The Thesis Committee for Amanda Lee Perme
Certifies that this is the approved version of the following thesis:

Measures of Narrative Performance in Spanish-Speaking Children on The Test of Narrative Language-Spanish

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor:
Lisa M. Bedore

Elizabeth D. Peña

Measures of Narrative Performance in Spanish-Speaking Children on The Test of Narrative Language-Spanish

by

Amanda Lee Perme, B.S.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2014

Dedication

I dedicate this thesis to my parents. Without your love, support, and encouragement these past 18 years of school, especially throughout graduate school, I would not be the person I am today. I owe it all to you! Thank you for always being there for me.

Acknowledgements

I would like to thank Dr. Lisa Bedore for all her advice and feedback during this project, as well as Dr. Elizabeth Peña for her collaborative efforts. Thank you to all those who aided me in the HABLA Lab, especially Mirza Lugo and Alexandra Estrella.

Many thanks and lots of love to my graduate school peers. I learned so much from each and every one of you and have so much respect for you all. I would have never gotten through graduate school without such an amazing group of people and will miss everyone so much. I can't wait to see what you all do in the future.

Thank you again to my family for all your encouragement and for listening to my endless phone calls about the ups and downs these past two years. I love you!

Abstract

Measures of Narrative Performance in Spanish-Speaking Children on
The Test of Narrative Language-Spanish

Amanda Lee Perme, M.A.

The University of Texas at Austin, 2014

Supervisor: Lisa M. Bedore

In the field of speech-language pathology, there is an increasing need for valid and appropriate assessment measures for Spanish-speaking students that can reliably be given by examiners with a wide range of Spanish knowledge and scoring experience. In order to determine the level of detail needed to reliably score a standardized measure of Spanish narrative ability, 15 participants, ranging in Spanish proficiency and experience levels, scored six samples of the experimental version of the Test of Narrative Language-Spanish (TNL-Spanish). Consistency and accuracy of scores were compared with Spanish proficiency levels, comfort levels, and presentation method (written transcript and audio-recorded samples). Results indicated no significant effect for any factor, indicating that examiners of varying levels of Spanish proficiency and experience level can reliably and efficiently score an assessment in Spanish when provided with clear and specific scoring procedures and information regarding the kinds of errors present in language disorders in Spanish.

Table of Contents

List of Tables

List of Figures

INTRODUCTION AND REVIEW OF THE LITERATURE

In the field of speech-language pathology, there is an increasing need for valid and appropriate assessment measures for Spanish-speaking students that can reliably be given by examiners with a wide range of Spanish knowledge and experience. In the United States, Hispanic children represent 21.4% of the early childhood population (Lopez, Barrueco, Feinauer, & Miles, 2007), leading to the largest minority population receiving speech and language services in pediatric settings (Roseberry-McKibbin, Brice, & O'Hanlon, 2005). By 2030, it is expected that the school-age population will include at least 40% of students whose first language is not English (Thomas & Collier, 2002).

In contrast to the number of CLD children to be served, only 3.5% of the American Speech-Language-Hearing Association (ASHA) membership reported the ability to deliver services in languages other than English (Langdon & Weddington, 2007). Because of this disproportion, the task of assessing and treating Spanish-speaking children with language disorders is particularly challenging. Despite these challenges, it is essential for speech-language pathologists (SLPs) to provide quality services to culturally and linguistically different clients. It is within the scope of practice that all clinicians provide culturally and linguistically appropriate services, as defined by the ASHA Principles of Ethics: "Individuals shall not discriminate in the delivery of professional services" (ASHA Principles of Ethics I, Rule C).

While assessment should be provided in a culturally and linguistically sound manner, this task often proves difficult for clinicians with limited knowledge of other languages. Best practices dictate that assessment in all of a child's languages is recommended, but it is currently unknown how much knowledge examiners need to have in order to reliably assess a child in a language other than English (Bedore & Peña, 2008;

Kohnert, 2010). It is, however, well established that examiners must be familiar with the linguistic and phonological system of the student's primary language, including vocabulary, syntax, grammar, and the sound-system of the language.

The current study will evaluate how much knowledge and information is needed for accurate scoring of the experimental version of the Test of Narrative Language-Spanish (TNL-Spanish) in order to provide a reliable and efficient measure of narrative abilities in Spanish-speaking children that can be given by any clinician. The sections that follow include a description of language disorders in Spanish-speaking children and the knowledge needed in order to identify errors, as well as what assessment manuals must include in order to achieve reliable measures of narratives of Spanish-speaking children.

