Clinical Aphasiology, Vol. 24, 1996, pp. 145–157

A Comparison of Aphasic Performance on the Standard and an Experimental Computerized Version of the Revised Token Test

Katharine H. Odell, Suzanne Bonneau Miller, and Charles Lee

Despite development of numerous computerized aphasia treatment programs, there is little research examining the use of microcomputers in aphasia assessment. Aspects of some aphasia tests, such as the Western Aphasia Battery (Kertesz, 1982) and the Porch Index of Communicative Ability (Porch, 1967) have been adapted for computer-assisted administration, with apparently positive results (Wertz, 1992). Only portions of these tests have been modified for computerization, because of the current limitation in computer recognition of speech, especially disordered speech, and written expression. Generally these tests are most useful in assessing the midrange of aphasia severity.

Measures for more mildly impaired individuals have not yet been adapted for computerized administration. One such test, the Revised Token Test (RTT) (McNeil and Prescott, 1978) is a sensitive measure of auditory language processing in brain-damaged individuals. The RTT is almost uniquely apt for computerization of stimulus presentation and response scoring. It assesses auditory language processing by means of gestural response, that is, token manipulation. No oral expressive language or writing, which are beyond the capabilities of current software to recognize or analyze, are involved. The test is standardized in both its presentation of stimuli and its application of a 15-point multidimensional scoring system. All these features facilitate a computer adaptation that is equivalent to the standard version.

Successful computerization of this particular test will be beneficial for several reasons. Computerized scoring will greatly simplify use of the multidimensional scale. Currently, a 25-hour training program is required

for clinicians to administer and score the RTT. Instrumental control of the timing of voice commands will enhance the standardization of test administration. If a computerized version of the RTT proves to be reliable and similar in measurement of response behaviors to the standard test, its use in clinical and research settings will be facilitated. Because the RTT is believed to be sensitive to subtle deficits in auditory processing, increased use might enhance initial evaluations as well as assessment of response to treatment. In addition, computerization may well encourage its use as an experimental measure in research efforts.

Thus, this study was undertaken to secure preliminary data on the degree to which the standard and a computerized version of the RTT are similar in measuring quantitative and qualitative aspects of auditory processing performance. To this end, the study compared the performance of aphasic adults on the RTT and on an experimental computerized version of this test (hereafter: CRTT, for computerized RTT). Evidence that subjects do perform similarly on the two versions will encourage undertaking a substantial software and hardware upgrade of the current CRTT program and acquiring normative data for it.

METHODS

Subjects

Seven aphasic subjects were tested. The selection criteria were:

1. Presence of aphasia, as determined by two measures: performance within the range of aphasic performance on the Shortened Porch Index of Communicative Ability (SPICA) (DiSimoni, Keith, and Darley, 1980) and the auditory comprehension subtests of the Boston Diagnostic Aphasia Exam (BDAE) (Goodglass and Kaplan, 1983);
2. Neuroradiological evidence of a single left cerebral hemisphere lesion, with sudden onset aphasia, as indicated by physician medical record notes and family report;
3. Absence of dementia or mental illness, indicated by a negative history of personality changes or gradual decline in memory or language abilities, as noted in the medical chart or by family report;
4. Native speaker of English;
5. Adequate hearing, as determined by four measures:
   - Normal otoscopy
   - Normal tympanometry

- Puretone air conduction thresholds that fell within one standard deviation on either side of the mean dBHL threshold considered normal for each test frequency for the subject's age (ISO 1984). Test frequencies were: 500 Hz and the octave frequencies 1–6 KHz;
- Word discrimination scores of 90% correct or better in sound field. Tests used were either the Northwestern University Auditory test No. 6 (NU6) (Tillman and Carhart, 1966) or Word Intelligibility by Picture Identification test (WIPI) (Ross and Lerman, 1970) (pass for WIPI was 96% correct);

6. Adequate vision capabilities, as indicated by the ability to successfully take the RTT and CRTT pretests.

Selected details of subject status on each measure are included in Table 1.

Technical Development

The CRTT was implemented on an IBM-PC, upgraded to the equivalent of an IBM XT, with a TECMAR color generation board. An RGB 13-inch monitor was interfaced with a TSD touch sensitive screen. A voice digitization system interfaced with the computer was played through an external speaker. Software was written in the C language.

