HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2020 No. 009

Third-Party Checking of 2019 Scaling and Equating for the Kentucky Performance Rating for Educational Progress (K-PREP) Tests

Final Report

Prepared for:	Kentucky Department of Education Office of Assessment and Accountability 300 Sower Boulevard Frankfort, KY 40601	Prepared under:	Contract #1900004339
Authors:	Bethany H. Bynum Arthur A. Thacker	Date:	January 10, 2020

Third-Party Checking of 2019 Scaling and Equating for the Kentucky Performance Rating for Educational Progress (K-PREP) Tests

Table of Contents

Executive Summary

Pearson and HumRRO independently calibrated, scaled and equated the 2019 Kentucky Performance Rating for Educational Progress (K-PREP) assessments and produced the raw-score-to-theta-score tables to be applied to students' test results. HumRRO further verified that scoring table were applied accurately by checking our scoring of the student sample against Pearson's. Results calculated by HumRRO were identical to those calculated by Pearson (M. Johnson, email communication, July 22, 2019 [Writing] and July 30, 2019 [Science]. Given that HumRRO's results were identical to those of Pearson, we are assured that Pearson did not commit processing errors.

Third-Party Checking of 2019 Scaling and Equating for the Kentucky Performance Rating for Educational Progress (K-PREP) Tests

Introduction

In 2012, Kentucky transitioned from the Kentucky Core Content Test (KCCT) to the K-PREP system for spring testing. This transition represented a significant departure from the prior assessment system. The 3-parameter logistic Item-Response Theory (IRT) model was replaced with a Rasch model, a new item-type (i.e., short-constructed-response) was added to the assessments, a new scale-score reporting system was developed for sub-scores, and new cut scores were identified for the reading and mathematics assessments. The transition was also accompanied by a new primary testing contractor, Pearson. As a result, HumRRO's third-party checking process underwent significant changes to accommodate the transition.[1]

Equating was added to the process in 2013 to permit comparison of the results across test years. The 2014 tests were equated to the 2013 tests using linking items. In this manner, comparable scores were produced for the 2014 and 2015 K-PREP. Forms for all subjects other than writing were repeated in 2016 from prior years, meaning that existing scoring tables could be used, and no equating was necessary. In 2017, new forms were constructed, and equating analyses were carried out for Reading and Mathematics in all grades. Beginning in 2015, scale scores were computed for the On-Demand Writing tests where simple number correct scores had been used in the past.[2] Writing tests were equated in 2016 and 2017. In 2018 we equated new Reading and Mathematics test forms and verified scoring tables for Social Studies and Writing forms that were repeated from 2016. We also calibrated items from the first operational administration of the new Science assessments and generated raw-to-theta conversion tables for use during standard setting.

This report describes how student test responses for the 2019 K-PREP assessments 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, but cooperatively, by both HumRRO and Pearson staff members. Several interim checks were conducted during the analyses and any discrepancies between the two companies was investigated and ultimately resolved. This process was conducted transparently among Pearson, HumRRO, KDE, and Kentucky's psychometric consultant (Dr. Bill Auty of Education Measurement Consulting) via frequent email communications and daily conference calls. The process was guided by a specifications document created by Pearson[3] and regularly updated based on decisions before and during calibration. This documentation is vital for ensuring consistency of processing across years and serves as a guiding document for subsequent years.

Changes in 2019

Beginning in spring 2019, the number of score points for each On Demand Writing (ODW) task was reduced from a total of eight possible points to a total of four possible points. Prior to 2019, each ODW task was given a score on a one to four scale by two independent scorers. Each

[1] For additional details on how the assessment system and third-party checking procedures changed, see Bynum and Thacker (2013).
[2] For additional information on how writing was calibrated and scaled, see K-PREP ODW Calibration and Scaling Specs v0.4.docx.
[3] Kentucky Spring 2019 Psychometric Analysis Specifications v1.3.docx.

student's ODW task score was then computed as the sum of the two independent scores (ranging from 0 to 8). Each student completed two ODW tasks resulting in an overall ODW score ranging from 0 to 16. In 2019, each ODW task was scored by only one scorer using the same one to four scoring rubric as previous administrations, resulting an overall ODW task score that ranged from 0-4. The two ODW tasks were combined to form a student's overall ODW score, which ranged from 0 to 8. Because of these changes, in 2019, we created new score tables for ODW based on the reduced score points.

