Equating Words-Correct-Per-Minute (WCPM) Scores in an Oral Reading Fluency Assessment

Jing Chen[1]
Mary Ann Simpson[1]

Abstract

Words-correct-per-minute (WCPM) scores from dissimilar passages are problematic for monitoring students' reading progress. We develop a method based on graph theory to identify the pairwise relationships between passages for equating. Such equated scores provide a better indication of students' oral reading fluency by accounting for differences in passage difficulty. Results from this study suggest that compared to the raw WCPM scores, equated scores have higher reliability and better reflect students' true reading ability as indicated by higher correlations with scores from an external reading measure.

Suggested citation: Chen, J., & Simpson, M. A. (2020). *Equating words-correct-per-minute (WCPM) scores in an oral reading fluency assessment.* Paper presented at National Council on Measurement in Education annual meeting, virtual.

[1] NWEA, Portland, OR, USA corresponding author email: jingchen@nwea.org

1. Introduction

Oral reading fluency (ORF) measures are designed to assess students' oral reading speed and accuracy as well as monitor their reading progress. Typically, a student reads one or more brief passages aloud for several minutes. The resulting words-correct per minute (WCPM) score serves as a formative indicator of a student's oral reading fluency (ORF). Raw WCPM scores are problematic for progress monitoring because passages vary in difficulty.

Readability formulas have been heavily relied on for selecting equivalently difficult passages (Ardoin et al., 2005). However, readability formulas have limited utility for predicting student actual reading performance across passages (Francis et al., 2008). Researchers have suggested psychometric equating of passages to account for such differences in passage difficulty and allow different passages to be used interchangeably (Ardoin et al., 2005; Christ & Ardoin, 2009; Francis et al., 2008; Poncy et al., 2005). However, ORF measures are often created according to a formative assessment model, which rarely meets the stringent assumptions for equating.

Along these lines, the goals of this study were 1) develop a method to simplify the task of equating a large number of passages, 2) evaluate equating results to see if equated scores provide a better indication of students' oral reading fluency, and 3) compare the results from three equating methods (mean, linear, and equipercentile) to identify the method works best in our sample.

2. Method

2.1. Data

Data were collected from a recently developed Oral Reading Fluency (ORF) assessment designed for students from grades K-3. In each test, students read two or three passages, with approximately 200 words each. The test was multi-stage adaptive. Students performing well on their first passage received a more difficult passage for their second and vice versa. Because of the adaptive nature of the oral reading forms, the passages are not equivalent with regard to their text difficulties. Passages were developed at grade- and term-appropriate levels of text complexity, as gauged by their Lexile® text measures. To make more meaningful comparison of students' WCPM scores from different passages, raw WCPM scores need to be adjusted to take passage difficulty into account so that scores from different passages are aligned on a common scale.

Data were collected from 56,343 K-3 students across 43 states in winter 2019 during the test window between Dec. 2018 to Feb. 2019. Twenty-six passages were administered across grades K-3. Table 1 lists the sample size and the percentage of students from each of the grade levels and the percentages of students grouped by gender, ethnicity, and social economic status across grades.

Table 1. Demographics of the Sample

Sample Characteristics	Subgroups	N	Percentage
Grade	Grade K	42	0.1
	Grade 1	12,092	22.3
	Grade 2	26,405	48.8
	Grade 3	15,564	28.8
Gender	Female	26,807	49.6
	Male	27,257	50.4
Ethnicity	American Indian or Alaskan	831	1.5
	Asian/Pacific Islander	3,866	7.1
	Black/African-American	9,118	16.9
	Hispanic	8,514	15.7
	Multiethnic	2,516	4.7
	Native Hawaiian/Other Pacific Islander	150	0.3
	Other/Not Specified	3,852	7.1
	White	25,256	46.7
SES	Low-SES	4,916	9.1
	Not low-SES	49,187	90.9

2.2. Equating Procedure

To equate all passages, equipercentile equating with loglinear pre-smoothing was applied to convert raw WCPM scores from a non-reference passage to those from the reference passage following the steps below:

1. Choose equating design
2. Identify the reference passage in each term
3. Define the shortest path to the reference passage
4. Choose equating method

All equating processes were conducted using data from students across grades because the relationship between the WCPM scores of passages is assumed to remain the same across grades. Otherwise grade-specific equating relationships using data from a specific grade would have been needed, which is difficult to implement and even more difficult to justify.