Table 1. Subject Status on Selected Measures of Enrollment Criteria

Subject	*Age*	*Gender*	*TPO*[1]	*SPICA*[2]	*BDAE*[3]	*Word Discrimination*[4]
1	70	M	36	14.35	70 – 100%ile[6]	100
2	69	M	10	13.44	60 – 80%ile	100
3	59	M	30	13.51	70 – 90%ile	94
4	57	M	13	12.24	70 – 100%ile	94
5	86	F	8	12.38	70 – 100%ile	86[5]
6	63	F	38	7.10	30 – 100%ile	90
7	53	F	39	12.96	70 – 80%ile	90

[1] TPO = Time post onset in months.
[2] SPICA = Shortened Porch Index of Communicative Ability (DiSimoni et al., 1980).
[3] Auditory comprehension subtests of the Boston Diagnostic Aphasia Exam (BDAE) Goodglass and Kaplan (1983). *Assessment of aphasia and related disorders).*
[4] Word discrimination = Percent correct on a sound field administration of the NU 6 (Tillman and Carhart, 1966).
[5] Percent correct on NU 6 (Tillman and Carhart, 1966); administration of the WIPI (Ross and Lerman, 1970) resulted in a score of 94% correct.
[6] Range of percentiles on auditory comprehension subtests on BDAE (Goodglass and Kaplan, 1983).

The CRTT was designed to be administered by computer with both prescribed and optional clinician intervention. The clinician enters all subject biographical data, administers a practice session, and can enter scores or anecdotal information about performance from the keyboard.

The entire software package includes several forms and components of the computerized test. The various versions of the CRTT are: standard (all 10 subtests, all 10 items per subtest); abbreviated (all 10 subtests, first five items per subtest); and modular (ability to present any single subtest in isolation or any sequence of any number of subtests in any order). The components for any of the test versions administered include: a personal data and test history file; a practice session, described in more detail below; a selected test version; playback (ability to recreate on the screen the test as performed by the subject); scoring (on-screen or hard-copy view of all the individual item scores and summary scores as dictated in the RTT manual; at this time, no percentile ranks have been included, as that must wait for future normative data).

The mandatory practice session involves two parts: In the first part, there is practice in manipulating the tokens via the touch screen without any test commands. Also at this time, monitor height, room and monitor illumination, and whether the subject will use his finger or a blunt tool, such as the smooth end of a pen, to touch and move the tokens are also determined. Part two is the computerized version of the standard RTT pretest, that gives subjects the opportunity to hear the voice through the speaker and to adjust to the test pace.

The 15-point multidimensional scoring system of the RTT was essentially unmodified for the CRTT. However, it was necessary to arrange for optional clinician intervention in the automatic computer scoring. Some aspects of response behaviors are invisible to the computer but not to the clinician; thus, the clinician can press a designated key on the keyboard and override a computer score or add information on the nature of the computer-generated score. For instance, vocal or subvocal behavior by the subject will be missed by the computer but can be signaled by the clinician by pressing designated keys. Also, in cases when the subject verbally or by gesture requests a repeat or cue of a command, the clinician must press designated keys to respond to the request. The entire RTT scoring scale is shown in Table 2. Those scores in boldface are ones that allow the clinician to intervene or elaborate the computer-generated score (i.e., 14–rehearsal; 13–delay or gestural uncertainty; 9–repeat; 8–cue; 5–rejection; 4–unintellgible differentiated; 3–unintelligible undifferentiated).

The touch sensitive screen registers even slight finger or pen movements which result in visible token movement. Thus, precise criteria were devised for determination of when a touch would be considered a sloppy touch but not an actual move of a token. It was decided that final movement of a token equal to one pixel more than one-half the diameter

Table 2. Behavioral Descriptions of Scores Used in Both the RTT and CRTT. Scores in Boldface Are Those That Allow or Demand Clinician Intervention in CRTT Scoring

Score	*Description of Response*
15	Complete, accurate, timely
14	**Rehearsal**
13	**Delay**
12	Immediate
11	Self-correction
10	Reversal
9	**Repeat**
8	**Cue**
7	Error
6	Perseveration
5	**Intelligible but wrong response; rejection of command**
4	**Unintelligible, wrong response that is differentiated**
3	**Unintelligible, perseverated**
2	Omission of one part of a two part command
1	No response

of the token would be recognized by the computer as a *move* not a *touch.* It was also necessary to specify the boundaries of correct token placement. Placement issues were pertinent in two sets of subtests: subtests V and VI, requiring tokens to be placed in various locations relative to other tokens—below, on etc.; and subtests VII and VIII, requiring tokens to be placed to the left or right of other tokens. Subjects had extensive opportunity during the practice session to make simple touches and actual moves.

