



2024 No. 149

## Third-Party Checking of 2024 Scaling and Equating for the Alternate Kentucky Summative Assessments (AKSA)

**Prepared for:** Kentucky Department of Education  
Office of Assessment and Accountability  
300 Sower Boulevard  
Frankfort, KY 40601

**Authors:** Gertrudes Velasquez  
Jiayi Deng  
Meng Fan  
Emily R. Dickinson  
Art A. Thacker

**Prepared under:** Contract #2400003213

**Date:** September 30, 2024

# Third-Party Checking of 2024 Scaling and Equating for the Alternate Kentucky Summative Assessments (AKSA)

## Table of Contents

Executive Summary .....	ii
Introduction .....	1
Brief Description of 2024 Assessments .....	1
Analysis Procedures .....	2
Sample Identification and File Construction .....	2
Calibration and Scaling Procedures .....	2
Equating Procedures.....	2
Raw-score-to-Scale-Score Procedures .....	3
Verification of 2024 Scoring Tables .....	3
Documentation .....	3
Conclusion .....	4
References .....	5
Appendix A – Sample Student Response Data (Math Grade 4) .....	6
Appendix B – Item Parameters (Math Grade 4).....	7
Appendix C – Raw-Score-to-Scale Score (Math Grade 4).....	8

## Executive Summary

Education Measurement Consulting and the Human Resources Research Organization (HumRRO) independently calibrated, scaled, and equated the 2024 Alternate Kentucky Summative Assessments (AKSA) and produced the raw-score-to-theta-score tables that were applied to students' test results. Results calculated by HumRRO were identical to those calculated by Education Measurement Consulting. Given that HumRRO's results were identical to those of Education Measurement Consulting, we are assured that no processing errors were committed.

# Third-Party Checking of 2024 Scaling and Equating for the Alternate Kentucky Summative Assessments (AKSA) Tests

## Introduction

The Alternate Kentucky Summative Assessments (AKSA) assesses student performance on the Alternate Assessment Targets. Alternate Assessment Targets represent limits to selected Kentucky Academic Standards (KAS). The AKSA is administered to students with moderate and significant disabilities (i.e., less than 1% of the total student population for whom traditional assessments would not provide a reliable and valid measure of progress). The Kentucky Department of Education (KDE) contracted with the Human Resources Research Organization (HumRRO) to provide independent, external replication of the psychometric processing activities for both the AKSA and the Kentucky Summative Assessment (KSA).<sup>1</sup>

In spring 2024, the AKSA was administered in reading, mathematics, science, social studies, editing and mechanics, and on-demand writing. Reading and mathematics were assessed in Grades 3 through 8 and Grade 10; science in Grades 4, 7, and 11; social studies, editing and mechanics, and on-demand writing in Grades 5, 8, and 11.

This report describes how student test responses on the 2024 AKSA were used to create scale scores and place students in Novice, Apprentice, Proficient, or Distinguished (NAPD) performance categories. The complex analyses to accomplish these tasks were conducted independently and cooperatively by both HumRRO and Education Measurement Consulting. Several interim checks were conducted during the analyses, and any discrepancies between the two companies were investigated and ultimately resolved. This process was conducted transparently among the University of Kentucky (the test vendor), Education Measurement Consulting, HumRRO, and the Kentucky Department of Education (KDE) via frequent email communications and conference calls. The process was guided by a specifications document created by Education Measurement Consulting and regularly updated based on decisions before and during calibration. This documentation is vital for ensuring consistency of processing across years and for guiding psychometric processing in future years.

## Brief Description of 2024 Assessments

The AKSA consists of picture-based, 3-option multiple-choice assessments that are fully scripted and read by a qualified test administrator. It is administered over two testing windows. For each content area, 15 items are administered in fall, and 15 items are administered in spring.

In the 2022-2023 academic year, the AKSA transitioned from a raw score to a scale score reporting model. An equating procedure was implemented to ensure that scores were comparable to those from the prior year. The 2023-2024 scores are reported on a scale of 150-250, based on the established score reporting model.

---

<sup>1</sup> The AKSA is composed of two tests. The Attainment Task parallels the KSA, whereas the Transition Attainment Record parallels the ACT. Accordingly, references to the AKSA in this report focus on the Attainment Task component only.

