



October 2023

Standardized Versus Naturalized: An Evaluation of Child Morphological and Syntactic Assessments

Jaclyn Shurman
jaclynshurman@gmail.com

Dorothy Leone
dleone@iona.edu

Follow this and additional works at: <https://digitalcommons.cwu.edu/ijurca>

Recommended Citation

Shurman, Jaclyn and Leone, Dorothy (2023) "Standardized Versus Naturalized: An Evaluation of Child Morphological and Syntactic Assessments," *International Journal of Undergraduate Research and Creative Activities*: Vol. 7: Iss. 1, Article 3.

DOI: 10.7710/2168-0620.1046

Available at: <https://digitalcommons.cwu.edu/ijurca/vol7/iss1/3>

This Article is brought to you for free and open access by ScholarWorks@CWU. It has been accepted for inclusion in International Journal of Undergraduate Research and Creative Activities by an authorized editor of ScholarWorks@CWU. For more information, please contact scholarworks@cwu.edu.

Standardized Versus Naturalized: An Evaluation of Child Morphological and Syntactic Assessments

Peer Review

This work has undergone a double-blind review by a minimum of two faculty members from institutions of higher learning from around the world. The faculty reviewers have expertise in disciplines closely related to those represented by this work. If possible, the work was also reviewed by undergraduates in collaboration with the faculty reviewers.

Abstract

Speech-language pathologists may choose to evaluate children's language using standardized or naturalized assessments. This study investigated if the *Clinical Evaluation of Language Fundamentals-Preschool 2 (CELF-P 2)*, a standardized assessment, and language sampling, a naturalized assessment, reveal the same information about children's linguistic competence and performance. Children ages 3.0-7.0 were assessed with specific focus on morphology and syntax. The participants completed four morphosyntactic-based subtests of the *CELF-P 2*. Additionally, play-based interactions, used to elicit natural language, were video-recorded. The *CELF-P 2* was scored and language samples were transcribed and analyzed. Mean length of utterance (MLU) scores showed a slightly more variable trend around the mean than *CELF-P 2* scores and there were no significant correlations between the two assessments. Furthermore, the two forms of assessment produced incongruous age equivalents for 66% of the participants (four out of six) and participants produced different morphosyntactic structures during each type of assessment. Thus, results indicated limitations and successes of the different assessment approaches. When used alone, either form of assessment did not provide a completely accurate representation of children's language acquisition. However, when used in conjunction, the two assessments may represent the linguistic competence and performance of children more accurately.

Keywords

morphology, syntax, language acquisition, standardized, naturalized

Acknowledgements

We would like to acknowledge the Iona College Honors Council which provided financial support via an American Express Grant to fund the necessary tools to complete the research. We would also like to acknowledge the participants and families who were a part of the study.

Editor's Note: Dr. Dorothy Leone, Ph.D., CCC-SLP, Assistant Professor, Department of Speech Communication Studies, Iona College, served as the faculty mentor for this work.

INTRODUCTION

Clinicians utilize a variety of evaluation tools to analyze and document a child's language development. According to Condouris, Meyer, & Tager-Flusberg (2003) assessments are used to measure language skills, identify children with language impairments, and monitor the efficacy of treatment. Assessments account for a great amount of a speech-language pathologist's (SLP) professional time and help justify decisions of treatment for clients. As noted by Huang, Hopkins, and Nippold (1997), SLPs spend "21% of their work time in evaluation, indicating that evaluation is a major professional activity" (p.12).

Methods to measure language acquisition include both naturalized and standardized assessments. While naturalized assessments provide insight into a child's language use in everyday settings and allow clinicians to judge a child's overall communication development, standardized assessments are highly structured, objective, and help clinicians to identify impairments through comparison to a normative sample (Tyler & Tolbert, 2002). Spontaneous language samples can be collected using conversation, free play, story generation, and interviews (Evans & Craig, 1992; Southwood & Russell, 2004). Contrary to naturalized assessments, standardized assessments follow a systematic structure (Condouris et al., 2003).

Brown (1973) provided a significant source of research for naturalized assessment by studying child language development through the transcription and analysis of spontaneous speech in a longitudinal study. Brown concluded that calculating mean length of utterance (MLU) is an appropriate tool to indicate syntactic and morphological development of preschool children. Brown developed an index of syntactical development in which

each successive stage accounts for new linguistic acquisition and grammatical growth. After the fifth stage, the reliability of MLU decreases because the production of longer utterances does not necessarily represent a child's linguistic knowledge (Johnston, 2001). MLU is a frequently-used and widespread measure of language development (Klee, Schaffer, May, Membrino, & Mougey, 1989). Furthermore, as highlighted by Miller and Chapman (1981), MLU is a good predictor of developmental achievement.

