Evidence of Reliability

CELF 5 METALINGUISTICS

The reliability of a test refers to its accuracy, consistency, and stability of test scores across situations (Anastasi & Urbina, 1997; Sattler, 2008). More specifically, reliability refers to the consistency of scores obtained with the theoretical concept of repeatedly testing the same student on the same test under identical conditions (including no changes to the student). Although this can never be done, various estimates of reliability are obtained in practice.

The difference between a student's hypothetical true score and the student's obtained score is called measurement error. Measurement error consists of both systematic and random errors. A reliable test will have relatively small random measurement error and provide consistent scores within and across administrations. The reliability of a test score and systematic error should always be considered in the interpretation of obtained test scores and differences between a student's test scores on multiple occasions. The reliability of CELF®-5 Metalinguistics was evaluated using internal consistency, test-retest stability, and inter-scorer reliability.

Evidence of Internal Consistency

One type of estimated reliability is internal consistency. Internal consistency reliability measures how consistently the items in the domain tested (e.g., a single test or a group of tests) measure one construct. Internal consistency reliability coefficients are used to describe the homogeneity of the items in a test.

The internal consistency of the CELF-5 Metalinguistics test and composite scores was examined using the split-half method. The split-half reliability coefficient is the correlation between the total scores of the two half-tests, corrected by the Spearman-Brown formula for the full test (Crocker & Algina, 1986; Li, Rosenthal, & Ruben, 1996). The composite score internal consistency reliability coefficients were calculated with the formula recommended by Guilford (1954), Nunnally and Bernstein (1994), and Brennan (2006).

As the data in the table below indicate, the average reliability coefficients of the CELF-5 Metalinguistics tests for the normative sample range from .80 (Conversation Skills) to .99 (Metalinguistics Profile).

CELF-5 Test	Average Reliability Coefficients (across target ages)
Metalinguistics Profile	.99
Making Inferences	.81
Conversation Skills	.80
Multiple Meanings	.90
Figurative Language	.89

Acceptable Good Excellent

Reliability of Index Scores

The average reliability coefficient is good for the Meta-Pragmatics Index (.86) and excellent for the Metalinguistics index (.94) and the Meta-Semantics index (.93).

CELF-5 Metalinguistics Composite Scores	Average Reliability Coefficients (Ages 9:0–21:11)
Metalinguistics Index	.94
Meta-Pragmatics Index	.86
Meta-Semantics Index	.93

Acceptable Good Excellent

Reliability coefficients by age are reported in the CELF-5 Metalinguistics Technical Manual.

Pearson

Evidence of Reliability

CELF-5 METALINGUISTICS

Evidence of Reliability for an English as a Second Language Group

The original research plan for this study included two groups of students who learned English as a Second Language (ESL)—those who had been speaking English for 5 years or less, and those who had been speaking English for more than 5 years. Once data collections began, it was evident that the group of students speaking English for 5 years or less could not speak English well enough to take the test. Therefore, collection of this sample was terminated. For this reason, the ESL study includes only students who had been speaking English for more than 5 years.

Evidence of internal consistency reliability was obtained from a sample of 33 students ages 9:0-21.11. Each student in the sample was identified by his or her parent as an English Language Learner (ELL) or English as a Second Language speaker, or the student was identified as a bilingual language speaker who was raised in a home where a language other than English was used frequently in everyday conversation. All students spoke English well enough to take the CELF-5 Metalinguistics test in the standardised fashion, and had been speaking English for more than 5 years. Detailed demographic information about the sample can be found in the CELF-5 Metalinguistics Technical Manual.

Ranging from .85 to .91, the internal consistency reliability coefficients of the tests for the ESL group were good to excellent. The test reliability coefficients of the ESL group are either higher than, or similar to, the coefficients reported for the normative sample. This indicates that the CELF-5 Metalinguistics is reliable for measuring the metalinguistic skills of students who have learned English as a second language and who speak English well enough to take the test in the standardized fashion.

CELF-5 Metalinguistics Test	English as a Second Language (n=33)
Making Inferences	.85
Conversation Skills	.85
Multiple Meanings	.91
Figurative Language	.89

Acceptable Good Excellent

Evidence of Reliability for Clinical Groups

Evidence of internal consistency reliability for clinical groups was obtained from a sample of 86 students in two groups: a language disorder (LD) group and an autism spectrum disorder (ASD) group.

The table below provides internal consistency reliability coefficients of tests for both of the two clinical groups who took the test. This table shows that the test reliability coefficients of most of the clinical groups are either higher than or similar to those coefficients reported for the normative sample, which suggests that CELF-5 Metalinguistics is equally reliable for measuring the language skills of students from the general population or students with clinical diagnoses.

Complete information about the studies can be found in the CELF-5 Metalinguistics Technical Manual.

