 2) metric invariance, and 3) scalar invariance. This multistep approach assessed a series of models ranging from less constrained to more constrained, and examined the (relative) model fit using changes in the comparative fit index (i.e., ΔCFI; Cheung & Rensvold, 2002). If adding constraints did not make the model fit significantly worse than the fit of the less constrained model (i.e., ΔCFI > − .01), invariance at the level of the more constrained model was established.

Across the subgroups studied, configural invariance models tested whether the dimensionality of LevelUp Math was the same across groups. Specifically, we first fitted a unidimensional CFA model to the data from each student subgroup separately and then simultaneously. Using gender as an example, we examined the one-factor structure for girls and boys separately, and then jointly using a single covariance matrix representing both groups. We then checked the absolute model fit indices to determine whether configural invariance could be established based on model-data fit, and if so, this model would then serve as the baseline model for testing metric invariance.

Metric invariance models assessed whether each IXL LevelUp Diagnostic math strand was an equally salient indicator of math achievement across groups. In other words, we tested whether corresponding strands' unstandardized factor loadings were equal across student groups. Using gender as an example again, the loadings from math proficiency to the *Geometry* strand were constrained to be equal for boys and girls, and so were the corresponding pairs of loadings of the other strands. We then checked the (relative) model fit using ΔCFI to determine whether metric invariance could be established; if yes, then this model would serve as the baseline model for scalar invariance testing.

Scalar invariance models examined whether each expected IXL LevelUp Diagnostic strand score was of the same magnitude across groups. We tested scalar invariance by constraining all the corresponding intercepts of the observed variables (i.e., math strands) to be equal across student groups. Using gender as an example again, the intercepts of the *Geometry* strand were constrained to be equal for boys and girls, and so were the intercepts of the other strands. We then checked the (relative) model fit using ΔCFI to determine if scalar invariance could be established.

Criterion Validity

Criterion validity is a type of validity that shows how well a test or measurement predicts or correlates with an external criterion or outcome. Concurrent validity is supported by high correlations between measures administered closer together in time without much regard to temporal order, whereas predictive validity is supported when the predictive assessment precedes the criterion measure and the two may be administered farther apart.

For both concurrent and predictive validity, we examined the correlation coefficients between the math scores from LevelUp Math and those from MAP Growth, using the Pearson product-moment correlation coefficient r. This coefficient measures the linear relationship between two variables and ranges from -1.00 to 1.00. An r value of 0 indicates no correlation, whereas an r value greater than 0.60 indicates a strong positive relationship (Ratner, 2009). High positive correlations between LevelUp Math and the criterion math assessment would support the validity of the IXL LevelUp Diagnostic.

For concurrent validity, we retrieved data from December 2024 to February 2025 and examined how well LevelUp Math scores correlate with MAP Growth scores within a 30-day window. For predictive validity, we examined how well LevelUp Math scores from the early portion of the 2024-25 academic year (August to November) correlate with scores from MAP Growth administered at least 90 days later (i.e., between December 2024 and February 2025).

Results

_____

INTERNAL STRUCTURE

Table 1 shows the results of the CFAs by grade band. All CFI values were greater than 0.95, indicating excellent fit. All RMSEA values were less than 0.08, indicating acceptable to excellent fit. All SRMR values were below 0.08, also indicating excellent fit. Lastly, there were no correlation residuals greater than|.10|, indicating no localized areas of misfit. As shown in Table 1, most factor loadings across grade bands and strands exceeded .70, with the *Geometry* strand in 3rd through 5th grade showing the lowest loading, slightly below .70. In terms of reliability, McDonald's omega was high,

exceeding .80 across all grade bands. These CFA results support the use of a unidimensional Rasch model with LevelUp Math in calibration and scoring.

Table 1. Confirmatory Factor Analysis Results

Grade band	K-2	3-5	6-7	8
N	65,843	108,894	136,089	72,889
Model fit indices				
χ^2	636.556	204.754	592.909	783.92
df	2	5	5	5
CFI	0.996	0.999	0.999	0.996
RMSEA	0.069	0.019	0.029	0.046
SRMR	0.010	0.005	0.005	0.009
Correlation residuals < |.10|	10 / 10	15 / 15	15 / 15	15 / 15
Factor Loadings				
Geometry	0.759	0.661	0.741	0.844
Number & Operations in Base Ten	0.865	0.751	-	-
Number & Operations Fractions	-	0.731	-	-
Operations & Algebraic Thinking	0.817	0.801	-	-
Measurement & Data	0.767	0.775	-	-
The Number System	-	-	0.859	0.811
Expressions & Equations	-	-	0.839	0.902
Ratios & Proportional Relationships	-	-	0.867	-
Statistics & Probability	-	-	0.795	0.800
Functions	-	-	-	0.753
Reliability				
McDonald's omega	0.885	0.862	0.912	0.913