MLU, as defined by Rice et al. (2010), is the number of morphemes, or the units of meaning, in each of a young child's natural utterances. Brown's grammatical morphemes appear throughout a young child's life and help shape the meaning of the child's utterances.

As a child transitions through Brown's stages of morphological development, utterances and sentence structure become more complex. Remarkably, children's acquisition of Brown's morphosyntactical stages is accomplished sans formal instruction (Berko Gleason & Bernstein Ratner, 2013; Pinker, 1994). Rice et al. (2010) compared MLU to age-referenced normative data for typically developing children and children with specific language impairment (SLI) (ages 2.6-9.0). A correlation between age progression and MLU was revealed, suggesting that MLU yields reliable and valid results of children's language acquisition in comparison to age (Rice et al., 2010).

As opposed to the naturalized assessment of language sampling and MLU calculation, standardized assessments provide rigid structure, including subtests, which often target specific areas of language, and allow for comparison with a normative sample. Depending on the area of language targeted, the subtests may utilize

tools such as picture identification, sentence completion, imitation, and question response to elicit language (Secord, Semel, & Wiig, 2013).

Although norm-referenced standardized assessments play an important role in language evaluation because they provide an objective form of measurement, the validity of the assessments may be questioned (McFadden, 1996). Limitations of standardized assessments include the potentially restricted comparison group, time constraints, and financial burdens. A comparison group represents an ideal population which may not be linguistically comparable in gender, culture, or socioeconomic status (Huang et al., 1997; McFadden, 1996).

While opinions of standardized testing vary, many researchers have concluded that naturalized assessments and standardized assessments alone are not an adequate representation of child language development (Huang et al., 1997; Southwood & Russell, 2004; Tyler & Tolbert, 2002). Rather, standardized assessments, used in conjunction with naturalized assessments, increase reliability of language evaluations and test all the components of language. For example, Condouris et al. (2003) provided a comparison of standardized and naturalized assessments for children with autism spectrum disorder (ASD). Results of the two forms of assessment revealed comparable data indicating that both standardized and naturalized assessments test the same underlying linguistic functions. Moreover, Tyler and Tolbert (2002) provided a comparison of standardized and naturalized assessments for a single subject within a 90-minute time span and found that combining the results of the assessments was ideal.

The present study examined if the results of the *Clinical Evaluation of Language Fundamentals-Preschool 2*

(*CELF-P 2*), a standardized assessment, and language sampling, a naturalized assessment, revealed the same fundamental morphosyntactic information about linguistic competence and performance. Ultimately, although both methods assessed similar components of child language development, it was hypothesized that language sampling may reveal a more accurate representation of a child's linguistic competence and performance.

METHOD

Participants

Six participants aged 3.0 to 7.0 participated in this study. All participants were typically developing, which was verified by caregivers' completion of a background questionnaire (see Appendix A). Half of the participants were female (ages 4.2, 4.2, and 6.11) and half were male (ages 3.4, 4.8, 6.0). All participants were monolingual speakers of American English. All language samples, with the exception of participant 1, were collected in familiar settings (e.g., home). Participant 1's assessments were conducted at Iona College's Speech Communication Studies Department.

Procedure

Prior to the study, Iona College Institutional Review Board approval was obtained. The legal guardians of the participants were asked to complete a background questionnaire regarding their child's developmental and language histories (see Appendix A). Informed consent was acquired from the legal guardians prior to testing. Following, language samples were collected and four subtests of the *CELF-P 2* were administered. Total testing time was approximately two hours. Follow-up and explanation of results

with the legal guardians was completed after testing.

Language Sampling

All language samples were approximately 30 minutes, involving a play-based interaction constructed around the interests of the participants. The interactions were recorded using a video camera. The participants were allowed to move freely throughout the setting. The researcher and legal guardians engaged the participants in conversation to elicit volitional speech following the principles of milieu teaching where the adults followed the child's attentional lead (Paul & Norbury, 2012). The adults utilized the participants' choice of toys, which included, for example, board games, dolls, and toy cars, to elicit and maintain communication. Furthermore, the researcher used open-ended prompts about the child's chosen item in order to further elicit language.

