Retesting oral language skills using identical or adapted versions of the clinical evaluation of language fundamentals-third edition (CELF-3)

ABSTRACT

One of the most commonly used standardised assessments, testing children for oral language skills, is called The Clinical Evaluation of Language Fundamentals (CELF-3 or 4).[1] There is a growing belief, especially among speech pathologists, that the original CELF-3 assessment package may need to be supplemented by 'adapted' versions when a child is retested, is an attempt to eliminate the possibility that variances such as memory and practice may distort the result.

This is a critical issue. The initial CELF-3 is a diagnostic clinical tool of oral language disorders as well as an indicator of the level of intervention needed by speech pathologists to help develop the child's skills in specific areas of oral language. Hence, future retesting of the child would need to be fairly accurate if it is to be relied on as a key tool in assessing the degree to which intervention has worked.

In this paper, I hypothesise that applying adapted versions of CELF-3 during retesting will produce the same outcome as retesting using the original CELF-3. In other words I propose that it makes no difference whether a child is retested using CELF-3 or a version of this test as long as the same oral language skills are being assessed.

To test my hypothesis I randomly chose 20 primary school children aged between 6 and 7, dividing them into two equal groups. For the purpose of the study I formulated one CELF-3 sub-test, rephrasing the questions, ensuring that the changes were minimal by replacing only the key words and the pictures.

There were two test sessions, one hour apart. In Session 1, I applied the original CELF-3 test to both groups. In Session 2, I applied the CELF-3 test to Group 1 and the Adapted Version to Group 2.

The results showed a close parallel between the two groups' performance in the CELF-3 in Session 1, and the retesting using CELF-3 and the Adapted Version in Session 2. Group 1 showed an improvement of 3 points over the same CELF-3 test and retest, while Group 2 showed a rise of 4 points

[1] For the purpose of this research paper I will be using CELF-3 as it is the test I am most familiar with. Also, by not using CELF-4, I wanted to eliminate the possibility that some of the subjects in my test group may have recently completed CELF-4 which is the latest version.

between the CELF-3 test and the Adapted Version. The difference was only one point, indicating that:

1. there was an improvement shown by both groups after retesting regardless of which test - CELF-3 or the Adapted Version - was used
2. the difference in improvement between retesting in CELF-3 and retesting using an Adapted Version (only 1 point) was negligible

The study strongly suggests that there is no real need for an Adapted Version of the CELF-3 test to be applied as a substitute for retesting purposes. The same CELF-3 test can be confidently used for both diagnostic and re-evaluation purposes.

INTRODUCTION

Standardised assessments like the CELF-3 test are used as an aid in evaluation and re-evaluation of school-age children with language skill deficits. When given at regular intervals, these assessments gauge the effectiveness of intervention programs created to help these children. Basically, standardised assessments are formulated by experts in the field and given under set conditions. The aim is to determine the ability of children (as well as adolescents and young adults) through a score. However, as these assessments are commonly administered at regular intervals, the rationale appears to be that using an identical test two or more times over a period of time may distort the outcome, giving a false reading of a child's progress in oral language skills.

This notion is grounded in the belief that a better performance on a *repeated*, *identical* test may be due to other variances and that the results will therefore not be accurately gauging whether a child has truly improved in specific oral language areas. Variances could include the intervention during testing of long and short term memory, motivation and test specific practice skills. For this reason, speech pathologists have expressed concern that repeating the same CELF-3 assessment at regular intervals may not be a viable way to gauge student progress in oral language skills. It has been suggested, therefore, that adapted versions, faithful to the original, be formulated and used in cases where a student needs to be re-assessed.

There is a consensus among psychologists that testing is a viable and important tool in gauging people's abilities. Fremer and Wall highlight the key value of testing as it provides "valuable information for decision makers in educational, employment and clinical settings," (*Fremer and Wall, 2003, p.3*) They also pinpoint the specific value of testing for diagnostic purposes: "Test results help educators, counsellors and other professionals plan individualised education programs for students or point out specific misconceptions or problem areas that hinder progress." (*Fremer and Wall, 2003, p.6*). This is a message promoted by the US Department of Education. In its 2001 paper 'Using data to influence classroom decisions,' the Department declared that

ongoing testing was a pivotal tool “to mark progress and highlight weakness.” (*US Department of Education, 2001, p.3*).