Note. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean squared error of approximation; SRMR = standardized root mean square residual

MEASUREMENT INVARIANCE

Gender

Table 2 summarizes the sample sizes by gender and grade band. Table 3 reports the results of measurement invariance across gender by grade band. Configural, metric, and scalar invariance

were established for all grade bands. Model fit indices were excellent in the configural invariance stage and did not become significantly worse in the metric or scalar invariance stages. Although not shown in the table, all factor loadings were statistically significant. In summary, the results provide clear support for the measurement invariance of LevelUp Math across gender.

Table 2. Sample Sizes by Gender and Grade Band

Grade	Boys	Girls
K-Grade 2	22,051	21,019
Grades 3-5	36,568	38,472
Grades 6-7	53,057	50,154
Grade 8	25,751	24,726

Table 3. Fit Statistics for the Models Testing Measurement Invariance Across Gender

Grade	Model	χ^2	*df*	*p*	CFI	RMSEA	SRMR	ΔCFI
K-Grade 2	Configural invariance	235.40	4	.000	.996	.052	.011	-
	Metric invariance	297.02	7	.000	.995	.044	.017	-0.001
	Scalar invariance	627.80	10	.000	.990	.054	.023	-0.005
Grades 3-5	Configural invariance	159.63	10	.000	.998	.020	.006	-
	Metric invariance	342.77	14	.000	.996	.025	.015	-0.002
	Scalar invariance	526.16	18	.000	.994	.027	.017	-0.002
Grades 6-7	Configural invariance	150.80	10	.000	.999	.017	.004	-
	Metric invariance	186.28	14	.000	.999	.015	.007	0.000
	Scalar invariance	596.48	18	.000	.997	.025	.011	-0.002
Grade 8	Configural invariance	558.81	10	.000	.995	.047	.011	-
	Metric invariance	592.41	14	.000	.995	.040	.014	0.000
	Scalar invariance	639.93	18	.000	.994	.037	.015	0.000

Race

Table 4 summarizes the sample sizes by race (i.e., White or non-White) and grade band. Table 5 reports the results of measurement invariance across racial groups by grade band. Configural, metric, and scalar invariance were established for all grade bands, with excellent model fit indices observed in the configural invariance stage and no significantly worse model fit in the metric and scalar invariance stages. All factor loadings were strong and statistically significant. In summary, the results provide clear support for the measurement invariance of LevelUp Math across racial groups.

Table 4. Sample Sizes by Race and Grade Band

Grade	White	Non-White
K-Grade 2	22,669	10,571
Grades 3-5	42,893	17,028
Grades 6-7	59,666	26,416
Grade 8	29,679	12,559

Table 5. Fit Statistics for the Models Testing Measurement Invariance Across Racial Groups

Grade	Model	χ2	df	p	CFI	RMSEA	SRMR	ΔCFI
	Configural invariance	160.72	4	.000	.996	.049	.010	-
K-Grade 2	Metric invariance	197.26	7	.000	.996	.040	.015	-0.001
	Scalar invariance	268.50	10	.000	.994	.039	.017	-0.002
	Configural invariance	117.75	10	.000	.998	.019	.006	-
Grades 3-5	Metric invariance	200.62	14	.000	.997	.021	.012	-0.001
	Scalar invariance	452.62	18	.000	.993	.028	.016	-0.004
	Configural invariance	134.06	10	.000	.999	.017	.004	-
Grades 6-7	Metric invariance	260.33	14	.000	.999	.020	.011	-0.001
	Scalar invariance	601.78	18	.000	.997	.027	.014	-0.002
	Configural invariance	422.14	10	.000	.995	.044	.011	-
Grade 8	Metric invariance	499.25	14	.000	.995	.041	.017	-0.001
	Scalar invariance	899.75	18	.000	.990	.048	.021	-0.004

Ethnicity

Table 6 summarizes the sample sizes by ethnicity (i.e., Hispanic or non-Hiapanic) and grade band. Table 7 reports the results of measurement invariance across ethnic groups by grade band. Configural, metric, and scalar invariance across ethnicity were established for all grade bands, with excellent model fit indices observed in the configural invariance stage and no significantly worse model fit in the metric and scalar invariance stages. All factor loadings were strong and statistically significant. In summary, the results provide clear support for the measurement invariance of LevelUp Math across ethnic groups.