In 2019, Grade 11 ODW moved from the paper-pencil format utilized since 2012 to an online format. The change in administration mode may have impacted the difficulty of the ODW tasks. To account for this possibility, we equated the 2019 item parameters to the 2016 scale. Additionally, in 2019, there were problems with accessing the online dictionary. About 2,000 students were not able to access the dictionary during the early days of the testing window. These students were removed from the calibration analyses.

The spring of 2019 was the first-year high school Science (grade 11) was administered operationally. As such, we estimated item parameter estimates and computed raw-score-to-scale score tables. Science grades 4 and 7 test forms were reused from a previous administration, so no calibration analyses were needed. However, we computed reporting category scoring tables for these grades and applied those to student scores.

All Reading and Mathematics test forms were reused from a previous administration year. Therefore, no calibration analyses were conducted for these subjects. The score tables were updated to divide the lower two performance level categories, Novice and Apprentice, into "low" and "high" breakouts.

Analysis Procedures

New item parameters were generated (i.e., calibrated) for high school Science, anchored item calibration analysis was conducted to compute new raw-score-to-scale score tables for Writing grades 5 and 8, and equating analyses were conducted for Writing grade 11. For each of these analyses, we followed the analysis specifications provided by Pearson, independently conducted analyses, and verified our results matched Pearson's results. Below we summarize HumRRO's processes and procedures for conducting these analyses.

Sample Identification and File Construction

We first applied exclusion rules to select the sample of student responses to include in the calibration analyses. Kentucky selects most of its student population for use in the calibration sample for scaling and equating. However, some students are purposefully exempted. KDE established a set of invalidation codes for excluding students in the calibration file. Kentucky's exemption rules only apply to students who receive accommodations (e.g., Braille forms, audio, large print, etc.) and students with duplicate records (the same identification number and name). The accommodated students receive scores but are simply omitted from the calibration sample. In addition to these exclusion rules, the students who were impacted by the Grade 11 writing dictionary issue were removed from the calibration analyses. Pearson and HumRRO verified n-counts after this step.

The next step was to format all the grade/subject files to be read into the Winsteps IRT program and create Winsteps[4] control files to read the student responses and estimate parameters. A sample control file is provided in Appendix A. HumRRO created specialized SAS programs to generate all input and control files automatically. The item documentation file was used to specify item types, location, keys, item use (e.g. field test vs. operational items), and other important information. HumRRO and Pearson did not share programming or methodology for creating the input, control or data files for Winsteps. However, both companies used the same raw student data files (containing all student responses). HumRRO followed the guidance provided by Pearson (with input from KDE) regarding the treatment of blank responses, condition codes, etc. in creating the input data files.

Calibration and Scaling Procedures

Once input and control files were prepared, Winsteps software was used to calibrate high school Science and grade 11 Writing items. Multiple-choice items were fit to the Rasch measurement model and constructed-response items (short constructed response and extended response items) were fit to the Partial Credit Model (PCM). Both types of items were simultaneously calibrated in Winsteps and item difficulty parameters (logits) were produced. "Step parameters" were also produced for constructed response items. Step parameters tell us how the various points possible on the item relate to the item's overall difficulty and are important for generating scoring tables. These parameters are produced on the 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 grade subject (logits and step parameters). Pearson and HumRRO verified item parameter estimates after this step.

Equating Procedures

Two types of equating occurred for the K-PREP: (a) forms equating within a given test administration year and (b) equating across test administration years using common anchor items. The first of these, forms equating, is accomplished by calibrating all the items for a given grade/subject together. By calibrating all the items together (i.e., across all forms), this effectively equates the various forms for a given grade/subject such that test scores on form 2 and form 3, for example, are interchangeable in terms of difficulty.