## Analysis Procedures

Item parameters were generated (i.e., calibrated) separately for each grade/subject combination. A common-item equating procedure was then used to put students' ability estimates onto the 2022-2023 scale. Finally, raw-score-to-scale (RSS) score tables were generated for all assessments. For each of these analyses, HumRRO followed the analysis specifications provided by Education Measurement Consulting, independently conducted analyses, and verified that the two sets of results matched. Calibration and equating procedures were conducted using the *mirt* (v1.4.1; Chalmers, 2012) package in R (R Core Team, 2024) statistical software. We summarize HumRRO's processes and procedures for conducting these analyses below.

### Sample Identification and File Construction

We received data with already applied exclusion rules to select the sample of student responses in the calibration analyses.<sup>2</sup> Kentucky selects most of its student population for use in the calibration sample for scaling and equating. However, as specified in the 2024 psychometrics specifications document, some students are purposefully excluded from the calibration samples.<sup>3</sup> Kentucky's exemption rules generally only apply to students who receive accommodations (e.g., Braille forms, audio, large print, etc.), students with duplicate records (the same identification number and name), and students with blank total test score values. HumRRO removed students with all missing responses. Education Measurement Consulting and HumRRO verified the n-counts for the calibration samples.

The next step was to format all subject/grade files to be read into R software for processing. A simulated raw response dataset for 15 Math Grade 4 items for six cases is provided in Appendix A as an example of typical response data. HumRRO and Education Measurement Consulting used the same raw student data files (containing all student responses) but did not share programming or methodology for creating the scored data files. Prior to item calibration, item responses were scored into correct/incorrect responses using response keys.

### Calibration and Scaling Procedures

Once scored data files were prepared, dichotomously-scored items were calibrated using the Rasch model (Rasch, 1960). This step produced item difficulty parameters. These parameters are produced on the student ability (theta) scale (a commonly used scale with a mean of 0 and a standard deviation of 1). Appendix B contains an example of item parameters for one subject-grade. Education Measurement Consulting and HumRRO verified item parameter estimates after this step.

## Equating Procedures

For each assessment, we also placed student ability estimates on the 2022 reporting scale. We implemented a common-item equating procedure to accomplish this. The anchor items used in the equating procedure were items common to the 2023 and 2024 AKSA administrations. The anchor item parameters are those from the 2023 calibration.

---

<sup>2</sup> Students who are excluded from calibration analyses are not excluded from scoring and reporting.

<sup>3</sup> Kentucky Spring 2024 Psychometric Analysis Specifications v1.0.

The following steps were applied for the calibration and equating procedures.

1. Data Exclusion: Remove any cases with missing responses to ensure the accuracy of the calibration process.
2. Item Calibration: Use the Rasch model and the anchor item parameters (after conducting an item stability check) to calibrate 2024 items.
3. Item Stability Check: Flag anchor items in the 2024 calibration for removal if they had negative item slopes (item difficulty parameters,  $a < -0.01$ ) or exhibited parameter drift based on Robust Z statistics (Huynh, 2000; Huynh & Meyer, 2010; Huynh & Rawls, 2009).
4. Theta Value Estimation: Use the EAPsum method to estimate the 2024 theta values, aiming for a single theta value for each raw score.
5. Application of Cut Scores: Implement the established cuts from 2022 onto the 2024 results to categorize or interpret the 2024 outcomes based on the previously set criteria.

### Raw-score-to-Scale-Score Procedures

The item parameter estimates from the 2024 calibration were used to generate person ability estimates. Ability estimates were reported in raw-score-to-theta-score tables.

Once theta scoring tables were obtained, they were linearly transformed to a reporting scale of 150-250 for all grades/subjects. The scale score transformation equation is given by

$$SS = 20 * \theta + 200,$$

where SS represents students' scale scores and  $\theta$  represents students' ability estimates. The scale score slope and intercept transformation constants are 20 and 200, respectively.

### Verification of 2024 Scoring Tables

After verification of the raw-score-to-scale score tables, scoring tables were generated to assign student performance level classifications (Novice, Apprentice, Proficient, and Distinguished; NAPD) based on cut points established in the 2022 reporting scale. HumRRO checked the 2024 scoring tables and verified that the correct scale score ranges were associated with each performance level. HumRRO matched Education Measurement Consulting on all subjects and grades.