Table 6. Sample Sizes by Ethnicity and Grade Band

Grade	Hispanic	Non-Hispanic
K-Grade 2	5,168	24,929
Grades 3-5	8,695	48,108
Grades 6-7	16,307	65,523
Grade 8	8,765	31,780

Table 7. Fit Statistics and Factor Loadings for the Models Testing Measurement Invariance Across Ethnic Groups

Grade	Model	χ^2	df	p	CFI	RMSEA	SRMR	ΔCFI
K-Grade 2	Configural invariance	165.36	4	.000	.996	.052	.011	-
	Metric invariance	195.19	7	.000	.995	.042	.014	-0.001
	Scalar invariance	298.28	10	.000	.993	.044	.016	-0.002
Grades 3-5	Configural invariance	117.64	10	.000	.998	.019	.006	-
	Metric invariance	183.18	14	.000	.997	.021	.010	-0.001
	Scalar invariance	305.16	18	.000	.995	.024	.012	-0.002
Grades 6-7	Configural invariance	120.68	10	.000	.999	.016	.004	-
	Metric invariance	200.76	14	.000	.999	.018	.008	0.000
	Scalar invariance	392.29	18	.000	.998	.023	.011	-0.001
Grade 8	Configural invariance	442.93	10	.000	.995	.046	.012	-
	Metric invariance	452.58	14	.000	.995	.039	.012	0.000
	Scalar invariance	489.64	18	.000	.994	.036	.013	0.000

English Learner (EL) Status

Table 8 summarizes the sample sizes by English learner status and grade band. Table 9 reports the results of measurement invariance across English learner status by grade band. Except for Kindergarten through 2nd grade, configural, metric, and scalar invariance were established for all other grade bands, with excellent model fit indices observed in the configural invariance stage and no significantly worse model fit in the metric and scalar invariance stages. For Kindergarten through 2nd grade, configural invariance was established with acceptable model fit, but only metric invariance was then established with no significantly worse model fit. It's worth noting that the sample size of EL for this grade band was quite small, which likely influenced the results. All factor loadings were strong and statistically significant. In general, the results provide support for the measurement invariance of LevelUp Math across English learner groups.

Table 8. Sample Sizes by English Learner Status and Grade Band

Grade	EL	Non-EL
K-Grade 2	61	975
Grades 3-5	98	2,424
Grades 6-7	1,652	13,598
Grade 8	826	6,333

Table 9. Fit Statistics for the Models Testing Measurement Invariance Across English Learner Status

Grade	Model	χ^2	df	p	CFI	RMSEA	SRMR	ΔCFI
K-Grade 2	Configural invariance	12.94	4	.012	.992	.066	.017	-
	Metric invariance	24.39	7	.001	.985	.069	.024	-0.007
	Scalar invariance	42.60	10	.000	.972	.079	.029	-0.013
Grades 3-5	Configural invariance	8.69	10	.562	1.000	.000	.007	-
	Metric invariance	28.97	14	.011	.996	.029	.011	-0.004
	Scalar invariance	37.05	18	.005	.994	.029	.013	-0.001
Grades 6-7	Configural invariance	26.12	10	.004	1.000	.015	.004	-
	Metric invariance	64.28	14	.000	.998	.022	.010	-0.001
	Scalar invariance	141.34	18	.000	.996	.030	.013	-0.002
Grade 8	Configural invariance	84.24	10	.000	.995	.046	.011	-
	Metric invariance	101.52	14	.000	.994	.042	.017	-0.001
	Scalar invariance	164.09	18	.000	.990	.048	.019	-0.004

Special Education Status

Table 10 summarizes the sample sizes by special education status (i.e., whether or not a student is eligible for an Individualized Education Program [IEP]) and grade band. Table 11 reports the results of measurement invariance across special education groups by grade band. Configural, metric, and scalar invariance were established for all grade bands, with excellent model fit indices observed in the configural invariance stage and no significantly worse model fit in the metric and scalar invariance stages. All factor loadings were positive and statistically significant. In summary, the results provide clear support for the measurement invariance of LevelUp Math across special education status.