SPECIFIC LANGUAGE IMPAIRMENT (SLI) IN SPANISH-SPEAKING CHILDREN: KNOWLEDGE BASE FOR THE EXAMINER

Children with specific language impairment (SLI) have been described as having significant language deficits and otherwise normal intelligence, hearing, and the "absence of gross neurological disabilities" (Leonard, Sabbadini, Leonard, & Volterra, 1987). SLI is typically identified in preschool or early elementary school. Language impairments, even when identified early, can have severe impacts on academics and social development (Beitchman et al., 2001; Johnson et al., 1999). Common characteristics of children with SLI include morphosyntactic issues, such as errors in bound-morphemes, tense agreement, and a shorter mean length of utterance when compared to their age and language-matched peers (Dromi, Leonard, & Shteiman, 1993; Lahey, 1988; Leonard, 1992; Leonard et al., 1987; Stekol & Leonard, 1979).

Many of the language difficulties of Spanish-speaking children with SLI have similar characteristics to that of their English peers. They may also make errors of

omission and substitution of pronouns, articles, and prepositions (Ambert, 1986; Juarez, 1983; Merino, 1983). However, there are some language specific characteristics of SLI. These include clitic (direct object) errors (Jacobson & Schwartz, 2002), omission of definite articles and gender agreement errors (Restrepo & Gutierrez-Clellen, 2001), overregularization, and person-number errors when combined with verbs (Ambert, 1986).

Because errors common among children with SLI extend to narratives, the knowledge needed by examiners to identify SLI in Spanish-speaking children is consistent for narrative assessment. Examiners must have a basic understanding of the linguistic aspects of the Spanish language, including articles, pronouns, verb tense, and agreement among gender and number (Anderson, 1995). SLPs must also have general knowledge about clitics in Spanish (Merino, 1992). When assessing narratives, examiners should have an understanding of the structure of narratives and where errors or inaccuracies may arise. SLPs should understand that narratives of children with SLI are often shorter, have fewer prepositions and clauses, lack well-established main ideas, and omit story grammar elements, such as character, setting, and problem (Bishop & Donlan, 2005; Merritt & Liles, 1987; Reilly, Losh, Bellugi, & Wulfeck, 2004). Children also make grammar errors (e.g. Gillam & Johnston, 1992; Norbury & Bishop, 2003; Scott & Windsor, 2000), and tell stories that are less cohesive (Liles, 1985; Norbury & Bishop, 2003), with shorter utterances and less complex syntax (e.g. Bishop & Donlan, 2005; Gillam & Johnston, 1992; Norbury & Bishop, 2003, Reilly et al., 2004).

Examiners with limited knowledge of Spanish grammar, including syntax and morphology, may find errors in narratives difficult to identify and analyze. Therefore, it is necessary for assessment manuals to provide explanations and examples of these elements so that examiners can rate correct or incorrect use during the evaluation.

As narratives have been identified as an ideal measure to aid in the identification of language impairment, they have become a focus of bilingual assessment development. Most notably, a Spanish version of the Test of Narrative Language (TNL; Gillam & Pearson, 2004) is currently being developed. The assessment identifies the absence or presence of language impairment in children aged 5;0 through 11;11 years. It also provides valuable information about how children sequence events in a story, use semantic and morphosyntactic knowledge, and how they recall and understand elements of stories. Both the English and Spanish versions incorporate three tasks each for narrative comprehension and production: no picture, sequenced pictures, and a single picture. This format provides information about the child's narrative comprehension and production abilities with decreasing examiner cues and structure. Specific rubrics were developed for the published English version, as well as the Spanish version. Narrative macrostructure elements, including character, setting, initiating event, plan, action, consequence, and internal response are included, along with microstructure elements, such as conjunctions, adverbs, and noun phrases. Each narrative production element is scored on a scale from 0 to 3, with 3 serving as the highest score. Narrative comprehension items are scored with a "1" if the child answers correctly and a "0" if the child answers incorrectly. The scoring rubrics include the precise words the child must use in order to receive credit for each item.