Test Conditions

Subjects were tested individually on two days. Each subject took both versions of the test. The order of presentation of each of the aphasia and experimental tests was semirandomly assigned across subjects.

Analysis

Three types of comparisons of performance in the two conditions (RTT and CRTT) were completed: overall (OA) scores, subtest scores, and the frequency distribution of score categories (e.g., scores of 15, 13, or 10). This last comparison was done because the overall and subtest mean scores are

summary scores that obscure the quality of behavior as captured in category scores (e.g., *13,* reflecting delay, and *10,* reflecting self-correction) assigned to response behavior relative to each critical element in RTT commands. To compare the quality of subject performance on the two versions, differences between conditions were examined with respect to the number of times a particular score was assigned.

Data for both the OA test comparison and the subtest comparison were subjected to nonparametric statistical analyses for two reasons: the small number of subjects, with the undoubted consequence of violation of the normalcy assumption required for parametric statistical tests, and the unknown distribution of severity for the population of aphasic individuals on measures such as these. The OA scores and the subtest scores across versions were each submitted to a Wilcoxon (matched pair) signed ranks test, the nonparametric analog to a *t* test of difference scores between related samples (Siegel and Castellan, 1988). The Wilcoxon signed ranks test uses information in the data relating to both direction of score difference between related pairs as well as size of the score differences between pairs. For the analysis of the overall score, the alpha level was set at .05. To more fully appreciate the relationship between group OA scores on the two versions, a nonparametric Spearman rho of OA scores were calculated, using an alpha level of .05.

For the 10 subtests' analyses, the familywise alpha was .05, but to account for the multiple Wilcoxon tests, a conservative alpha level of $p < .01$ was adopted. While dividing the alpha level does reduce the possibility of Type I errors, it simultaneously increases the risk of Type II errors, specifically, the failure to detect a difference that exists. Accepting this heightened risk in this preliminary study was judged reasonable.

The tests were established as one-tailed, to reflect the a priori notion that CRTT scores would be slightly lower than RTT scores. The CRTT was expected to impose stricter scoring criteria than the RTT, causing lower scores, in instances such as *delay* (score of *13*), because the computer is more precise than humans in temporal measurement. Similarly, the strict criteria for token placement and movement might negatively affect CRTT scores. Lower CRTT scores might also be expected because subjects are not familiar with computers, resulting in hesitancy or self-corrections.

RESULTS

Overall and Subtest Scores

Table 4 displays the results of the statistical tests. OA mean score between versions was not significantly different ($T = +16$). As shown in Table 3, overall mean score on the RTT was higher than on the CRTT by .12

Table 3. Overall and Subtest Mean Scores for All Subjects for the RTT and CRTT

	Overall		*Subtest I*		*Subtest II*		*Subtest III*		*Subtest IV*		*Subtest V*	
	RTT	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*
S1	14.13	13.30	15.00	15.00	13.90	14.50	14.33	14.30	13.00	14.40	14.13	12.20
S2	11.52	12.90	14.60	15.00	14.10	14.60	12.13	13.00	11.90	13.30	12.27	11.80
S3	13.04	13.00	15.00	15.00	14.20	15.00	12.87	14.30	12.40	13.30	12.90	11.90
S4	13.13	11.60	13.80	11.90	13.60	12.70	14.33	11.90	12.30	8.70	10.80	12.53
S5	12.52	12.60	14.20	15.00	13.80	13.00	11.40	12.20	11.40	10.10	12.07	12.40
S6	8.96	8.20	12.13	13.00	11.20	10.80	10.13	11.80	8.96	10.60	8.80	5.00
S7	11.84	12.70	14.60	14.50	11.70	15.00	12.80	14.00	11.84	13.60	10.90	12.00
Avg.	12.16	12.04	14.19	14.20	13.21	13.66	12.57	13.07	11.69	12.00	11.70	11.12