Interestingly there is concern, particularly in the US where testing is common practice in practically every area of education and employment, that test results may not necessarily gauge the subject’s true ability. Harris expresses concern that schools could train students to perform well by “teaching to the test”, leading to test results that provide inaccurate readings of a student’s academic, intellectual or other abilities. Harris warns that teaching to the test “strips the assessment of its value and short changes the education of students.” (*Harris, 2003, p.3*) This concern may be valid to my research as it suggests that repeating the same type of test questions, as teachers presumably do when they ‘teach to the test’, can tamper with the outcome, giving a false finding.

In the CELF-4 ‘Examiner’s Manual’, Semel, Wiig and Secord noted that “re-evaluating with the same test may raise concerns about practice effects” (*Semel, Wiig, et. al., 2003, p.13*). They define ‘practice effects’ as “a gain in score points from test to retest…a result of learning from the administration of the initial test, not learning new information since administration of the initial test” (*Semel, Wiig, et. al., 2003, p.13*). However, the authors also concede that an improvement in test results, particularly in young children, following retesting could also be due to other factors such as normal intellectual and other development: “With young children, rapid language acquisition can produce real score gains due to further development of language skills” (*Semel, Wiig, et. al., 2003, p.13*). Semel, Wiig et. al. also observe that there has been no research conducted to date to gauge the minimal test-retest interval needed to eliminate most of what they label “intervening events”, such as practice effects. (*Semel, Wiig, et. al., 2003, p.13*).

Reeve and Lam also address the problem of practice effects during retesting and hence the problem of setting the same test. They point out that if the same test produces false results because of intervening factors or variables such as practice, the outcome is negating educators’ and others’ “assumption of “invariant measurement operations across unit of observations, time and conditions.” (*Reeve and Lam, 2005, p.536*) However, after conducting research to investigate this problem, Reeve and Lam concluded that “practice does not alter the nature of latent ability constructs assessed by ability tests” (*Reeve and Lam, 2005, p.536*) Reeve and Lam tested 158 undergraduate students at a US university, exposing all students to the same test over the same time intervals of one week. The same test was applied three times. Reeve and Lam concluded that: “the reliabilities of the composite factor scores do not change appreciably over repeated administrations.” (*Reeve and Lam, 2005, p.545)* The writers also make an interesting observation that not only vindicates reassessment using the same test, but also promotes it as possibly beneficial to the subjects. They cite Anastasi (1981) who suggested that “brief practice (e.g. an example set of items) may increase the construct validity of ability tests by increasing familiarity and reducing confusion and anxiety” (*Reeve and Lam, 2005, p.546)*

This raises an important consideration: could retesting using the same test be aiding rather than hampering the outcome, giving us a better reading of a child's ability as it eliminates negative variables like anxiety? Another relevant point raised by Reeve and Lam is that although the retest results in their study showed a marked improvement over the three tests, the higher score was due to "non-ability factors…not associated with the criterion" that the test was based on. In other words, if the score was improved due to memory or other "non-ability" factors, the result is still reliable as these factors are outside of the areas of ability being tested. (*Reeve and Lam, 2005, p.545)* The authors concede that these non-ability factors are important nonetheless and point out that further research is needed in this area. Ultimately, in relation to my hypothesis and investigation, Reeve and Lam's findings suggest that it does not matter whether retesting of a subject is conducted using the same test. Although there were improvements in their test results following retesting, it appears that the higher score was due to factors external to the abilities being assessed through the test, and hence irrelevant to the outcome.

Finally, it needs to be pointed out that there appears to be no research or literature available that could guide me in regard to the vital question: does an adapted test produce different results from the original? Yet, it is an important question as those recommending the replacement of the original CELF-3 with an adapted version for the purpose of retesting children are basing their view on the assumption that an adapted test will produce identical results to the original as it tests the same abilities. Otherwise, if the outcome differed for whatever reason then the results would be invalid.