A narrative assessment for Spanish-speaking children generates several distinct challenges for examiners with limited knowledge of Spanish. While it is possible for examiners to listen for the intended target during narrative comprehension questions, the child may use a synonym or correct answer that is not included in the list of appropriate answers. While examiners who are proficient in Spanish would understand the child's

answer and relate it to their understanding of the narrative and question, other clinicians may find it difficult to consider any other answer but those included within the manual.

Another challenge of this type of assessment concerns ratings of narrative production. For example, it is nearly impossible for examiners with little to no knowledge of the Spanish language to accurately judge whether the child uses correct syntax and vocabulary, including grammar, verb tense, and pronoun reference. Furthermore, it is difficult to rate the child's ability to describe the problems within the story, if a consequence or resolution is provided, and if a detailed conclusion was included if the examiner cannot understand the story. These are behaviors that are more difficult to define.

These challenges present a difficult task for examiners and test developers. In order to ensure that the assessment can reliably be given by clinicians of all proficiency and experience levels, the scoring procedures must be clear and detailed. For narrative comprehension tasks, several potential answers—including synonyms and additional appropriate answers—must be included specifically in the manual. Detailed explanations of narrative production items, including a basic overview of the concept and examples of possible answers for each score (0-2) is necessary. For example, examiners should be given a brief explanation of verb tense in Spanish and potential incorrect or correct responses the child may use. An example of an item evaluating verb tense is included in Figure 1.

Figure 1. TNL-Spanish Item Example

23. Utiliza el mismo tiempo verbal a través del cuento; *Uses the same verb tense throughout story (remember, some Spanish verb changes are valid—i.e. imperfect [ongoing] vs. preterite [interrupting event in past])*
    0   Cambia tiempo verbal dos o más veces; *Changes verb tense 2+ times*

Figure 1 (continued)

1   Cambia tiempo verbal una vez; *Changes verb tense once* [e.g. "El perro estaba en la calle. De repente, pasa un coche." The use of the imperfect (estaba) and preterite is incorrect because "pasa" is present tense).

2   Utiliza el mismo tiempo verbal; *Uses same verb tense* [e.g. "El perro estaba en la calle cuando un coche pasó." The use of imperfect (estaba) and preterite (pasó) is correct].

The item includes a brief explanation of verb tense changes and examples of incorrect and correct tense changes that may be included within the child's narrative. These detailed explanations and examples give examiners with limited Spanish knowledge more information in order to ensure a more accurate and reasoned score.

In order to accurately score the test, examiners need detailed explanations like the item described above. While examiners may not understand the child's narrative in full, they should still be able to accurately rate narratives and answers given detailed scoring procedures and explanations.

HYPOTHESIS

Bilingual children are in need of a quality form of assessment in order to ensure accurate enrollment in speech and language services. Ideally, the experimental Spanish version will provide speech-language pathologists with a reliable and efficient measure of narrative ability for Spanish-speaking children. The current study will attempt to evaluate the knowledge base for accurate scoring of the Spanish version of the Test of Narrative Language in order to provide a valuable and efficient measure of narrative comprehension and production in Spanish-speaking children. Based on the detail provided within the manual and scoring sheets, it is expected that examiners with high and low proficiency and experience levels will be able to score the test accurately. Scoring accuracy should also be similar when scored through an on-line manner as well

as through transcript analysis in order to ensure testing efficiency. The results of this study will inform the specificity of criterion needed to be able to score the test on-line, as this is the most ideal and efficient condition. In addition, the scorers' knowledge of Spanish will be examined along with their scores in order to determine the extent to which Spanish proficiency influences the ability to apply scoring criteria accurately. Finally, scoring comfort levels will be compared with scores to determine the correlation between experience level and scoring accuracy.

METHODS

PARTICIPANTS

15 students from the Communication Sciences and Disorders department at the University of Texas at Austin participated in the study. Participants were recruited through e-mails and announcements in upper level courses within the department. All participants had a background in speech-language pathology and some knowledge of basic assessment procedures. Subjects ranged from second year undergraduate students to graduate students. Participants also had varying levels of Spanish proficiency and experience with giving and scoring assessments, as determined by online measures of Spanish proficiency, the Bilingual Language Profile (BLP; Birdsong, Gertken, & Amengual, 2012) and the Oxford Placement Test—Spanish (OPT—Spanish; Oxford University Language Centre, 2010). Experience level was determined by records of previous evaluation experiences. All methods and procedures were approved by the IRB Committee.