The language samples were transcribed by the researcher and checked by a SLP to verify the transcription. Utterance contour, pauses greater than two seconds, and inhalation served as utterance boundaries (Miller & Chapman, 1981). Approximately 120 child utterances were transcribed for each of the participants. For 67% of the samples, the medial portion was transcribed, which is the recommended best practice for language sample analysis (Paul & Norbury, 2012). The first 120 utterances were used for two participants (4 and 6) due to the robust quality and quantity of the utterances in the beginning of the sample. MLU was computed by dividing the number of morphemes in an utterance by the number of utterances produced during the language sample (Brown, 1973). Morphology and syntax were analyzed according to Brown's index of syntactical development (see Table 5.3 of Berko Gleason & Bernstein Ratner, 2013, for complete data). For each language

sample, unintelligible phrases and hedges, such as "umms," were not included in the calculation of MLU. Incomplete phrases defined as those containing unintelligible words were not used in the calculation of MLU. All contractions except *let's*, *don't*, and *won't* were counted as two morphemes. Fillers, such as *like*, were also omitted from calculation. Reformulations, false starts, and repetitions were also not included. Phrases such as *uh oh*, *uh huh*, *nuh uh*, *yup*, and *mhmm* were counted in the calculation because they hold semantic meaning. Two SLPs also calculated each transcript's MLU. Reliability between the experimenter who calculated the original MLU and the two SLPs who later analyzed each sample was 83%. If a discrepancy occurred, an additional SLP was asked to respond and the most frequent response was selected.

CELF-Preschool 2

The *CELF-P 2* was utilized as the standardized assessment for this study. The participants were required to complete four subtests of the *CELF-P 2*: Sentence Structure, Word Structure, Recalling Sentences, and Recalling Sentences in Context. The abovementioned subtests focus on the participant's morphosyntactic development. The Sentence Structure subtest requires the examiner to read a sentence which corresponds to a set of photos. Participants choose which photograph the sentence described and thus demonstrate their receptive language abilities. The Word Structure subtest examines a participant's expressive mastery of word structure in relation to tense, comparative suffixes, derivational suffixes, possessive, and other grammatical forms. The examiner reads a sentence which corresponds to a picture. The examiner reads another sentence and the participants complete the sentence based on the preceding sentence's structure. For the

Recalling Sentences subtest, the examiner reads a sentence and the participants repeat the sentence verbatim. Success in repetition depends upon the participant's number of errors. The Recalling Sentences in Context subtest requires the participant to repeat sentences of a story. The sentences in the subtests vary in structural complexity and length. Both the Recalling Sentences and Recalling Sentences in Context assess the participant's expressive and receptive abilities.

During the evaluation, the researcher read directions to the participants prior to the beginning of each subtest. All protocols were followed as outlined in the *CELF-P 2* instruction manual. Following test administration, all individual subtest raw scores were calculated as standard scores ($M = 10$, $SD = 3$).

DATA ANALYSIS

Language Sampling

Each participant's language sample was analyzed independently. Utterances ranged in length from one morpheme (participants 1-6) to 46 morphemes (participant 6). Lengthier utterances, contributing to a higher calculation of MLU, were attributed to rambling, explanation of board games, and story-telling. Shorter utterances were attributed to yes/no questions and responses such as "okay." Specific linguistic background is listed in Table 1 for each participant. Only one MLU calculation derived from the language sample (participant 2) was judged to be an inaccurate representation of the child's linguistic competence and performance. While participant 2 produced long utterances, which created a larger MLU, the participant also expressed consistent word and sentence structure errors (e.g., substituted "them" for "they are") throughout the sample.

Table 1. Participants' Language History, MLU, and Age Equivalents

	Participant #1	Participant #2	Participant #3	Participant #4	Participant #5	Participant #6
Age	6.0	3.4	4.2	4.2	6.11	4.8
Gender	Male	Male	Female	Female	Female	Male
Testing Setting	Unfamiliar: Iona College Speech & Hearing Center	Familiar: Participant's Home	Familiar: Participant's Home	Familiar: Participant's Home	Familiar: Participant's Home	Familiar: Participant's Home
Siblings	Yes: younger brother, 3.4	Yes: older brother, 6.0	Yes: younger sister, 1.0	Yes: older sister, 6.11 & younger brother, 1.3	Yes: younger sister, 4.2 & younger brother, 1.3	Yes: younger brother, 2.10
Typical Day	School, home with dad, mom, and babysitter	Preschool, home playing with parents, brother, and babysitter	½ day school, home with mother and sister	In school, with family	In school, with parents	School (daycare)
Speech Therapy	No	No	No	Yes, as per parents' request, not specifically needed	No	No
Languages Spoken	English	English	English	English	English	English
MLU	11.50	7.73	4.09	5.05	9.38	5.46
MLU Age Equivalents	>5.0	>5.0	3.9	4.5	>5.0	4.75

CELF-Preschool 2

Results varied among subtests and participants for the *CELF-P 2*. Raw scores, scaled scores, and age equivalents (AE) for participants are listed in Table 2.