Clinical Group Test	Language Disorder (*n*=54)	Autism Spectrum Disorder (*n*=32)	Average r_{xx}
Metalinguistics Profile	.99	.98	.99
Making Inferences	.82	.88	.85
Conversation Skills	.80	.96	.91
Multiple Meanings	.93	.95	.94
Figurative Language	.89	.94	.92

Acceptable Good Excellent

Evidence of Reliability

CELF 5 METALINGUISTICS

Evidence of Test-Retest Stability

Another way of estimating the reliability of a test is to examine its test-retest stability. Test-retest stability is the correlation between the test and retest scores and is a direct measure of test stability for repeated testing. To examine retest stability, the student is given the same test twice, each time under conditions that are as similar as possible. The student will not perform exactly the same way in each of the two test sessions. The time interval between the test and retest is as short as possible, to minimise changes in the individual, yet long enough for any practice or memory effects to dissipate.

The CELF-5 Metalinguistics test-retest reliability was evaluated in a study in which the test was administered to a group of students on two separate occasions and then the scores were compared. The sample used to assess the stability of CELF-5 Metalinguistics scores over time included 68 examinees (ages 9:0 to 21:11, with a mean age of 13.5 years) selected from the standardisation sample. The sample included 38 females and 30 males. In the sample, 50% of the students were white, 27.9% were African American, 10.3% were Hispanic, 3% were Asian, and 8.8% were students of other races/ethnic origins. The parent/caregiver education level of the sample was as follows: 10.4% had no high school diploma or GED, 27.9% had a high school diploma or GED, 33.8% had some college or technical school, and 27.9% had a bachelor's degree or higher.

After being tested as part of the standardisation study, these students repeated the test within a range of 7 to 30 days (mean of 16.3 days), with both tests administered by the same examiner. The test-retest reliability was estimated using Pearson's product-moment correlation coefficient. As the data indicate, the CELF-5 Metalinguistics test scores possess adequate stability across time for the two age groups tested (9:0 to 12:0 and 13:0 to 21:11). The average corrected stability coefficients for all ages for the Making Inferences and Conversation Skills tests are adequate, at .72 and .73, respectively. The average stability coefficients are good for the Metalinguistics Profile (.85) and the Multiple Meanings and Figurative Language tests (.87). The average corrected stability coefficients of the composite scores are shown below.

Index Score	Corrected *r*
Total Metalinguistics Index	.89
Meta-pragmatics Index	.73
Meta-semantics Index	.91

Acceptable Good Excellent

This test-retest study was conducted to evaluate stability of test scores. The shortest test-retest interval that will not result in significant practice effects on CELF-5 Metalinguistics has not been determined; however, one or more of the following criteria should be met before CELF-5 Metalinguistics is administered again:

1. Retesting should be conducted after the student no longer remembers the test items and/or his or her responses when tested previously.
2. Retesting should be conducted when the examiner thinks the child has made progress since the previous test administration; otherwise, there is no reason to retest.
3. Retesting can be conducted when the student's age at testing requires the next-age norms table to score.
4. Retesting can be conducted when other factors negatively affecting the student's performance (e.g., illness, inattention) cause you to question the accuracy of previous test results.

Evidence of Reliability

CELF 5 METALINGUISTICS

Evidence of Inter-Scorer Agreement

The CELF-5 Metalinguistics tests require familiarity with different scoring criteria that require clinical judgment and qualitative and quantitative judgments about student responses. Familiarity with varieties of American English is also important, so the clinician can identify the variety of American English spoken by the student being tested. Because there is room for interpretation on the subjectively scored items, it is necessary to evaluate the extent to which these interpretations are consistent from one scorer to another. Scoring rules were developed for the Making Inferences, Conversation Skills, Multiple Meanings, and Figurative Language tests and scorers were trained in applying the rules before the standardisation protocols (i.e., record forms) were scored. The CELF-5 Metalinguistics tests were scored by a team of four trained scorers under the supervision of the test developers. To ensure accuracy of scoring before analysis of the test data, two different scorers from the team were randomly selected to score each protocol independently. The scores were compared, and a third independent scorer resolved any differences.

Double scoring for the Multiple Meanings and Figurative Language tests continued for three weeks. The average scorer agreement during this period was .95. Item-level agreement rates were used in the analysis. After this period, responses on the Multiple Meanings and Figurative Language tests were scored by a single scorer, with checks to prevent scorer drift.

Due to the complexity of responses on the Making Inferences and Conversation Skills tests, double-scoring with resolution was continued for a longer period. To determine inter-scorer reliability, reliability coefficients were calculated according to appropriate intra-class correlation procedures. Total test raw scores were used in the analysis. Inter- scorer reliabilities were .95 for Making Inferences and .90 for Conversation Skills.

These results demonstrate that although these tests require more judgment, they can be scored reliably. Additional information about internal consistency, standard error of measurement, test-retest stability, and inter-scorer agreement can be found in the CELF-5 Metalinguistics Technical Manual.

For more information about CELF-5 Metalinguistics, please visit Pearsonclinical.co.uk/meta

Pearsonclinical.co.uk