Among the participant group, Oxford Placement Test (OPT) scores ranged from 12 to 44 (on a scale of 0-50). There was one "absolute beginner," six "lower intermediate," four "intermediate," and four "upper intermediate," resulting in a wide range of proficiency levels. An upper and lower half was determined and split between scores of 23 and 25. According to the BLP survey, four participants had been speaking Spanish from birth, two began learning Spanish before age 10, seven began learning in middle and high school, and one had never taken Spanish classes.

The correlation between comfort level and scoring was determined by using self-ratings of comfort level with giving and scoring speech and language assessments. Subjects were asked to rate their comfort level on a scale of 1 to 10. Four students gave a level of "2," four gave a level of "3," one rated a level of "4," four rated a level of "5,"

one gave a level of "8," and one a level of "9." The students in the upper range had scored assessments before. These numbers indicate a wide range of comfort and experience levels.

Existing Data/Materials

For the stimuli used for scoring, the study utilized existing data from two previous research projects, Phenotype Assessment Tools for Bilingual Children, awarded to Peña and Bedore, and the Cross-Language Outcomes of Language Impairment, awarded to Peña, Bedore, and Griffin. In these studies, the investigators collected narrative test data from the experimental version of the Test of Narrative Language (TNL), which was then transcribed and deidentified. Samples were collected by undergraduate and graduate research assistants in several Austin, Texas elementary schools between 2010 and 2011.

Six TNL samples were selected from the existing data set, which were then divided between transcript and audio recording stimuli. Samples were selected based on difficulty level. Samples with few errors were labeled "easy", samples with a medium number of errors were labeled as "medium", and those with a large number of errors were labeled "difficult." In order to determine the difficulty level for each sample, the highest score within the data (80%) was determined and divided by three in order to establish an upper, intermediate, and lower third. Two comparable scores were then selected within each range to incorporate a transcript and recording sample for each of the low, medium, and high range of scores. Raw score percentages and item analysis were used to ensure that each sample within the series was similar to its counterpart. For example, in the lower range, two samples with scores of 25% were selected based on equivalent scores

and similar responses on the test items. One was chosen as the recorded stimulus while the other was selected as the transcript stimulus.

TNL samples had previously been scored by research assistants working on the Phenotypes and Bilingual Outcomes projects. Inter-rater reliability was completed again by the principle investigator to ensure accuracy of scoring.

Transcripts were completed on the Systematic Analysis of Language Transcripts (SALT) software (Miller & Chapman, 1998) by the primary investigator and a trained undergraduate research assistant. These recordings and transcripts were uploaded to a password protected cloud-sharing space ("UTBox:" http://www.utexas.edu/its/cloudstorage/), which could be accessed by participants online.

Additional materials included testing forms and booklets from the experimental Spanish version of the TNL. These forms included the examiner's manual, including administration and scoring instructions, as well as scoring sheets. These indicate correct and incorrect answers for comprehension tasks, checklists for story retelling, and rating scales for narrative generation tasks (i.e. Indicates the setting: 0—No information; 1—General Information, such as castle, forest; 2—Names the setting, such as Dallas, Parque Alta Loma).

To measure participant ability levels, online measures of Spanish proficiency and rating sheets regarding experience level were utilized. These included the Bilingual Language Profile (BLP; Birdsong, Gertken, & Amengual, 2012) and the Oxford Placement Test—Spanish (OPT—Spanish; Oxford University Language Centre, 2010). The BLP is a self-rating measure of language use and ability, while the OPT measures accuracy of grammatical use in Spanish. Record sheets of previous experiences with giving and scoring speech and language assessments and comfort levels were also provided to participants prior to participation in the study.

PROCEDURES

Once participants consented, they were sent links to the Bilingual Language Profile (BLP: Birdsong, et al., 2012) and the Oxford Placement Test—Spanish (OPT—Spanish; Oxford University Language Centre, 2010). Participants were also asked to complete a questionnaire regarding previous experiences with assessments, including types of assessments given and/or scored, as well as their perceived comfort level on these tasks. Participants were then sent a training presentation on the Spanish TNL manual, including the purpose, test items, basic scoring procedures for each task, and an overview of SALT transcripts. The training presentation was prepared by the principle investigator. Participants were trained only on information presented in the manual to ensure that results emulate future results in which examiners only have access to the manual. Each participant was then given access to the samples online, as well as a copy of the TNL manual and six record forms. Each participant scored the six selected TNL samples. Three samples were presented as audio recordings and three as SALT transcripts, all of which were uploaded to UTBox. Each condition included two samples with few errors, a medium number of errors, and a large number of errors. Samples were presented to each participant in a counterbalanced manner. While the presentation order varied among subjects, all participants scored the same six samples. To ensure independent scoring, participants conducted scoring procedures on separate computers and were not allowed to see others' scores. To control the recording condition, scorers listened to each audiotaped sample a maximum of three times. Upon completion, participants returned all six scored record forms to the principle investigator.