Sentence Structure. The Sentence Structure subtest assessed receptive language skills by analyzing the child's ability to understand spoken sentences. Scaled scores ranged from eight to 14. There were no congruous errors among the participants. However, four participants (2, 3, 4, and 6) inaccurately responded to items that addressed the subordinate clause (e.g., before she ate the sandwich) and the passive tense (e.g., is being followed).

Word Structure. The Word Structure subtest evaluated expressive language skills by assessing the child's ability to produce morphological markers and pronouns. Scaled scores ranged from nine to 13. All participants incorrectly identified the contractible/auxiliary copula

Table 2. CELF-P 2 Raw Scores, Scaled Scores, and Age Equivalents

Participant	Score Type	Sentence Structure (SS)	%	Word Structure (WS)	%	Recalling Sentences (RS)	%	Recalling Sentences (In Context)	%	Participant Age Equivalents	
										SS	>7.0
1	Raw	21	74	22	75	34	91	37	70	SS	>7.0
	Scaled	12		12		14			-	WS	>7.0
									79	RS	>7.0
2	Raw	15	91	16	84	13	84	31	95	SS	4.8
	Scaled	14		13		13			-	WS	4.8
									99	RS	4.0
3	Raw	17	75	18	75	19	63	17	45	SS	5.3
	Scaled	12		12		11			-	WS	5.2
									49	RS	4.8
4	Raw	16	63	17	63	19	63	23	65	SS	5.3
	Scaled	11		11		11			-	WS	5.2
									69	RS	4.8
5	Raw	21	75	23	84	27	37	31	35	SS	>7.0
	Scaled	12		13		9			-	WS	>7.0
									39	RS	6.1
6	Raw	13	25	14	37	-	-	18	30	SS	4.2
	Scaled	8		9		-			-	WS	4.2
									34	RS	-

of “They are.” Four participants (2, 3, 4, and 6) inaccurately identified the irregular past tense of “fell.” However, of those same participants, only one participant (6) inaccurately identified a subsequent irregular past tense question (blew).

Recalling Sentences. The Recalling Sentences subtest assessed expressive language skills by analyzing a child’s ability to repeat sentences without altering word or sentence structure and meaning. Scaled scores ranged from nine to 14. Participant 6 received a score of zero for incompleteness of the subtest. All participants incorrectly recalled sentences which included an active declarative with a relative clause (e.g., the dad brought a book for his son who likes funny stories). Four of the five participants inaccurately repeated sentences which included the active declarative with negation (e.g., the kindergartner cannot cross the street by himself), active declarative with noun modification (e.g., the big, brown dog ate all of the cat’s food), and active declarative with a subordinate clause (e.g.,

because tomorrow is Saturday, we can stay up late tonight).

Recalling Sentences in Context.

The Recalling Sentences in Context subtest is a supplementary subtest similar to the Recalling Sentences subtest but includes contextual cues through a story. In the subtest, five participants inaccurately repeated the sentence containing an active declarative with a relative clause (e.g., I am very happy that we finally found you, Grandma). Additionally, four participants (2, 3, 4, and 5) were unable to repeat sentences with active declarative with an infinitive clause and negation (e.g., I can’t wait to have Grandma come to our house). Furthermore, four participants (2, 3, 5, and 6) were unable to recall the active declaration with coordination (e.g., I fell and dropped my juice).

Language Sampling and Standardized Testing Comparison

Scores at an individual participant level from both the *CELF-P 2* and MLU were compared to analyze the variance

around the mean. This was done by calculating z-scores for each participant's individual score. Each participant's MLU was compared to the sample's mean because a population mean was not available. Subtest scores were compared to the population mean of the *CELF-P 2*. As shown in Table 3, MLU scores showed a slightly more deviant trend around the mean than *CELF-P 2* scores. However, most individual scores were within ± 1 standard deviation around the mean. Scores above one standard deviation were only slightly more deviant, with the highest MLU score being 1.5 standard deviations above the mean and the highest *CELF-P 2* score being 1.33 standard deviations above the mean.

Table 3. Comparison of Participants' Performance on *CELF-P 2* and MLU

CELF-P 2 Measures		MLU	Total	
		Normal*	> ± 1 SD	
Word Structure	Normal	4	2	6
	> ± 1 SD	0	0	0
Sentence Structure	Normal	3	2	5
	> ± 1 SD	1	0	1
Recalling Sentences	Normal	3	1	4
	> ± 1 SD	0	1	1

*Normal- within 1 SD of mean

In addition, Pearson's product moment correlation was used to test the relationship between *CELF-P 2* and MLU measures within the current sample. None of the correlations between MLU and the *CELF-P 2* subtests reached significance (sentence structure: $r = .36$, $p = .48$; word structure: $r = .47$, $p = .35$; recalling sentences: $r = .36$, $p = .55$).