The present investigation aims to establish that an adapted CELF-3 test produces the same results as the original during retesting and that it makes no difference if an identical CELF-3 test is conducted at regular intervals.

I therefore predict that there is no need for an Adapted Version of the CELF-3 test to replace the original for retesting purposes.

METHOD

Design: I used the CELF-3 sub-test 'Word Structure' and created an adapted version, formulating questions similar to those in the original sub-test, only changing the key words and the accompanying pictures. For instance in the Objective Pronouns section I changed the original **"The girl has a notebook. The notebook belongs to… (her)"** to **"The girl has a ball. The ball belongs to… (her)."**

Although the CELF-3 for the age group tested (6 to 8 years) incorporates a total of 6 sub-tests, I could only select one sub-test, given the time restrictions I was compelled to work within. The 'Word Structure' sub-test assesses elements of Expressive Oral Language. The questions cover a variety of syntax and grammar related topics such as Pronouns, Verb Tenses, Nouns and Adjectives.

There are 32 questions in the CELF-3 sub-test and I set 32 questions in my Adapted Version. The original test includes three 'trials' prior to the actual question as well as 'demonstration' items to familiarise students to the focus and concerns of a particular question, helping the child to "determine the morphological rule targeted." *(Semel, Wiig, et. al., 2003, p.14)* I formulated similar trials and demonstration items for the adapted version. The CELF-3 test provides three assessment options – 1 for a correct response, 0 for an incorrect response and NR for no response. I followed this assessment procedure for both the CELF-3 test and the Adapted Version.

Participants: There were 20 children selected at random. They are students in two Year 1 classes at a north-western suburban Catholic primary school. Ten students were selected randomly from each class. The children are aged between 6 and 7, the youngest being 6 years and 2 months and the oldest 7 years 5 months. There are 10 males and 10 females. Four of the children have been identified with oral language difficulties through speech pathology assessments. Two are currently seeing a speech pathologist and the school has a variety of intervention programs in action to help these children both at home and at school. The other two children have just been assessed and intervention has not yet started. Additionally, these two children appear to fit the criteria of Severe Language Disorder (SLD). One of the two children already undergoing intervention also has an SLD. Finally, one child in the test group has been diagnosed with Asperger's Syndrome, 3 children were previously involved in a reading recovery program and 2 children wore glasses - one child has a turned eye (see Table 1).

Materials: I used the CELF-3 assessment tool, referring specifically to the 'Word Structure' sub-test (Appendix 1), including the coloured illustrations in the stimulus manual. The illustrations in my Adapted Version (Appendix 2) were also coloured. The CELF-3 test incorporates a section for assessment, placed in the right hand margin. The Adapted Version was a photocopy of the original test with the questions slightly altered. The CELF-3 and the Adapted Version each contained 32 questions and there is an accompanying coloured image as stimulus for each question.

Procedure: My first step was to collect detailed information about each of the 20 children - gender, age, sensory impairments, severe language disorder, previous intervention and any other relevant details (Table 1). I gathered this information from the classroom teachers and the Special Needs Coordinator. I believed this information was vital to explain the test outcomes as some factors may have influenced the results.

Table 1: Student Information

Student Number	***Student Initials***	***Sex m/f***	***Age***	***HI yes/no***	***VI yes/no***	***SLD/ID/other***	***Previous Intervention***	***Other***
1	PJLP	M	7;0	no	No	No	no	No
2	OV	F	6;4	no	No	No	no	No

3	EB	F	7;0	no	Glasses-turned eye	No	no	No
4	MH	M	7;5	no	No	No	no	No
5	JS **ABSENT**	M	6;11	no	No	No	no	No
6	J O'C	M	7;2	no	No	No	no	No
7	C DM	F	6;8	no	No	No	no	No