Participants' scores for samples were then compared with the previously determined score in order to determine percent accuracy for each sample. Scoring

accuracy was then compared for several variables: (a) Spanish proficiency, (b) comfort level, and (c) audio recording (on-line) vs. written transcript.

ANALYSIS

Once scores were submitted, scores for all six subtests for each sample were recorded and totaled. Total scores were then compared to previously determined scores and percent accuracy was calculated for the narrative comprehension and oral narration portions of each sample. These percentages were then averaged for each sample and recorded.

To compare narrative comprehension and oral narration scores of written transcript and audio recording samples, a correlation coefficient was calculated. This was done by obtaining the total scores for narrative comprehension and oral narration from each sample. Total scores were then averaged, resulting in an average narrative comprehension and oral narration score for the written transcript samples, audio recording samples, and the total number of samples. The correlation between audio-recorded and written samples was then calculated for both the narrative comprehension portion and the oral narration subtest.

To compare scores based on Spanish proficiency, scores were organized based on participant results on the Oxford Placement Test. The test is scored on a scale from 0 (absolute beginner) to 50 (advanced). For the participant group, there was a natural break between scores of 23 and 25—the middle of the scoring range—so scores were split at this point to create a "low proficiency" and a "high proficiency" level. A one-way analysis of variance (ANOVA) was used to determine the differences in scoring accuracy

based on these two proficiency groups for audio recorded samples, written transcript samples, and total scores for all samples.

Differences in scoring accuracy based on experience were determined in a similar manner. On the experience survey, subjects indicated their "level of comfort" with giving and scoring assessments on a scale of 1 to 10. To create an upper and lower half, participants were split between comfort levels of 4 and 5. A one-way ANOVA procedure was again used to determine the impact of scoring experience on scoring accuracy of the TNL-Spanish.

These three analyses were completed to determine differences in scoring accuracy based on three factors: (1) sample presentation in written transcript and audio recording formats, (2) Spanish proficiency level, and (3) comfort level.

RESULTS

The primary purpose of this study was to determine the differences in scoring accuracy of narrative tasks using the experimental version of the Test of Narrative Language—Spanish based on scorer's Spanish proficiency and experience level, in order to determine the level of information needed when scoring an assessment in Spanish. Additionally, assessment samples were compared based on whether they were presented as written transcripts or audio recordings. In order to address these questions, participant's scores were recorded and totaled. Percent accuracy for the narrative comprehension and oral narration portions were calculated using the previously determined scores for each sample. Average scores were then recorded for each portion and analyzed.

Written Transcripts vs. Audio Recordings

To evaluate differences in scoring accuracy based on the presentation of samples through written transcripts and audio recordings, an agreement ratio was calculated for the average scores of the 6 narratives rated by each participant in each format. To determine averages, a ratio for each sample was calculated using each subject's score and the previously determined target score. The averages for the 3 transcribed samples and 3 audio-recorded samples were then calculated. Observation of the calculated ratios for narrative comprehension and oral narration for both the transcribed and audio-recorded samples revealed that most subjects were within 10% of the target (between .9 and 1.1). For narrative comprehension, 10/15 subjects were within 10% of the target on both transcript and audio-recorded samples. For oral narration, 8/15 subjects were within 10% of the target for both formats. Average scores and accuracies for participant scores are

reported in Table 1. These calculations indicate that there is a high level of agreement between the target score and participant scores on both recorded and written samples, indicating that most subjects were able to score the samples similarly and accurately, regardless of Spanish proficiency and experience levels when presented with transcribed narratives as well as audio-recorded samples.

The overall array of averages for audio and written samples was also compared between narrative comprehension and oral narration subtests. Correlation coefficients were determined from these combined averages. For narrative comprehension, the correlation coefficient between the written samples and recorded samples was significant at 0.91. For oral narration, the correlation coefficient was also large, with a value of .87. These values signify that examiners with varying levels of experience and Spanish proficiency can reliably score assessments when presented with transcribed narratives as well as audio-recorded samples.