Table 10. Sample Sizes by Special Education Status and Grade Band

Grade	Receiving IEP	Without IEP
K-Grade 2	265	940
Grades 3-5	668	1,821
Grades 6-7	2,399	12,685
Grade 8	1,123	6,122

Table 11. Fit Statistics for the Models Testing Measurement Invariance Across Special Education Status

Grade	Model	χ2	df	p	CFI	RMSEA	SRMR	ΔCFI
	Configural invariance	10.96	4	.027	.995	.054	.014	-
K-Grade 2	Metric invariance	25.83	7	.001	.987	.067	.038	-0.008
	Scalar invariance	37.08	10	.000	.981	.067	.043	-0.006
	Configural invariance	17.32	10	.068	.998	.024	.011	-
Grades 3-5	Metric invariance	20.05	14	.129	.998	.019	.014	0.000
	Scalar invariance	34.90	18	.010	.994	.027	.020	-0.004
	Configural invariance	31.55	10	.000	.999	.017	.005	-
Grades 6-7	Metric invariance	67.75	14	.000	.998	.023	.011	-0.001
	Scalar invariance	97.13	18	.000	.997	.024	.013	-0.001
	Configural invariance	84.74	10	.000	.995	.045	.012	-
Grade 8	Metric invariance	86.31	14	.000	.995	.038	.013	0.000
	Scalar invariance	99.63	18	.000	.994	.035	.014	-0.001

CRITERION VALIDITY

Concurrent Validity

We found strong positive correlations between IXL LevelUp Diagnostic math scores and concurrent MAP Growth math scores in Kindergarten through 8th grade (rs > .60, ps < .001; Table 12).

Table 12. Concurrent Validity for LevelUp Math by Grade Using MAP Growth as Criterion

Grade	N	r	p
Kindergarten	103	.69	<.001
Grade 1	279	.82	<.001
Grade 2	492	.81	<.001
Grade 3	500	.78	<.001
Grade 4	470	.82	<.001
Grade 5	1,238	.82	<.001
Grade 6	856	.79	<.001
Grade 7	801	.75	<.001
Grade 8	640	.75	<.001

Predictive Validity

We found strong positive correlations between IXL LevelUp Diagnostic math scores and subsequent MAP Growth math scores in Kindergarten through 8th grade (rs > .60, ps < .001; Table 13).

Table 13. Predictive Validity for LevelUp Math by Grade Using MAP Growth as Criterion

Grade	N	r	p
Kindergarten	174	.62	<.001
Grade 1	877	.69	<.001
Grade 2	1,560	.72	<.001
Grade 3	1,475	.70	<.001
Grade 4	1,459	.81	<.001
Grade 5	2,308	.80	<.001
Grade 6	2,038	.84	<.001
Grade 7	1,752	.84	<.001
Grade 8	1,515	.80	<.001

Conclusion

_____

The present study provides preliminary evidence supporting the validity of the IXL LevelUp Diagnostic for math as an efficient and reliable measure of student math achievement from Pre-K through 8th grade. Our analyses indicate that LevelUp Math demonstrates a coherent internal structure, with factor loadings reflecting its effectiveness as a comprehensive tool for evaluating grade-level math achievement. Furthermore, the Diagnostic shows strong measurement invariance across key student subgroups, including gender, race, ethnicity, English learner status, and special education status, suggesting that the assessment provides fair measurement across diverse populations. Finally, LevelUp Math scores are strongly and positively correlated with both concurrent and subsequent MAP Growth performance in math, supporting the assessment's criterion validity and confirming its effectiveness in measuring student progress toward grade-level standards.

Taken together, these findings demonstrate that LevelUp Math is a valid and effective assessment, providing meaningful and equitable insights into students' math abilities. By accurately reflecting progress toward grade-level standards, it enables educators to make informed, data-driven decisions, identify areas for growth, and tailor instruction to support all learners. In this way, LevelUp Math helps promote both individual student achievement and overall learning outcomes.

References

_____

An, X. & Schonberg, C. (2024). *The impact of IXL on students' math self-efficacy* (pp. 1–13). https://www.ixl.com/materials/us/research/The_Impact_of_IXL_on_Students_Math_Self-Efficacy.pdf

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. S*tructural Equation Modeling: A Multidisciplinary Journal, 9,* 233–255.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Ratner, B. (2009). The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing, 17(*2), 139–142. https://doi.org/10.1057/jt.2009.5