Student Number	***Student Initials***	***Sex m/f***	***Age***	***HI yes/no***	***VI yes/no***	***SLD/ID/other***	***Previous Intervention***	***Other***
8	BS	F	6;2	no	No	No	no	No
9	MO	F	6;7	no	No	No	no	No
10	CG	F	7;0	no	No	No	no	No
11	OC	M	7;3	no	glasses	Possible SLD (waiting on psych Ax)	SP-oral language Ax – no intervention as yet Reading Recovery	No
12	JK	M	7;1	no	No	Possible SLD (waiting on pscyh Ax)	SP- oral language Ax- no intervention as yet Reading Recovery	No
13	RM	M	6;9	no	No	No	no	No
14	N DL	M	6;11	no	No	No	SP-oral language; intervention at school & home Reading Recovery	No
15	IS	F	6;9	no	No	No	no	No
16	JW	M	6;3	no	No	No	no	No
17	JG	F	7;1	no	No	SLD	SP- oral language Ax; intervention at school & home	No
18	IEH	F	6;5	no	No	No	No	No
19	ASB	F	7;2	no	No	Asperger's	No	No
20	AM	M	7;0	no	No	No	No	No

I used the same procedure with each child. I tested the children individually. Each child came to the room alone. I specifically asked for a quiet setting, to minimise any distractions that could have affected the outcome. I tested all 20 children individually using the CELF-3 Word Structure sub-test. At the end of this session I gave the children a one-hour break. Each child then returned to the test room in the same order as in the first session. I administered the CELF-3 sub-test to the first ten children and the Adapted Version to the last ten. There was no time limit set for each test. As specified in the CELF-3 manual, I repeated a question if the child appeared to be finding difficulty answering (*Semel, Wiig, et. al., 2003, p.24).* Also, as stipulated in the CELF-3 manual, I pointed to the related section of the stimulus material as I read out each question (*Semel, Wiig, et. al., 2003, p.24).*

Before each test in both sessions, I spent about a minute talking casually to each child to set them at ease. Using a positive, enthusiastic tone, I described the tests as a 'language activity'. Before the second test, each child was told that the activities were going to be 'the same' or 'similar' depending on which test (CELF-3 or Adapted) the child was going to take.

In regard to gauging the result for each test, I gave a numerical score out of 32 for each child. In other words I used a raw score. In a CELF-3 test, one is able to convert a raw score into a normed score, and I did this for the 10 students who repeated the CELF-3 test. However, as I could not do this with the10 students who sat for the adapted version, I used only the raw scores to interpret the results.

One child in Group 1 was absent for both sessions (number 5 on the list), so in effect, I only tested 19 students. I was unable to get a replacement as, due to time restrictions, I was not able to organise for a Parent Consent Form to be sent out and returned. As my results were based on an <u>average</u> score per group per test, I divided the total scores by 9 for Group 1 and by 10 for Group 2.