Age Equivalents (AEs). AEs were determined based on the model provided by Miller and Chapman (1981), which denotes the age at which most children have an MLU equal to that of the children included

in the present study. Age-equivalence scores were included so that MLU and *CELF-2 P* results could be compared. Five out of six participants (83%) had MLU AEs representative of their chronological ages (CA) (see Table 1). Age-equivalence scores were reported as a reliable and age-validated measure of syntactic growth in children with and without SLI (Rice et al., 2010). Results of that study conducted by Rice et al. (2010) revealed MLU calculation is sensitive to language impairment throughout the range for which MLU is considered a reliable index.

MLU calculation and results of the Word Structure subtest produced different AEs for 66% of the participants (four of six participants). Both measures produced comparable AEs for the older participants (participants 1 and 5). For the four younger participants, MLU calculation and results of the *CELF-P 2* produced different AEs for 75% of the participants (three of four participants). Although participant 2 had a comparable MLU and Word Structure subtest score, results of the measures were not comparable to the participant's CA (3.4).

Overall, approximately 83% of the participants (five of six participants) had a MLU AE comparable to their CAs. For these five participants, MLU was indicative of their CAs and language acquisition. However, only approximately 33% of the participants (two of six participants) had a standardized assessment AE comparable to their CAs. The two participants were aged 6.0 and 6.11.

Furthermore, participant 2 (age 3.4) had a MLU AE of >5.0 and a Word Structure AE of 4.8. While these results are much higher than expected for a child aged 3.4, the results demonstrate consistency between both measures. Overall, while the MLU AE scores appeared to reflect the participant's CA, the AE results of the *CELF-P 2* indicated an older AE score for

50% of the participants (three of six participants), a younger AE for 16% of participants (one of six participants), and an accurate representation of CA for 33% of the participants (two of six participants). See Table 4 for a comparison between the forms of assessment and CA.

Table 4. Comparison of Age Equivalents: Word Structure Subtest (*CELF-P 2*) and MLU

Participant	Participant's chronological age	Computer MLU Age Equivalent	Computer Word Structure Age Equivalent
1	6.0	>5.0	>7.0
2	3.4	>5.0	4.8
3	4.2	3.9	5.2
4	4.2	4.5	4.11
5	6.11	>5.0	>7.0
6	4.8	4.75	4.2

Morphosyntactic Structures. A list of morphosyntactic categories assessed in the Word Structure subtest of the *CELF-P 2* was created. Participants' errors were categorized according to the morphosyntactic categories (e.g., future tense, preposition) listed in Table 5. Participants' language samples were analyzed based on accurate use of the abovementioned morphosyntactic structures. Responses were reported for one time accurate usage or 100% accuracy. On the subtest, specific items within categories were tested either once or twice. Categories containing only one item were classified as "used item with 100% accuracy."

A total of 17 expressive morphosyntactic categories were analyzed (see Table 5). Congruent results between the *CELF-P 2* and language samples were found for 17.6% of items (e.g., prepositions, irregular past tenses, objective pronouns) when both the one time accurate usage and 100% accuracy groups were analyzed.

Congruent results between the *CELF-P 2* and language samples were found for 29.4% of items (e.g., objective pronouns,

uncontractible copula/auxiliary, irregular past tense, third person singular, prepositions) when only the 100% accuracy group was analyzed for both forms of assessment. An increase in congruent results were found for 47.1% of items (i.e., subjective pronouns, objective pronouns, contractible copula, irregular past tense, regular past tense, progressive *-ing*, regular plural, prepositions) when only the one time accurate usage group was analyzed.

The *CELF-P 2* subtests only contained certain items from each morphosyntactic category. For example, although the *CELF-P 2* only tested a limited number of pronouns (e.g., her, him, hers, he, she, herself), the participants utilized a variety of additional pronouns in the naturalized context (e.g., I, you, yours, we, they, my, itself, etc.). Furthermore, the *CELF-P 2* tested morphosyntactic categories that were not frequently used during the language samples (e.g., noun derivation).