	Subtest VI		*Subtest VII*		*Subtest VIII*		*Subtest IX*		*Subtest X*	
	RTT	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*	*RTT*	*CRTT*
S1	15.00	15.00	14.40	13.80	13.90	13.10	13.70	9.80	14.68	14.00
S2	14.60	15.00	11.50	12.60	11.80	11.80	7.30	12.80	7.44	12.50
S3	15.00	15.00	14.00	12.70	11.43	11.20	14.20	13.80	12.28	11.90
S4	13.80	11.90	13.85	12.87	13.10	9.30	14.20	13.40	13.96	13.40
S5	14.20	15.00	11.03	13.00	12.35	11.60	12.00	13.90	13.64	13.90
S6	7.45	5.00	8.93	6.00	5.38	8.90	10.70	6.70	8.08	4.20
S7	10.00	9.80	11.60	12.50	10.60	11.40	12.60	13.00	11.20	10.70
Avg.	12.86	12.39	12.19	11.92	11.22	11.04	12.10	11.91	11.61	11.51

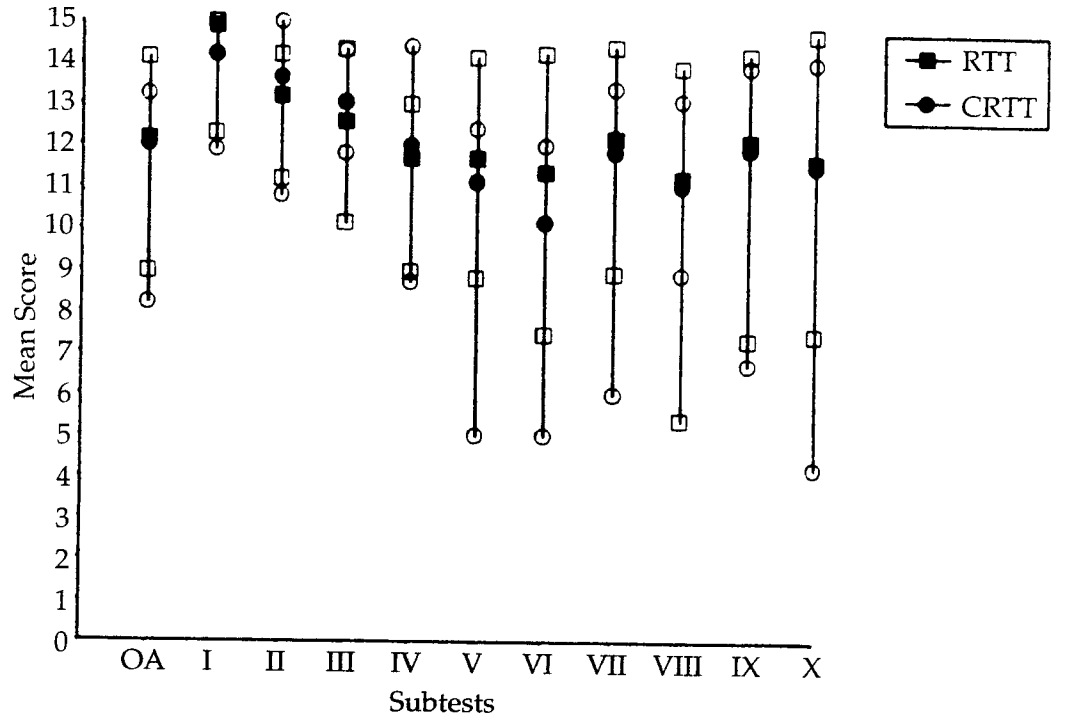

Figure 1. Group overall (OA) and subtest mean scores and ranges. Mean scores are represented by filled symbols; ranges are represented by unfilled symbols.

points. The general level of performance across subjects clustered between 11 and 13, but the range of performance level across subjects was substantial, as detailed in Table 3 and illustrated in Figure 1. The Spearman rho also was not significant ($r_s = + .5$), indicating that the rank order of each subject's OA score did not remain the same in the two versions. The failure of the Spearman procedure to attain statistical significance is interpreted as further support for the conclusion that these subjects did not, as a group, perform differently on the two versions. Although OA scores did vary randomly (i.e., unpredictably) in rank across versions, these changes in rank did not affect the significance of the group mean OA differences.