Table 1. Participant accuracy on transcript and audio recordings

Participants	Avg. Transcript Scores		Average Audio Scores		Total
	NC	ON	NC	ON	
P1	0.92	0.96	1.11	1.14	1.03
P2	0.93	0.87	0.87	0.65	0.83
P3	0.97	0.87	0.97	0.74	0.89
P4	0.98	0.65	0.79	0.47	0.72
P5	0.86	0.99	1.05	0.95	0.96
P6	1.18	1.08	0.96	0.84	1.02

Table 1 (continued)

P7	0.99	0.72	0.88	0.72	0.83
P8	1.02	1.05	0.95	0.82	0.96
P9	0.98	0.93	1.01	0.97	0.97
P10	1.06	0.78	1.02	0.87	0.93
P11	1.01	0.85	1.01	0.68	0.89
P12	0.95	0.89	0.97	0.88	0.92
P13	0.94	0.97	0.98	0.97	0.97
P14	1.01	0.91	1	0.92	0.96
P15	0.98	1.03	0.98	0.82	0.95

Note: P=Participant Number, NC= Narrative Comprehension, ON= Oral Narration. Values of 1.0 indicate 100% agreement with target sample scores.

Spanish Proficiency Levels

As explained above, the participants were divided into higher and lower proficiency groups. Participants with high proficiency scored above 25 on the Oxford Placement Test and participants with lower proficiency ranged between 12 and 23. A one-way ANOVA yielded no significant effect for higher and lower proficiency groups. For audio recorded samples, ANOVA revealed no significant effect, $F(1, 13) = 1.071$, $p = .319$. Similarly, ANOVA showed no significant difference in the accuracy of the written samples, $F(1, 13) = .803$, $p = .386$. For the total number of samples, ANOVA results yielded no significant effect, $F(1, 13) = 1.271$, $p = .280$. Results of this comparison are reported in Table 2. Accuracy for each factor was also calculated in order

16

to determine group differences for Spanish proficiency level. For audio recordings, total accuracy was .90. Accuracy was slightly higher for the "higher proficiency" group (.93) than the "lower proficiency" group (.87). Similar trends were observed in written samples. Participants with higher proficiency were slightly more accurate (.96) than the group with lower proficiency levels (.93). When total accuracy was calculated, the "higher proficiency" group achieved higher accuracy (.94) than the "lower proficiency" group (.90). However, these differences are not significant and reveal that scoring accuracy is generally similar and high across proficiency groups. Accuracy results are reported in Table 3. In sum, there was no difference in scoring accuracy based on the participants' proficiency as determined by ANOVA procedures and accuracy comparisons.

Table 2. Differences between Spanish proficiency and scoring accuracy

		Sum of Squares	df	Mean Square	F	Sig.
Total	Between Groups	0.008	1	0.008	1.271	0.28
	Within Groups	0.085	13	0.007		
	Total	0.093	14			
Audio Samples	Between Groups	0.015	1	0.015	1.071	0.319
	Within Groups	0.176	13	0.014		
	Total	0.191	14			
Written Samples	Between Groups	0.004	1	0.004	0.803	0.386
	Within Groups	0.072	13	0.006		
	Total	0.077	14			

Note: df= Degrees of Freedom, F= F-value, Sig.= Significance.

Table 3. Average scoring accuracy between higher and lower proficiency groups

		Number in Each Group	Average Scoring Accuracy	Std. Deviation
Total Samples	Lower Proficiency	7	0.8971	0.11481
	Upper Proficiency	8	0.9442	0.02844
	All Participants	15	0.9222	0.08152
Audio Samples	Lower Proficiency	7	0.8667	0.16266
	Upper Proficiency	8	0.929	0.04968
	All Participants	15	0.8999	0.11666
Transcript Samples	Lower Proficiency	7	0.9264	0.10035
	Upper Proficiency	8	0.961	0.04107
	All Participants	15	0.9449	0.07401

*Note: Average scoring accuracy of 1.0 indicates 100% average accuracy for all samples when compared to the target scores. Values within .90 and 1.1 are within 10% of the target score.