Overall, the participants showed a trend of producing morphological and syntactic structures during language sampling that they inaccurately produced during administration of the *CELF-P 2*. All of the participants were unable to produce the contractible copula, "they are," in the Word Structure subtest of the *CELF-P 2*. In the language sample, each participant produced some form of the contractible copula (e.g., they are, he is). Additionally, individual participants produced morphosyntactic structures during the naturalized assessment that he or she inaccurately produced during the *CELF-P 2*. For example, participant 3 was unable to produce the superlative "fastest" during the standardized evaluation but produced "biggest" during volitional speech. Some forms were produced accurately by participants during the *CELF-P 2* but contained errors during language sampling. For example, participant 2 produced the

Table 5. Percent of Participants Who Used Morphosyntactic Items Accurately on the *CELF-P 2* Subtest and Language Sample

	Subtest:		Sample:		
	Used item(s) at least 1x	Used item(s) 100%	Used item(s) at least 1x	Used item(s) 100%	No attempt of category
Prepositions	100.0	83.3	100.0	83.3	0.0
Regular Plural	N/A	100.0	100.0	50.0	0.0
Possessive Noun	N/A	83.3	66.7	50.0	33.3
Progressive –ing	100.0	100.0	100.0	83.3	0.0
Third Person Singular	100.0	66.7	83.3	66.7	0.0
Future Tense	N/A	66.7	100.0	83.3	0.0
Regular Past Tense	N/A	83.3	83.3	66.7	16.7
Irregular Past Tense	83.3	33.3	83.3	33.3	0.0
Contractible Copula	N/A	100.0	100.0	33.3	0.0
Uncontractible/Auxiliary Copula	83.3	0.0	100.0	0.0	0.0
Objective Pronoun	100.0	83.3	100.0	83.3	0.0
Possessive Pronoun	N/A	66.7	50.0	50.0	33.3
Subjective Pronoun	100.0	50.0	100.0	83.3	0.0
Reflexive Pronoun	N/A	50.0	33.3	16.7	50.0
Noun Derivation	N/A	66.7	16.7	16.7	83.3
Comparative	N/A	83.3	50.0	50.0	50.0
Superlative	N/A	66.7	33.3	16.7	33.3

irregular past tense “blew” in the Word Structure subtest, but consistently used overregularization errors in conversational speech (e.g., taked, broked).

Time Commitment. Language sampling and standardized testing were both time-consuming measures to administer, score, and analyze. Administration and scoring of the *CELF-P 2* took approximately 45 minutes per participant (20-30 minutes to administer; 20 minutes to score). Transcription and MLU calculation took approximately 223 minutes (3.7 hours) per participant. It is important to note, however, that only four of the 11 subtests of the *CELF-P 2* were administered and scored. It would take over 120 minutes (two hours) to administer and score all subtests of the *CELF-P 2*. In comparison to an analysis of the morphosyntactic subtests of the *CELF-P 2*, language sampling took approximately triple the time to administer and score.

DISCUSSION

The primary purpose of this study was to investigate if standardized and naturalized assessments produced comparable results of linguistic competence and performance of children. A larger sample size may show that the standardized and natural measures are measuring the similar language skills while providing slightly different information about the child being tested (Bornstein & Haynes, 1998; Ukrainetz & Blomquist, 2002).

Challenges with the *CELF-Preschool 2*

Inconsistency of results was noted across the subtests of the *CELF-P 2*. Discrepancies between subtests may lead to inconclusive results of a child’s language abilities. Even though the Recalling Sentences subtest and the Sentence Structure subtest both tested acquisition of syntax, the two subtests did not reveal congruent results

(e.g., participants struggled with the passive tense during Sentence Structure but accurately produced it during Recalling Sentences). Furthermore, while participant 5 scored in the 75th and 84th percentile for the Sentence Structure and Word Structure subtests, the participant scored in the 37th percentile for the Recalling Sentences subtest.

Additionally, discrepancies within an individual subtest may lead to inaccurate conclusions regarding language acquisition. For the Word Structure subtest, all participants incorrectly identified the contractible/auxiliary copula of “They are.” However, within the same context, five of the six participants all correctly identified the contractible/auxiliary copula of “She is.”

Furthermore, the complexity and appeal of completing subtests of standardized assessments may influence a child’s performance. For example, the Recalling sentences and Recalling Sentences in Context subtests challenged participants to reproduce sentences with multiple clauses (e.g., the dad bought a book for his son who likes funny stories). The length of the utterances and attentiveness of the participants may have increased the number of errors. Participant 6 refused to complete the subtest, claiming it was “boring.” Thus, the inconsistencies noted between and among subtests may lead clinicians to question the test’s results.

Age Equivalents

Inconsistency in results was noted not only across the subtests of the *CELF-P 2*, but also between the naturalized and standardized assessments. As compared to the calculation of MLU, which produced accurate AE for five out of six participants, the AE calculated from the *CELF-P 2* were only accurate for two of the six participants. The two participants that the *CELF-P 2* provided congruent results with CA were the

two oldest participants. Thus, language sampling may be more indicative of a child's true morphosyntactic abilities than standardized results for younger children; however, both language sampling and the *CELF-P 2* may provide valid morphosyntactic results for older children. The findings are congruent with Rice et al. (2010) who established age-referenced MLU as a reliable and valid measurement of language acquisition. Akin to the present study, Rice et al. (2010) documented that MLU increased with age progression.