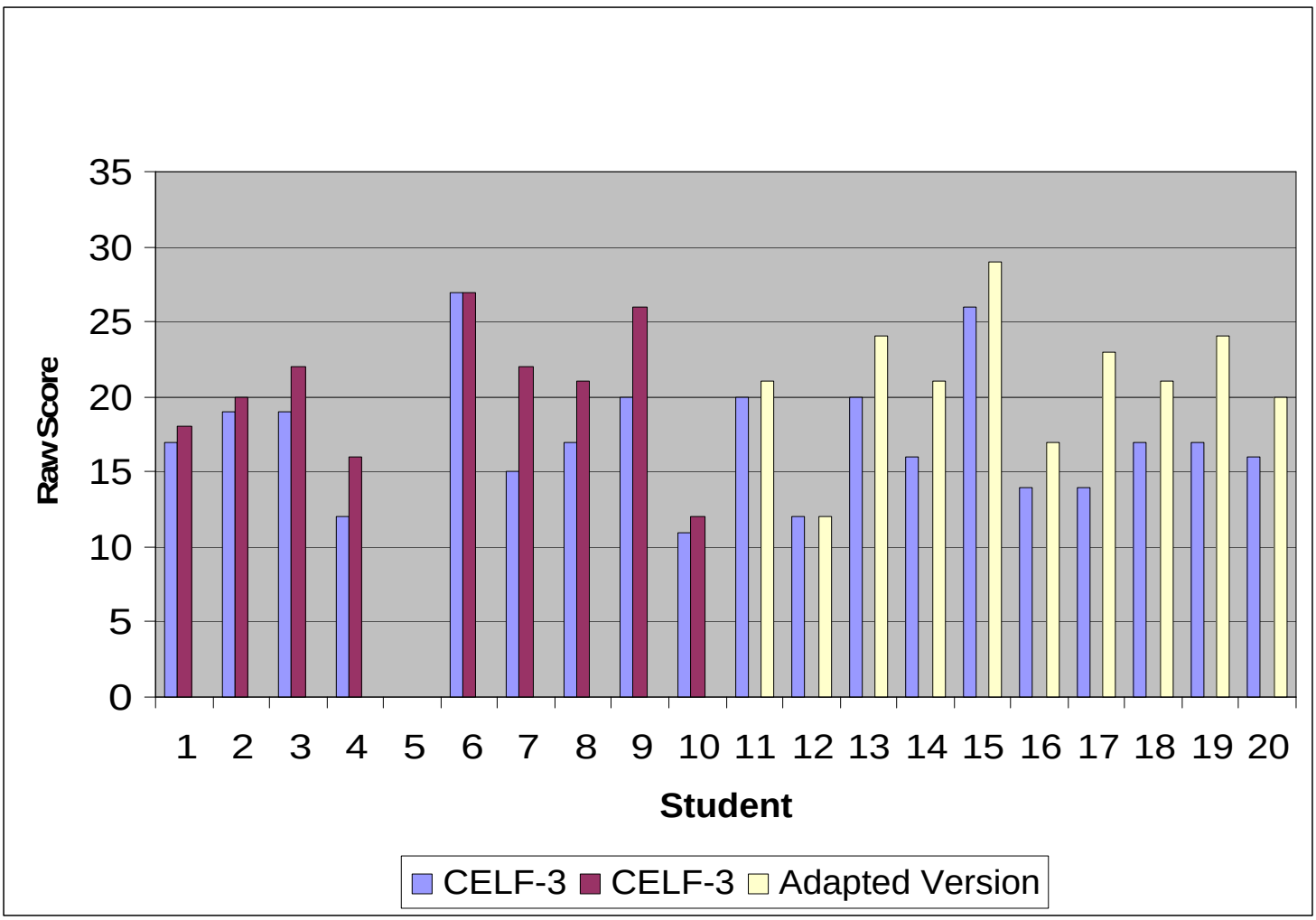
Another important factor is that only one child in the first group had an identifiable issue that may have affected the result – she had a turned eye. However, in the second group there were 5 children with a language or sensory difficulty.

I collated the raw scores for each child (out of 32) and placed them in a Table (Table 2). I then created a bar graph, using the colour blue for the first set of CELF-3 scores, purple for the second set of CELF-3 scores and the colour yellow for the Adapted Version scores (Chart 1). The average score per group per test was then calculated (Table 3).

RESULTS

Table 2: Raw Score (out of 32) outcomes of Sub-Tests administered.

Student Number	*CELF-3*	*CELF-3*	*Student Number*	*CELF-3*	*Adapted Version*
1	17	18	**11**	20	21
2	19	20	**12**	12	12
3	19	22	**13**	20	24
4	12	16	**14**	16	21
5	-	-	**15**	26	29
6	27	27	**16**	14	17
7	15	22	**17**	14	23
8	17	21	**18**	17	21
9	20	26	**19**	17	24
10	11	12	**20**	16	20



Group 1 (students 1-10)	**CELF-3**	**CELF-3**
Average Score (out of 32)	17.44	20.44
Group 2 (students 11-20)	**CELF-3**	**Adapted Version**
Average Score (out of 32)	17.20	21.20

In the CELF-3 test (Session 1), out of the maximum score of 32, Group 1 averaged a score of **17.44** (total score of 157 divided by 9 students). In the CELF-3 test (Session 1) Group 2 averaged a score of **17.20** (total score of 172 divided by 10 students).