For grade 11 writing, we also needed to equate the current year's scores to be comparable to scores from prior years. To accomplish this, we computed a shift estimate taking into account the differences in parameter estimation between the 2016 administration, which was administered via paper and pencil, and the 2019 administration, which was administered online. The shift estimate was then applied to the 2019 item parameter estimates to put the estimates onto the 2016 scale. These estimates were then used to create the raw-score-to-scale-score tables. Since 2019 was the first year of operational administration for the new high school Science test, no year-to-year equating was conducted.

Raw-Score-to-Scale-Score Procedures

We conducted an anchor item calibration for Writing grades 5 and 8 using Winsteps software to produce raw-score-to-scale-score tables. Because the number of categories were reduced from 2016 to 2019, fewer step parameters were needed for each item. To reduce the number of step parameters, Pearson conducted an equating study using the 2016 data, where they calibrated

[4] HumRRO used Winsteps version 3-73-00 for this project.

item parameter independently for each of the two scorers. This resulted in two sets of step parameters (one for each scorer) for each item. These two sets were then averaged to obtain the reduced number of step parameters needed to score the 2019 data. We provided these item parameters to Winsteps as anchor values to compute the raw-score-to-scale score tables.

The item parameters estimates from the initial calibration for high school science and the item parameters estimates following the equating analyses for grade 11 writing were used to create scoring tables. We used Winsteps to estimate the raw-score-to-scale-score tables by providing the final item parameter estimates as anchor values.

For writing grades 5 and 8, we only conducted an anchor item calibration to produce raw-score-to-scale-score tables. Because the number of categories were reduced from 2016 to 2019, fewer step parameters were needed for each item. To reduce the number of step parameters, Pearson conducted an equating study using the 2016 data, where they calibrated item parameter independently for each of the two scorers. This resulted in two sets of step parameters (one for each scorer) for each item. These two sets were then averaged to obtain the reduced number of step parameters needed to score the 2019 data. These values were provided to HumRRO by Pearson and Humrro used these item parameters as anchor values to compute the raw-score-to-scale score tables using Winsteps.

Once theta scoring tables were obtained, they were linearly transformed to a reporting scale of 100-300 for all grade subjects. Performance levels (Novice, Apprentice, Proficient, and Distinguished; NAPD) were also assigned to each score. Cut scores for the performance levels were determined following a standard setting workshop conducted by Pearson. The results of that workshop included cut scores on the theta metric that can be used to assign NAPD categories to students. Scale score cuts were used, as opposed to theta cuts, to assign performance levels to students' scale scores. Using these cuts allowed the scale scores associated with each performance level to be fixed across test administrations. HumRRO verified the raw-score-to-scale-score tables and the associated performance levels. In 2019, the performance levels for Reading and Math were updated to include a "low" and "high" breakout for Novice and Apprentice. We added these breakouts to the existing raw-score-to-scale-score tables and verified our results against Pearson's.

In addition to overall scores, Kentucky also reports cluster scores (subscores based on subsets of items within each test). The generation of cluster scores uses the previously estimated item parameters and is accomplished by generating scoring tables in Winsteps on the theta metric, based on the specific items identified for each scoring cluster. These theta scores are then transformed in exactly the same manner as the full test scores.

HumRRO generated raw-score-to-scale score tables were compared to Pearson's raw-score-to-scale score tables for all grades and subjects. HumRRO matched Pearson's score tables for all grades and subjects.

Verification of 2019 Scoring Tables

After the final scoring tables were constructed, scoring tables were applied to the 2019 student data. For grades/subjects that repeated forms previously administered in 2016, existing scoring tables were verified. HumRRO checked the 2019 scored student data to verify that the scoring tables are being appropriately applied to the data and to check the distribution of students falling into each performance level. HumRRO verified Reading, Math, Social Studies, Science and

Writing performance level distributions. HumRRO matched Pearson on the number and percent of students assigned to each performance level by subject and grade

Documentation

As HumRRO and Pearson completed each step of the process described above, Winsteps control, item parameter, score, and output files were shared to check for inconsistencies. Winsteps output contained the number of cases in the calibration sample, item-level information (e.g., p-values, parameters), and the theta scoring tables. A sample of the output files are appended to this document. They include:

1. Winsteps Control Files (Appendix A). These files contain the item parameter estimation specifications and important information for reading the student score files. It also specifies the output file names. The appendix includes an example control file for the initial item parameter estimation, equated item parameter estimation, and estimation of the cluster scores.
2. Winsteps Item Parameter Files (Appendix B). These files contain the item parameters for the operational items. Each multiple-choice item has one parameter, a logit difficulty (named Measure in the Winsteps files). Each constructed-response item has an overall difficulty parameter and a number of step parameters indicating how the points for the item are distributed along the theta scale. The file included in the appendix is an example of a final item parameter file. Initial item parameter files are in similar formats.
3. Winsteps Anchor File (Appendix C). The file includes the 2019 item parameter values for each anchor item with the equating shift estimate applied to the overall difficulty measure. The file is read by Winsteps and used to fix the item parameter values and estimate final score files.
4. Winsteps Score File (Appendix D). The file contains the raw score to theta estimation and includes the distribution of student scores.
5. Comparison of Files Output (Appendix E). This is a SAS output file from HumRRO's comparison program that checks scoring table results against Pearson's results. The files match if all comparison values are 0.

Conclusion

Pearson and HumRRO independently calculated the scaled/equated raw-score-to-scale-score tables for the 2019 K-PREP Writing and High School Science assessments and verified the application of scoring tables for Reading, Mathematics, Science (grades 4 and 7), and Social Studies assessments. No differences were found between Pearson's and HumRRO's parameter estimations or raw-score-to-scale-score tables. Given that HumRRO's and Pearson's scaling and equating results were identical, HumRRO is confident that Pearson did not commit processing errors.

References

Bynum, B. H., & Thacker, A. A. (2011). *Third-party checking of calibration scaling and equating of the 2011 Kentucky core content test* (FR-11-65). Human Resources Research Organization.

Bynum, B. H., & Thacker, A. A. (2013). *Third-party checking of 2012 scaling and equating for the Kentucky Performance Rating for Educational Progress (K-PREP) Tests* (2013 No.11). Human Resources Research Organization.

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). Available online: 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 E. V. Smith Jr. & G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting*. JAM Press.

Pearson, Inc. (2012). *Kentucky performance rating for educational progress performance standards workshop: Performance level descriptor creation and standard setting, v1.1*. Author.

Appendix A – Control File (Writing Grade 5, Task 1)

```
;Winstep Control file h05WRF1T1_v0
; HumRRO
&INST
Item1 = 25
NI = 3
TABLES = 00100000000001000001000000001
CODES = 01234
FITP = 3.0
FITI = 3.0
XWIDE =1
HLINES = Y
data=h05WRmopv0.dat
IFILE= h05WRv0_F1T1.ITM
ISFILE = h05WRv0_F1T1.ISF
SFILE =  h05WRv0_F1T1.CSF
IAFILE = h05WRv0_anchorsEQ.IAF
SAFILE = h05WRv0_anchorsEQ.SAF
SCFILE = h05WRv0_F1T1.RSS
mprox=10
mucon=100
rconv=.50
lconv=.01
models=r
groups=0
stkeep=n
realse=n
stbias=n
target=n
extrsc=0.25
udecim=4
uimean=0
uscale=1
upmean=0
;uanchor=y
ptbis=y
ILFILES = *
I0001
I0002
I0003
*
IDELETE = 3;
 &END
END NAMES
```

Appendix B – Winsteps Item Parameter Files (Writing Grade 5)

Item parameters - Writing Grade 5, Task 1 (h05WRv0.ITM)

```
; ITEM  C:\Data\Kentucky\KPREP2019\WINSTEP\h05WRv0_F1T1.con  Jul 23 16:50 2019
;ENTRY MEASURE ST   COUNT    SCORE  ERROR IN.MSQ IN.ZST OUT.MS OUT.ZS  DISPL PTBISE WEIGHT OBSMA EXPMA DISCRM
LOWER UPPER PVALU PBE-E  RMSR G M R NAME
     1  -.9218  2 51155.0 107682.0  .0132    .55  -9.90    .41  -9.90  .0454   -.03   1.00  92.1  86.5   1.25
.00  4.00  2.11  -.07   .25 0 R . I0001
     2 -2.5987  2 21500.0  47347.0  .0179   1.01    .67   1.01    .60 -.0860    .53   1.00  81.7  81.1    .98
.00  4.00  2.20   .53   .38 0 R . I0002
     3   .0000 -3     1.0       .0  .0000   1.00    .00   1.00    .00  .0000    .00   1.00 100.0 100.0   1.00
.00  1.00   .00   .00   .00 0 R . I0003
```