Morphosyntactic Structures

Results of the *CELF-P 2* and measures derived from natural language samples yielded the greatest congruency when morphosyntactic structures were analyzed for accurate usage one time. The *CELF-P 2* only tests items once or twice. However, during language sampling, participants may use morphosyntactic items multiple times. Therefore, accurate usage of an item once may not reflect accurate usage of that item throughout the language sample. For example, participant 2 only used irregular past tense accurately one out of four times during the language sample; whereas the participant used irregular past tense accurately one out of two times on the *CELF-P 2*.

The abovementioned concept was evident when morphosyntactic structures were analyzed for both one time accurate usage and 100% accurate usage. More participants used morphosyntactic forms accurately one time on the *CELF-P 2* and during the language sample than with 100% accuracy throughout the naturalized assessment. Furthermore, although results may be congruent across participants, results may not reflect accurate usage of the morphosyntactic form (i.e., no participants used the uncontractible copula/auxiliary with 100% accuracy).

Standardized assessments may test forms that are not readily used during volitional speech. For example, although noun derivation is a morphosyntactic category on the *CELF-P 2*, five out of six participants did not attempt to use noun derivation during the language sample. However, although standardized assessments test a gamut of morphosyntactic forms, they also limit the variety of items tested. For example, the *CELF-P 2* only assesses a limited amount and type of pronouns. Although the participants may not have used the tested pronouns (e.g., her, him, hers, he, she, herself) accurately on one or both forms of assessment, generally, he or she utilized a greater variety of pronouns during the language sample (e.g., I, you, yours, we, they, their, my, it, itself, etc.) than standardized testing.

Overall, the participants produced certain morphological and syntactic structures during language sampling that they inaccurately produced during the standardized assessment. Furthermore, the participants did not always produce certain morphosyntactic categories during the language sample that were tested during administration of the *CELF-P 2*. These findings, as well as previous research (Huang, et al., 1997), suggest that although participants may not produce certain forms during standardized testing, they may still have the linguistic knowledge of the forms.

Time Commitment

As found in the present study and supported by Huang et al. (1997), evaluation is a time-consuming aspect of a SLPs workload. Language sampling took triple the time than standardized testing when analyzing morphology and syntax. Thus, language sampling may be unrealistic in certain clinical settings due to the time demands.

Limitations and Future Directions

Limitations of the study included the sample size and reporting of AE scores. A larger sample size would allow for greater generalization of results and possibly yield additional findings. In addition, there are certain limitations associated with the use and reporting of AE scores (Maloney & Larrivee, 2007). Although standard scores are normally distributed, AE scores are not. Therefore, participants who score within a “normal range” may have an AE score that reflects a much lower- or higher- equivalent than their performance. Thus, an AE score reflects the number of items answered accurately as opposed to the quality of the responses.

To improve reliability of MLU calculation, future directions of the study should include a larger number of participants, particularly younger participants. MLU is considered an inaccurate measure of syntactic growth after Brown’s fifth stage of morphological development. Therefore, a larger sample of younger participants would increase the ability to assess the validity of the study.

Conclusion

Assessment is a critical component of a SLP’s workload. Previous research (Condouris et al., 2003; Huang et al., 1997; Tyler & Tolbert, 2002) suggests that analyzing a child’s language should be completed both standardly and naturally. The present study adds to the literature by extending the results to the evaluation of morphosyntactic components of language specifically. While the present study provides some examples of challenges associated with standardized testing, standardized testing when combined with naturalized testing appeared to yield the most accurate picture of a child’s morphosyntactic abilities. Either method of assessment, when used individually, may

not accurately represent a child’s language abilities. Thus, SLPs may seek to employ both a psychometric and descriptive approach during evaluation of linguistic competence and performance when assessing morphology and syntax to obtain the most accurate results.