In the identical CELF-3 test (Session 2), out of the maximum score of 32, Group 1 averaged a score of **20.44** (total score of 184 divided by 9 students). In the Adapted Version (Session 2), Group 2 scored an average of **21.20** (total score of 212 divided by 10 students.)

In the CELF-3 test (Session 1), Group 1 achieved a score of 0.24 higher than Group 2. In Session 2, Group 1 (CELF-3 repeated) achieved a score of 0.76 lower than Group 2 (Adapted Version).

Group 1 scored an average of **3 points higher** when re-sitting the CELF-3 test (from 17.44 to 20.44). Group 2 scored an average of **4 points higher** when re-sitting using the Adapted Version (from 17.20 to 21.20).

These results indicate that there was a significant improvement in performance when students sat for the Oral Language Assessment a second time, whether it was in the CELF-3 or the Adapted Version. Students in Group 2, who sat for the CELF-3 followed by the Adapted Version, performed slightly better, scoring an average of 1 point higher over the two tests than students in Group 1.

These results support my hypothesis as they indicate that the performance of students does not change markedly whether a CELF-3 test is repeated for the second testing or an Adapted Version is used as a substitute. As my hypothesis stipulates, it makes no difference whether the CELF-3 is used, or an Adapted Version, when assessing students' progress and as such, the CELF-3 test could be confidently used for all further testing procedures without the need to set up alternatives.

DISCUSSION

The study indicates that it makes no difference if a child is retested using the original CELF-3 test or the Adapted Version. This is clearly shown if one compares the results of Group 1 and Group 2.

The scores for the first session where all 19 students sat for the same CELF-3 test show minimal difference between the two groups as far as performance in Language Skills. Group 1 scored an average of 17.44 points out of 32 while Group 2 scored 17.20 points out of 32. The difference was a mere 0.24 points, showing that the ability of both groups in Language Skills was virtually identical.

This is an interesting result as in Group 2 there were 5 children with a sensory, social language or oral language difficulty whereas in Group 1, there was only 1 child with an identifiable issue. I was expecting a significantly lower score for Group 2 in both tests. Although an investigation of the possible reasons for this unexpected result are not within the scope of this study, it is important to mention that the results for Group 1 may be closer to those of Group 2 because there are students within Group 1 who may not have been identified as having oral language or sensory difficulties. It is telling that in Group 1, Student 10 had the lowest score of all 19 subjects yet the student had no record of any identified difficulties. Additionally, the performance of the 5 special needs children in Group 2 could have been enhanced by their previous experience of language assessments which may have reduced factors such as anxiety levels and practice effects likely to shape the outcome.

Overall, the two groups were surprisingly close even at the higher performance level, with Student 6 in Group 1 scoring well above average – (27 points out of 32 in both first and second sessions of the CELF-3) and Student 15 in Group 2 also scoring well above average (26 and 29 on the CELF-3 and the Adapted Version respectively).

The most important discussion in regard to this study is the close result achieved between the groups' performances on the CELF-3 and the retesting on CELF-3 and the Adapted Version. Group 1 showed an improvement of 3 points over the same CELF-3 test and retest while Group 2 showed a rise of 4 points between the CELF-3 test and the Adapted Version. The difference was only one point, indicating that:

3. there was an improvement shown by both groups after retesting whether it was CELF-3 or the Adapted Version (see discussion below)
4. the difference in improvement between retesting in CELF-3 or the Adapted Version (only 1 point) was negligible

These results would show that there is no need to be concerned about the risk of a repeated CELF-3 producing an unreliable or invalid outcome. The belief that repetition of an identical test may distort the outcome, hampering accurate diagnosis of a child's progress in language skills, appears to be

unfounded. This has been a point of contention both here in Australia and in the US where researchers, educators and other specialists have criticised the 'teaching to the test' methods in schools as it is believed that retesting using the same test produces results that do not truly gauge the student's ability or progress. My study appears to negate this view as it reveals no marked difference between a child's performance at the first or original test and his or her performance whether it is an identical or an adapted version of that test that is used.