Step parameters 2019 - Writing Grade 5, Task 1 (h05WRv0.CSF)

```
; STRUCTURE MEASURE ANCHOR FILE FOR C:\Data\Kentucky\KPREP2019\WINSTEP\h05WRv0_F1T1.con Jul 23 16:50 2019
; ITEM CATEGORY  Rasch-Andrich threshold MEASURE
     1  0   .0000
     1  1 -9.5721
     1  2 -2.8075
     1  3  3.3991
     1  4  8.9805
     2  0   .0000
     2  1 -10.288
     2  2 -3.0311
     2  3  3.4312
     2  4  9.8885
```

Appendix C – Winsteps Anchor File (Grade 5 Writing, Task 1)

Item Anchor File (h05WRv0anchorsEQ.IAF)

```
1  -0.9218
2  -2.5987
3  -2.4353
```

Step Parameter Anchor File (h05WRv0anchorsEQ.SAF)

```
; ITEM CATEGORY  Rasch-Andrich threshold MEASURE
 1 0      0.0000
 1 1     -9.5721
 1 2     -2.8075
 1 3      3.3991
 1 4      8.9805
 2 0      0.0000
 2 1    -10.2885
 2 2     -3.0311
 2 3      3.4312
 2 4      9.8885
 3 0      0.0000
 3 1    -11.9340
 3 2     -3.2783
 3 3      4.2749
 3 4     10.9372
```

Appendix D – Winsteps Score File (Grade 5 Writing, Task 1)

PERSON SCORE FILE FOR C:\Data\Kentucky\KPREP2019\WINSTEP\h05WRv0_F1T1.con Jul 23 16:50 2019 USCALE=1.00

SCORE	MEASURE	S.E.	INFO	NORMED	S.E.	FREQUENCY	%	CUM.FREQ.	%	PERCENTILE
0	-14.1285	2.2410	.20	201	48	25	.0	25	.0	1
1	-11.6971	1.6689	.36	253	36	608	1.2	633	1.2	1
2	-8.0914	2.4307	.17	330	52	7364	14.4	7997	15.6	8
3	-4.6825	1.5614	.41	403	33	3923	7.7	11920	23.3	19
4	-1.4696	2.2229	.20	472	48	20028	39.2	31948	62.5	43
5	1.6559	1.5172	.43	539	32	4654	9.1	36602	71.6	67
6	4.7727	2.2019	.21	605	47	12210	23.9	48812	95.4	83
7	7.6851	1.4307	.49	667	31	742	1.5	49554	96.9	96
8	9.6763	2.1532	.22	710	46	1601	3.1	51155	100.0	98

Appendix E – Comparison of Files Output (Writing Grade 5)

All RSSS Differences – Writing Grade 5

Obs	grade	formnum	wr_task	raw_ score	theta_ diff	se_diff	SS_diff	pl_diff
1	05	01	01	0	-.0005	.0000	0	0
2	05	01	01	1	-.0002	.0001	0	0
3	05	01	01	2	0.0000	.0000	0	0
4	05	01	01	3	0.0000	.0000	0	0
5	05	01	01	4	0.0000	.0000	0	0
6	05	01	01	5	0.0000	.0000	0	0
7	05	01	01	6	0.0000	.0000	0	0
8	05	01	01	7	0.0000	.0000	0	0
9	05	01	01	8	0.0000	.0000	0	0
10	05	01	02	0	0.0000	.0000	0	0
11	05	01	02	1	0.0000	.0000	0	0
12	05	01	02	2	0.0000	.0000	0	0
13	05	01	02	3	0.0000	.0000	0	0
14	05	01	02	4	0.0000	.0000	0	0
15	05	01	02	5	0.0000	.0000	0	0
16	05	01	02	6	0.0000	.0000	0	0
17	05	01	02	7	0.0000	.0000	0	0
18	05	01	02	8	0.0000	.0000	0	0