Comfort Level

Scoring accuracy for audio-recorded samples, written samples, and the total number of samples was also calculated based on each participant's comfort level group. For audio samples, accuracy for the "upper comfort level" group was the same as the "lower comfort level" group (.90). For the written transcript samples, the accuracy for the "upper comfort level" group was .95 and the accuracy for the "lower comfort level" group was .94. For the total number of samples, the group most comfortable with scoring was only slightly more accurate (.93) than the lower comfort level group (.92). Results of accuracy comparisons are reported in Table 5. To explore the difference in scoring accuracy between comfort level and scoring, a one-way ANOVA procedure was conducted for audio-recorded samples, written transcript samples, and total samples. The

ANOVA demonstrated no significant effect for comfort level on audio-recorded samples $F(1, 13) = .018, p = .895$. Similarly, ANOVA revealed no significant effect for comfort level on written samples, $F(1, 13) = .081, p = .780$. For total samples, ANOVA again revealed no significant effect for comfort level and accuracy, $F(1, 13) = .056, p = .816$. Results are reported in Table 4. These results indicate that overall, examiners can effectively score the assessment given detailed instructions independent of comfort level and/or experience with giving and scoring speech and language assessments.

Table 4. Differences between comfort level and scoring accuracy

		Sum of Squares	df	Mean Square	F	Sig.
Total	Between Groups	0	1	0	0.056	0.816
	Within Groups	0.093	13	0.007		
	Total	0.093	14			
Audio Samples	Between Groups	0	1	0	0.018	0.895
	Within Groups	0.19	13	0.015		
	Total	0.191	14			
Transcript Samples	Between Groups	0	1	0	0.081	0.78
	Within Groups	0.076	13	0.006		
	Total	0.077	14			

Note: Note: df= Degrees of Freedom, F= F-value, Sig.= Significance.

Table 5. Scoring accuracy between upper and lower comfort level groups

		Number in Each Group	Average Scoring Accuracy	Std. Deviation
Total Samples	Lower Comfort	8	0.9174	0.07185
	Higher Comfort	7	0.9277	0.09703
	All Participants	15	0.9222	0.08152
Audio Samples	Lower Comfort	8	0.896	0.11748
	Higher Comfort	7	0.9044	0.12494
	All Participants	15	0.8999	0.11666
Transcript Samples	Lower Comfort	8	0.9396	0.05731
	Higher Comfort	7	0.9509	0.09418
	All Participants	15	0.9449	0.07401

*Note: Average scoring accuracy of 1.0 indicates 100% average accuracy for all samples when compared to the target scores. Values within .90 and 1.1 are within 10% of the target score.

DISCUSSION

This study assesses the amount of linguistic knowledge and information needed in order to accurately and consistently score measures of narrative ability in Spanish speaking children, regardless of examiner experience and Spanish proficiency. This question was examined using the experimental version of the Test of Narrative Language—Spanish (TNL—Spanish), which assesses Spanish-speaking children's abilities to comprehend and produce narratives. While all test items and responses are in Spanish, it is anticipated that the evaluation will be able to be scored by a wide variety of speech-language pathologists, including those with little to no Spanish proficiency and/or assessment experience as a result of detailed scoring procedures and examples. As time limitations are a common complaint of evaluating SLPs, the assessment should also have the capability of being scored on-line during the assessment period with the assistance of a recorder rather than through transcription.

These factors were explored by comparing participant's scores for six samples— three written transcripts and three audio recordings—to previously determined scores for each sample. Overall, correlations were consistent with the hypothesis. Correlation coefficients for audio recorded and transcribed samples revealed a high relationship between the two methods. When scores were compared based on proficiency levels, statistics indicated that there was not a significant difference between upper and lower proficient groups. Finally, scores compared on the basis of scoring comfort level showed that there was not a significant difference between scores with higher and lower comfort levels. These results provide encouraging information about the Test of Narrative Language—Spanish and other similar Spanish assessments. Regardless of Spanish proficiency and previous testing experiences, examiners have the ability to accurately and