Similarly, as shown in Table 4, subtest mean scores between versions did not differ significantly (alpha = $p < .01$). The ranges of performance on each subtest in both versions was again considerable, as shown in Table 3 and Figure 1. RTT ranges were greatest for subtests IX and X (e.g., on IX, a low of 7.3 to a high of 14.2; from a low of 7.4 to a high of 14.7 on X). On the CRTT, ranges were greatest for X (4.2 to 14).

RTT scores exceeded CRTT scores in six of the 10 subtests. Only in the first four subtests did group CRTT performance exceed RTT performance.

Table 4. Results of Statistical Tests

Comparison	*Test*	*Test Result Value*	*Significance*
Mean overall scores between versions	Wilcoxon signed ranks	T = +16	NS[1]
Relationship of overall scores between versions	Spearman rank-order correlation coefficient	r_s = + .50	NS[1]
Subtest mean scores between versions	Wilcoxon signed ranks	Subtest I: T = +6 Subtest II: T = +11.5 Subtest III: T = +8 Subtest IV: T = +9 Subtest V: T = +18 Subtest VI: T = +27 Subtest VII: T = +17 Subtest VIII: T = +12.5 Subtest IX: T = +15.5 Subtest X: T = +20	NS[2]
Relationship of Severity (SPICA OA) and RTT	Spearman rank-order correlation coefficient	r_s = + .50	NS[1]
Relationship of Severity (SPICA OA) and CRTT	Spearman rank-order correlation coefficient	r_s = +1.0	S[1]

[1] Nonsignificant at $p < .05$
[2] Nonsignificant at $p < .01$

Three of these subtests are often considered among the easiest of the entire battery; the exception is subtest IV. CRTT scores fell behind in the other subtests involving movement, that is, token placement.

Regarding individual performance, across the first four subtests on which the RTT ranked lower for the group as a whole, three of the seven subjects performed consistently with the group trend on subtest I, four performed like the group on subtest II, and five of the seven performed like the group on subtests III and IV, as shown in Table 5. On subtests V–X, in which the RTT outranked the CRTT, three, four, or five subjects behaved consistently with the group performance (Table 5). Perusal of the data does not highlight factors that might be invoked to explain this inter-subject variability. Influences such as aphasia severity or TPO do not appear to bear a consistent relationship to the variability.

Table 5. Tabulation of Instances in Which the RTT Mean Score Was Higher Than the CRTT Mean Score for Individual Subjects and the Group

	OA	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*
S1	+	=	−	+	−	+	=	+	+	+	+
S2	−	−	−	−	−	+	−	−	=	−	−
S3	+	=	−	−	−	+	=	+	+	+	+
S4	+	+	+	+	+	−	+	+	+	+	+
S5	−	−	+	−	+	−	−	−	+	−	−
S6	+	−	+	−	−	+	+	+	−	+	+
S7	−	+	−	−	−	−	+	−	−	−	+
Group performance	−	−	−	−	−	+	+	+	+	+	+
Number of Ss with performance consistent with group	3	3	4	5	5	4	3	4	4	4	5

Relationship of Aphasia Severity and Performance

The relationship between aphasia severity, as determined by the SPICA overall scores, and OA performance was extremely strong for the CRTT ($r_s = 1.0$), but not significant for the RTT ($r_s = .5$) (see Table 4).

Score Category Frequency

Because score categories reflect qualities of response, a difference in category frequency between the CRTT and the RTT would indicate a change in the quality of response behavior as a function of test version. The frequency with which each score was assigned in the two versions was tabulated for each subject, as shown in Table 5. Cell entries for each score category are not identical across subtests or subjects because the frequency with which each score was assigned varied for each individual; in addition, each subtest includes different numbers of linguistic elements (e.g., nouns, adjectives) to be scored. However, each version of the total test registers 290 scores (580 across both versions). All subjects received 290 scores per test; the exception was subject S3, whose score category totals for subtests IX and X on the CRTT could not be retrieved from the computer, resulting in a total of 490 scores on the CRTT; an adjustment of the totals on the RTT for this subject was made for calculation purposes.