Outliers were excluded from the data when building the equating relationship between a pair of passages to build more reasonable relationships. They were identified by the Mahalanobis distance of < -10 or > 10, a statistic that helps find observations that are outlying on all variables involved in an analysis. Linear interpolation was used to identify integer score points in chained equatings, and linear extrapolation was used to build the equating relationship beyond the range of the scores in the data to produce plausible results. Conversion tables for reported scaled WCPM (SWCPM) scores were capped at 20 SWCPM at the low end and 170 SWCPM at the high end because very low SWCPM scores are likely unreliable, and the high-end caps were introduced to prevent over-interpretation of SWCPM scores. Oral reading fluency manifests a "good enough" quality beyond which extra speed offers little further benefit to the reader.

2.2.1. Choose Equating Design

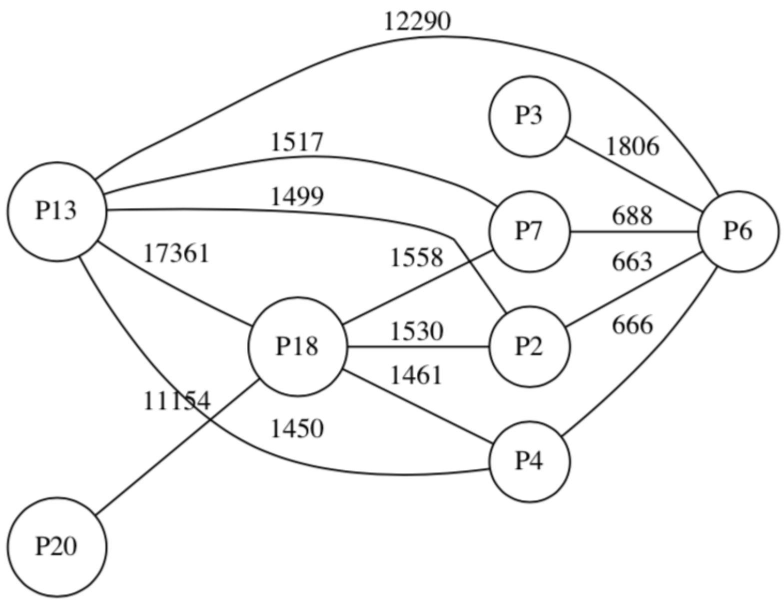
A reference passage (i.e., anchor passage) was needed to place all 26 passages on a common scale. Given the adaptive nature of the test, a single group design was adopted where the equating relationship is built based on the scores from a pair of passages read by the same student during one test event. Although no official "reference passage" had been worked into the original test design, we developed a method to identify a representative reference passage to which all the other passages could be equated. With this design, a new passage can be equated as long as enough students read the new passage and an existing passage has already been equated. This design does not require administering the exact same test forms or the exact same reference passages.

2.2.2. Identify the Reference Passage

Graph theory is a mathematically structured way to visually networks of objects (exSTEMsions, 2019). In graph theory, objects in a graph comprise nodes and edges. In this analysis, nodes are passages, and the edges identify pairs of passages read by the same students. To identify the reference passage in each term, all possible passage pairs were first identified. Figure 1 shows a graph representations of passage pairs (only 8 passages included for illustration purposes). The number along each line is the number of students that took each passage pair and received valid WCPM scores on both passages (i.e., WCPM score >0). For example, 12,290 students took Passage 13 and Passage 6. A small proportion of students with invalid WCPM scores from either passage in each pair were excluded from all data analyses. All the passages administered in the term and all students across grades that took each passage pair were included.