### Documentation

As HumRRO and Education Measurement Consulting completed each step of the process described above, item parameters, anchor items, scores, and other output files were shared to check for inconsistencies. Output files included the number of cases in the calibration sample, item-level information (i.e., item parameters), and the theta scoring tables. Samples of input and output files are appended to this document. They include:

1. Student Response Data contains students' raw responses on each AKSA subject and grade. Appendix A provides simulated raw responses for 15 Math Grade 4 items for six cases as an example of typical response data.
2. Item Parameter files contain operational item parameters. The file in Appendix B is an example of an item parameter file for five items for Math Grade 4.



3. Raw-Score-to-Scale Score and Performance Level files include raw-score-to-scale score tables and corresponding performance levels. The file included in Appendix C is an example of a raw-score-to-scale score table for Math Grade 4.

### Conclusion

Education Measurement Consulting and HumRRO independently calculated the scaled/equated raw-score-to-scale-score tables for 2024 AKSA reading and math (Grades 3-8, 10), social studies and writing (Grades 5, 8, 11), and science (Grades 4, 7, 11). No differences were found between the two sets of parameter estimations or raw-score-to-scale-score tables. Given that HumRRO's and Education Measurement Consulting's scaling and equating results were identical, HumRRO is confident that no processing errors were committed.

## References

- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.  
<https://doi.org/10.18637/jss.v048.i06>.
- Huynh, H. (2000, June). *Guidelines for Rasch Linking for PACT*. Memorandum to Paul Sandifer on June 18, 2000. Available from Author.
- Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15(2).  
<http://pareonline.net/getvn.asp?v=15&n=2>.
- Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting*. JAM Press.
- R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.



## Appendix A – Sample Student Response Data (Math Grade 4)

Grade	Math A1	Math A2	Math A3	Math A4	Math A5	Math B1	Math B2	Math B3	Math B4	Math B5	Math C1	Math C2	Math C3	Math C4	Math C5
4 c	b	c	c	a	b	a	c	a	c	b	c	a	a	c	
4 a	b	b	c	c	c	b	c	a	c	a	a	b	a	b	
4 c	a	a	a	c	b	b	c	b	c	b	a	c	b	b	
4 a	b	a	c	b	a	b	c	a	a	c	b	c	a	c	
4 c	b	b	c	c	b	b	a	b	a	c	b	a	c	c	
4 a	c	b	c	b	c	b	b	a	c	c	a	c	b	c	

## Appendix B – Item Parameters (Math Grade 4)

item	class	name	parnum	value
A1	dich	a1	1	1.00000000
A1	dich	d	2	-0.47242900
A1	dich	g	3	0.00000000
A1	dich	u	4	1.00000000
A2	dich	a1	5	1.00000000
A2	dich	d	6	-1.17741139
A2	dich	g	7	0.00000000
A2	dich	u	8	1.00000000
A3	dich	a1	9	1.00000000
A3	dich	d	10	-0.16207711
A3	dich	g	11	0.00000000
A3	dich	u	12	1.00000000
A4	dich	a1	13	1.00000000
A4	dich	d	14	0.07284874
A4	dich	g	15	0.00000000
A4	dich	u	16	1.00000000
A5	dich	a1	17	1.00000000
A5	dich	d	18	-0.39346208
A5	dich	g	19	0.00000000
A5	dich	u	20	1.00000000

## Appendix C – Raw-Score-to-Scale Score (Math Grade 4)

RawScore	Theta	ScaleScore	PL
0	-1.2688	175	N
1	-1.1443	177	N
2	-1.0244	180	N
3	-0.9087	182	N
4	-0.7967	184	N
5	-0.6880	186	N
6	-0.5824	188	N
7	-0.4794	190	N
8	-0.3788	192	N
9	-0.2802	194	A
10	-0.1834	196	A
11	-0.0881	198	A
12	0.0060	200	A
13	0.0991	202	A
14	0.1913	204	A
15	0.2830	206	P
16	0.3743	207	P
17	0.4655	209	P
18	0.5568	211	P
19	0.6483	213	P
20	0.7403	215	P
21	0.8330	217	P
22	0.9267	219	D
23	1.0215	220	D
24	1.1177	222	D
25	1.2155	224	D
26	1.3153	226	D
27	1.4173	228	D
28	1.5218	230	D
29	1.6291	233	D
30	1.7396	235	D