REFERENCES

- Berko Gleason, J., & Bernstein Ratner, N., (2013). *The Development of Language*. Boston, MA: Pearson Education Inc.
- Bornstein, M.H. & Haynes, M. (1998). Vocabulary competence in early childhood: Measurement, latent construct, and predictive ability. *Child Development*, 69, 654-671. DOI: [10.2307/1132196](https://doi.org/10.2307/1132196)
- Brown, R. (1973). *A First Language*. Cambridge, MA: Harvard University Press. DOI: [10.4159/harvard.9780674732469](https://doi.org/10.4159/harvard.9780674732469)
- Condouris, K., Meyer, E., & Tager-Flusberg, H., (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology*, 12, 349-358. DOI: [10.1044/1058-0360\(2003/080](https://doi.org/10.1044/1058-0360(2003/080)
- Evans, J.L., & Craig, H.K., (1992). Language sample collection and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech and Hearing Research*, 35, 343-353. DOI: [10.1044/jshr.3502.343](https://doi.org/10.1044/jshr.3502.343)
- Huang, R., Hopkins, J., & Nippold, M.A., (1997). Satisfaction with standardized language testing: A survey of speech-language pathologists. *Language, Speech, and Hearing Services in Schools*, 28, 13-

29. DOI: [10.1044/0161-1461.2801.12](https://doi.org/10.1044/0161-1461.2801.12)
- Johnston, J.R., (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech, Language, and Hearing Research*, 44, 156-164. DOI: [10.1044/1092-4388\(2001/014\)](https://doi.org/10.1044/1092-4388(2001/014))
- Klee, T., Schaffer, M., May, S., Membrino, I. & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language impaired preschool children. *Journal of Speech and Hearing Disorders*, 54, 226-233. DOI: [10.1044/jshd.5402.226](https://doi.org/10.1044/jshd.5402.226)
- Maloney, E.S. & Larrivee, L.S. (2007). Limitations of age-equivalent scores in reporting results of norm-referenced tests. *Contemporary Issues in Communication Science and Disorders*, 34, 86-93.
- McFadden, T., (1996). Creating language impairments in typically achieving children: The pitfalls of “normal” normative sampling. *Language, Speech, and Hearing Services in Schools*, 27, 3-9. DOI: [10.1044/0161-1461.2701.03](https://doi.org/10.1044/0161-1461.2701.03)
- Miller, J.F., & Chapman, R.S., (1981). The relation between age and mean length of utterance in morpheme. *Journal of Speech and Hearing Research*, 24, 154-161. DOI: [10.1044/jshr.2402.154](https://doi.org/10.1044/jshr.2402.154)
- Paul, R. & Norbury, C. (2012). *Language Disorders from Infancy through Adolescence: Listening, Speaking, Reading, Writing, and Communicating* (4th ed.). St. Louis, MO: Elsevier Inc.
- Pinker, S., (1994). *The Language Instinct*. New York: Harper Perennial Modern Classics.
- Rice, M.L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M., (2010). Mean length of utterance levels in 6 month intervals for children 3-9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53, 333-349. DOI: [10.1044/1092-4388\(2009/08-0183\)](https://doi.org/10.1044/1092-4388(2009/08-0183))
- Secord, W.A., Semel, E., & Wiig, E.H., (2004). *Clinical Evaluation of Language Fundamentals-Preschool 2*. Pearson Education Inc.
- Southwood, F., & Russell, A.F., (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research*, 47, 366-376. DOI: [10.1044/1092-4388\(2004/030\)](https://doi.org/10.1044/1092-4388(2004/030))
- Tyler, A.A., & Tolbert, L.C., (2002). Speech-language assessment in the clinical setting. *American Journal of Speech-Language Pathology*, 11, 215-220. DOI: [10.1044/1058-0360\(2002/022\)](https://doi.org/10.1044/1058-0360(2002/022))
- Ukrainetz, T.A. & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy*, 18, 59-78. DOI: [10.1191/0265659002ct227oa](https://doi.org/10.1191/0265659002ct227oa)

Appendix A: Parent Questionnaire

Parent Questionnaire

Participant # _____

Child's Age: _____

Please list who your child lives with (i.e. mother, brother, etc), along with their ages, and primary language(s) spoken:

Relationship to Child	Age	Language(s) Spoken

Please describe how your child spends most of his/her day: (i.e. in school, with parents, with siblings, with babysitter).

If your child is enrolled in school:

- Current grade?
- Languages spoken in school?
- Special services in school?

DEVELOPMENTAL HISTORY

Please indicate the approximate age when your child completed the following:

Sat up alone _____

Crawled _____

Walked _____

Spoke first word _____

Fed him/herself _____

Toilet trained _____

LANGUAGE HISTORY

Does your child combine words?

YES NO

If yes, how long are your child's sentences?

Approximately how many words can your child say?

Please describe 3 typical utterances that your child produces regularly:

1. _____
2. _____
3. _____

Does your child have difficulty understanding you? _____

Do you have difficulty understanding your child? _____

Do you have any concerns about your child's speech and language abilities? If yes, please describe your concerns.

Has your child ever received speech and language therapy? If yes, what was the nature of the therapy?

Has your child had a recent hearing screening? If yes, what were the results?

Please describe any additional information about your child's speech and language that you think would be helpful.