A reference passage was then selected based on having a greater number of edges in the graph, adequate sample sizes for most pairs, and medium text difficulty. A larger number of edges for a passage indicates more frequent pairing with other passages. Large sample sizes reduce equating errors. Average text difficulty ensures the WCPM score distribution of this passage overlaps the score distribution of other passages to a large extent so that typically no single passage would have substantially many equated scores that extend beyond the reference passage scale. In this example, Passage 13 (P13) is selected to be the reference passage because of these properties. All other passages administered were equated to this reference passage.

Figure 1. Graph representation of all pairwise relationships between eight passages



2.2.3. Define the Shortest Path to the Reference Passage

As shown in Figure 1, two passages can be connected through more than one path, which makes it complicated to equate scores from one passage to the scores of the reference passage. Graph theory provides a way to model this type of pairwise relationships by defining the shortest path between two objects. Based on graph theory, NWEA defined the shortest path from a passage to the reference passage as the path with the most data points and fewest connecting nodes. This shortest path results in the least amount of equating error because the equating relationship will be built based on more direct connections and larger sample sizes, making it desirable. For example, in Figure 1, the shortest path from passage 3 to passage 13 is through passage 6 rather than other paths. The WCPM scores of Passage 3 can be equated to the WCPM scores of Passage 6 and then to the scores of Passage 13, the reference passage, through the shortest path.

2.2.4. Choose Equating Method

Finally, we conducted mean, linear, and equipercentile with loglinear presmoothing equatings to the reference passage. For the equipercentile method, we applied the loglinear pre-smoothing method (Holland & Thayer, 2000) to obtain lower SEEs. The sample size of each passage pair was approximately 1500 or more, which is sufficient for these equating methods in a single-groups design (Kolen & Brennan, 1995). Linear interpolation and extrapolation were employed to produce plausible results at the ends of the score scale. All data preparation, smoothing and equating were accomplished with R and the equate, R package (Version 2.0.7, 2018).

3. Results

3.1. Reliability Before and After Equating

The reliability of the fluency measure can be evaluated by the correlations between the WCPM scores from the passages that the same student read in one sitting to evaluate the consistency of the scores across passages. The same student's equated scores from different passages have higher correlations than the correlations from his/her raw scores. This indicates that explicit equating is essential to make WCPM scores from different passages equivalent and more reliable. The three equating methods yield almost identical correlations for different passage pairs. No method outperforms the others in terms of producing more reliable equated scores.

Table 2 presents the correlations between the WCPM scores from the three passages that the same student read. For example, 'Passage 1 & 2' is the first passage and the second passage a student read. The first, second, and third passages presented to different students are different depending on the test form that the student took. The 'Raw', 'Mean', 'Linear', and 'Equipercentile' columns present the correlations of the WCPM scores before equating and after equating based on each of the three equating methods.

Table 2. Correlations between Students' Raw WCPM Scores and Equated WCPM Scores from Three Equating Methods

Passage Pairs	N	Raw	Mean	Linear	Equipercentile
Passage 1 & 2	44,012	0.897	0.910	0.910	0.910
Passage 1 & 3	41,963	0.842	0.888	0.888	0.888
Passage 2 & 3	41,471	0.850	0.907	0.907	0.906

3.2. Equating Accuracy

The reference passage, P13, was administered together with each of 22 passages[2] as a pair on a test form. This means some students took the reference passage, P13, and a non-reference passage together on a test form. So, their actual WCPM scores from the reference passage and their WCPM scores from a non-reference passage equated to the reference passage scale can both be derived and compared to evaluate the accuracy of equating. Table 3 shows the residual sum of squares between students' scores on the reference passage and their scores on a non-reference passage before and after equating averaged across students.