Admittedly, the time interval (one hour) in this study was dramatically shorter than in real life where the time between the first CELF-3 test and a repeated test could be as long as 12 months. However, the much shorter time interval used in my study could be perceived as a positive factor. It eliminates at least some of the possible variances or 'practice effects' such as long term memory, intervention and natural language acquisition over time that speech pathologists and educators believe could interfere with an accurate assessment of the child's progress. These are what Semel, Wiig et al referred to as "intervening events." (*Semel, Wiig, et. al., 2003, p.13*) Still, the time interval I worked with fails to eliminate other variances such as short-term memory and test specific practice skills that may distort the outcome. In fact, the marked improvement in results from Group 1 and Group 2 during the second test (3 and 4 points increase respectively) would support the notion that intervening factors like test practice skills and memory may influence the subject's performance during retesting.

Overall, my findings appear to support the work done by Reeve and Lam in the US (see Introduction) who concluded that applying the same test a number of times is a valid method of assessment as the practice variant does not distort the "nature of the latent ability constructs" being tested (*Reeve and Lam, 2005, p.536*). Reeve and Lam also suggested that retesting using an identical test may even prove a bonus as it eliminates factors like anxiety and confusion, factors that may hamper a student from producing the optimum result. In my study, it appears that retesting using the same CELF-3 test may have eliminated some of the anxiety experienced by students in Group 1, and, added to other variances like short term memory, helped to produce a better result.

From what Reeve and Lam stipulate, however, one would conclude that the students in Group 2 would not have had their anxiety alleviated as they did not sit for the same test but a version of it. Yet, the results, quite unexpectedly, show a marked improvement in results between the test in CELF-3 and the retest in the Adapted Version. Is this negating the views of Reeve and Lam? I could explain this apparent anomaly by reiterating that I deliberately set about to make the students feel at ease. Ultimately, however, the anxiety factor appears to be negligible when we consider that:

1. the content of the Adapted Version is so similar to the original – students are not sitting for a 'new' test, merely one that is slightly different
2. the difference in improvement between Group 1 and Group 2 over the two tests was only 1 point (Group 1 an increase of 3 points to Group 2 an increase of 4 points)

My findings would suggest that educators, speech pathologists and other professionals working with children in the language skills area can be assured that they can reliably retest students using the same language assessment tool. The study indicates that there is no real need for an Adapted Version of the CELF-3 test to be applied as a substitute for retesting purposes. The same test can be confidently used for both diagnostic and re-evaluation purposes.

Nonetheless, there is a need to explore this area further, perhaps through conducting research that:

1. tests children over more than one retesting session
2. uses a longer time interval
3. tests a greater number of children
4. uses more than one sub-test

Ultimately, I still believe that such research would produce very similar results to the ones in this study.

REFERENCES

Fremer, J. & Wall, J. (2003). Why use Tests and Assessments?. In, Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators. U.S: North Carolina.

Harris, W.G. (2003). Current Issues in Educational Assessment: The Test Publisher's Role. In, Measuring Up: Issues for Teachers, Counselors, and Administrators. U.S: North Carolina.

Reeve, C.L. & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence,* 33 535-549.

Semel, E., & Wiig, E.H., & Secord, W.A. (1995) Clinical Evaluation of Language Fundamentals 3: Third Edition. The Psychological Corporation. Harcourt Brace & Company. United States of America.

Semel, E., & Wiig, E.H., & Secord, W.A. (2003) Clinical Evaluation of Language Fundamentals 4: Fourth Edition. The Psychological Corporation. Harcourt Brace & Company. United States of America.

U.S. Department of Education. (2001). Using Data to Influence Classroom Decisions. In, No child left behind Act. U.S.

APPENDICES

Appendix 1: CELF-3 Assessment Form

Appendix 2: Adapted Version Assessment Form and Illustration Stimuli

Appendix 3: Examples: Student 1 Assessment Form
Student 11 Assessment Form

This document was created with Win2PDF available at http://www.daneprairie.com.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.