Some notable differences were apparent in the quality of performance across versions, as reflected in the frequency of score category assignment (Table 5), however, visual inspection of the data does not reveal any consistent trends across subjects and subtests. For S1, the score pattern indicates the subject was faster and more accurate (more 15s, fewer 13s) on the RTT, but also required more repeats (9s), suggesting a negative aspect of increased speed might be reflected in occasional less efficient processing. A similar profile was noted for S5. Subject 4 received more scores reflecting fast and accurate performance (15s) on the RTT but also made more errors (7s). In contrast, S7 received a greater number of accurate (15s) scores on the CRTT but also more delayed (13s) and error scores (7s).

On both versions, by far the greatest number of scores assigned fell into four categories: 15s (complete, accurate, timely), 13s (delayed), 9s (repeat), and 7s (error). One subject (S6), incidentally the subject with the lowest SPICA score, received a large number of 5s on the CRTT, which in this case indicated the subject exceeded the time limit on these commands. There were considerably more 13s (delay) assigned by computer in the CRTT than by the clinician in the RTT.

DISCUSSION

This study examined the similarity in measurement of two versions of the RTT, as a preliminary step in further development of a computerized version of the test. The bulk of the data indicate that subjects in this study performed similarly on the standard and computerized version of the RTT. Trends revealing different styles of response on one version or the other were not apparent. Intersubject variability, in comparison with the group performance, was noted; however, the variability was not easily explained by factors such as severity of aphasia. Similarity in performance on the CRTT and RTT is certainly not requisite to justify continued refinement of the CRTT, however, it does support the argument that the computer version measures auditory processing performance comparably to the standard version.

That performance on the two versions did not significantly differ was a welcome but somewhat surprising event. It was expected that scores would be notably lower on the CRTT than the RTT, due to stricter computer scoring of responses and the novelty of computer use for the subjects. Results of this study may well be influenced by the small number of subjects. Additionally, the fact that most of these subjects, except S7, were rather mildly aphasic may have influenced the results; more moderately impaired subjects may not perform as similarly on the two versions as the subjects in the current study.

Before the CRTT can be used for clinical or for research purposes, additional development is required. The CRTT must be upgraded to be compatible with current software and hardware capabilities. Future efforts must also address the limitations of this study, namely, the small sample size, the consequent lower statistical power, and the lack of a wide range of aphasic severity in the subjects. The temporal reliability of the CRTT must be assessed. Finally, new normative data on the CRTT version must be acquired.

REFERENCES

DiSimoni, F., Keith, R., & Darley, F. (1980). Prediction of PICA overall score by short versions of the test. *Journal of Speech and Hearing Research, 23,* 511–516.

Dunn, O. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52–64.

Goodglass, H., & Kaplan, E. (1983). *Assessment of aphasia and related disorders.* Philadelphia: Lea & Febiger.

International Standard 7029. (1984). *Acoustics—Threshold of hearing by air conduction as a function of age and sex for otologically normal persons.* International Organization for Standardization. [Ref. No. ISO-7029-1984 (E)].

Kertesz, A. (1982). *Western Aphasia Battery.* New York: Grune & Stratton.

McNeil, M., & Prescott, T. (1978). *Revised Token Test.* Baltimore: University Park Press.

Porch, B. (1967). *Porch Index of Communicative Ability.* Palo Alto: Consulting Psychologists.

Ross, M., & Lerman, J. (1970). A picture identification test for hearing-impaired children. *Journal of Speech and Hearing Research, 13,* 44–53.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences,* (2nd ed.). McGraw-Hill: New York.

Tillman, T., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CVC monosyllabic words (Northwestern University Auditory Test No. 6).* Technical report, SAM-TR-66-55, USAF School of Aerospace Medical Division, (AFSC) Brooks Air Force Base, TX.

Wertz, R. (1992). Matreshka in aphasiology: Less is more. *Clinical aphasiology* (Vol. 21, pp. 1–7). Austin, TX: PRO-ED.