Differences between students' scores on the reference passage and another passage equated to it are smaller than differences between students' raw scores. The equipercentile method produce the closest results for most passages as indicated by the lowest mean residual sum of squares among three methods.

[2] Only three passages were not administered together with the reference passage on a test form.

Table 3. Mean Residual Sum of Squares between Students' Actual Scores on the Reference Passage and Their Scores from Another Passage Equated to the Reference Passage

passage	N	Raw	Mean	Linear	Equipercentile
P1	1,460	210.653	198.878	191.259	**189.886**
P2	1,499	211.916	175.113	169.378	**168.524**
P4	1,450	214.268	**152.301**	153.001	152.372
P5	1,514	215.191	179.375	179.659	**177.769**
P6	12,290	159.021	**157.835**	162.241	162.336
P7	1,517	216.231	153.745	153.546	**152.999**
P8	1,431	215.945	167.953	168.038	**167.720**
P9	1,511	186.654	157.799	**152.422**	152.514
P10	1,450	176.931	154.570	**149.361**	149.548
P11	1,472	198.035	152.088	147.730	**146.868**
P12	1,495	219.951	**141.205**	142.167	141.809
P14	1,429	183.264	173.022	151.964	**150.955**
P15	1,492	277.426	163.210	156.032	**154.313**
P17	1,503	227.675	170.148	**166.854**	167.868
P18	17,361	191.495	146.418	134.448	**133.816**
P19	1,393	265.247	161.212	141.358	**140.052**
P21	1,337	472.970	160.170	160.441	**159.356**
P22	1,291	493.293	196.178	167.164	**166.246**
P23	1,414	316.668	155.631	139.574	**138.699**

3.3. Validity Evidence Before and After Equating

Compared with raw scores, equated scores have higher correlations with students' reading scores from an external reading measure, which suggests that equated scores may reflect students' true reading ability better. The difference between the correlations from 3 equating methods is minimal. Around half of the students took the reading fluency measure also took an external reading assessment. The correlation between students' reading scores from an external measure and their raw WCPM scores, and those from mean, linear, and equipercentile equating methods were 0.5910, 0.6158, 0.6163, and 0.6163 respectively. The WCPM score was calculated as the average of the WCPM scores of all the passages that the student read.

4. Conclusion and Implications

We introduced a new method to select reference passages for equating of WCPM scores. This method overcomes some design limitations of typical ORF assessments and makes equating a large number of passages feasible. Additionally, our results provide evidence of higher reliability and validity of equated passage scores than those of the raw scores. The equipercentile method slightly outperformed linear and mean equating as it produced scores closest to students' raw scores on the reference passage for most passages. Generally speaking, equipercentile equating is preferable given that it is the most general approach and can accommodate any degree of nonlinearity across forms. However, the choice of equating method under other situations should be based on the characteristics of the score distributions and the available sample size to build the equating relationship.

Passage equating can potentially reduce the assessment burden for students and teachers by reducing the number of passages students need to read to get a reliable and valid ORF score. Passage equating can help make more meaningful comparison across passages to monitor real progress over time. More precise and reliable WCPM scores that account for text difficulty differences are more acceptable to be used for making high-stakes educational decisions as well as providing more precise information for classroom teaching and learning.

References

Ardoin, S. P., Suldo, S. M., Witt, J. C., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20*(1)*,* 1–22.

Christ, T.J. & Ardoin, S.P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probeset development. *Journal of School Psychology, 47,* 55-75. doi: 10.1016/j.jsp.2008.09.004.

exSTEMsions. (2019, May 2). What is graph theory, and why does it matter so much? [Blog post]. Retrieved from, https://exstemsions.com/blog/graphtheory.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3)*,* 315–342.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–83.

Kolen, M. J., & Brennan, R. J. (1995). *Test equating: methods and practices.* New York: Springer-Verlag.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute. *Journal of Psychoeducational Assessment, 23*, 326–338. doi:10.1177/073428290502300403.