consistently score high quality assessments when provided with sufficient information about narrative and linguistic elements in Spanish. Given thorough scoring procedures, even examiners with basic knowledge of Spanish can score assessments in the language, which may be more efficient than seeking out a bilingual SLP or interpreter to complete the evaluation. When scoring language and narrative samples of Spanish-speakers, examiners must have knowledge about the types of errors that distinguish between children who are typically developing or those who present with language impairment. By providing specific information about the kinds of acceptable and unacceptable responses within the scoring manual, examiners are able to gather information about the manifestation of language impairment in Spanish, regardless of their prior understanding or knowledge of these elements. Results of this study were expected based on the type of information detailed within the manual. In order to analyze a narrative sample, SLPs must evaluate the child's grammar, syntax, and vocabulary, as well as macrostructure elements, including organization and cohesiveness and story grammar, such as setting, consequences, and resolution (Stein & Glenn, 1979; Justice et al, 2006). While analysis of these elements is difficult for speakers unfamiliar with Spanish, explanations and examples of both microstructure and macrostructure domains of narratives were incorporated into the scoring procedures, giving examiners access to this crucial information. With access to this information, examiners with a wide variety of Spanish proficiency levels and experience have the ability to score assessments accurately and reliably.

Audio Recordings vs. Written Transcripts

The correlation coefficient between the scores for the audio-recorded samples and the written samples showed a high association for both the narrative comprehension (.91) and oral narration (.87) portions of the samples presented. This indicates that, regardless of the manner in which the assessment is presented, scores are comparable. While most participants provided feedback that the transcripts were easier to score than the audio recordings, as they were able to visually compare acceptable responses with the written samples of the children's narratives, the scores did not show a significant difference between the two. Further, participants also believed that the narrative comprehension questions were simpler to score than oral narration due to the lack of grammatical judgments in the question and response format (i.e. examiners listened for specific responses outlined in the manual and did not have to judge syntax or use of grammatical elements, such as pronouns and adverbs). However, the data showed no significant difference in either format, revealing that both portions of the test were scored in similar manners on both the audio recordings and written transcripts, presumably due to the level of detail included within the manual.

These data provide valuable information about the assessment measure and the explanations of scoring procedures included in the manual. Most notably, the ability to score the assessment from an audio recording in the same way as a transcript indicates that the test can be scored on-line with the assistance of an audio recorder. Therefore, it is not necessary for the examiner to transcribe the child's responses in order to effectively score the test. This is likely due to the level of detail provided to examiners about scoring procedures. Given precise wording and examples, examiners are able to listen for specific elements and consistently rate narrative samples without reading transcripts. For

example, examiners can refer to the manual during the assessment period and listen for specific information the child is expected to say, including consistent verb tense, pronoun reference, adverbial clauses, and other grammatical and vocabulary elements. Possible alternate answers and synonyms were also included in order to eliminate confusion for raters with limited Spanish proficiency. Ideally, the SLP should be aware of these linguistic components when assessing narrative samples, but may find this task difficult and tedious with little knowledge of Spanish. Given detailed scoring procedures, this step is eliminated and thus leads to more efficient scoring.

According to surveys of speech-language pathologists (Huang, Hopkins, & Nipplod, 1997; ASHA, 2012), assessments are most often chosen based on their efficiency. Because of the lack of time allotted for evaluations, common tests selected for use in the schools are typically quick to administer, interpret, and report. Examiners do not have time to return to an assessment to transcribe and score later to interpret the results. Rather, they must be able to score most of the test during the evaluation session, with the exception of some analysis of recorded language and narrative samples. The results of this study indicate that assessments with specific scoring information can be scored reliably and efficiently without the need for written transcriptions.

The high scoring relationships between narrative comprehension and oral narration tasks also highlight the capacity for various narrative test items to be included on the assessment without concerns that they might be scored with significant differences. In this study, comprehension questions, story retells, sequenced picture story generation, and single picture narrative productions were all scored relatively consistently across formats, as indicated by the similar correlation coefficients.

The data resulting from this study reveal that the assessments with detailed scoring procedures have the potential to be a reliable and efficient narrative assessment used for the identification of Specific Language Disorder in Spanish-speaking children.

Spanish Proficiency

It was hypothesized that scoring would be similar among participants, regardless of Spanish proficiency level. According to ASHA, bilingual service providers must have specific knowledge and skills in another language, including native or near-native proficiency, knowledge of typical language development, and the ability to give and interpret assessments in the second language (2004). While this is ideal, access to capable bilingual SLPs is limited in most areas, leading to alternative options for clinicians serving speakers of other languages. One such viable option involves the production of detailed assessment procedures, allowing examiners without the appropriate qualifications to serve as a bilingual SLP the ability to effectively assess children in languages other than English. For example, although the TNL-Spanish is given and responded to entirely in Spanish, it was expected that scorers could effectively score the tasks using the manual and scoring sheet, both of which were generated with attention to detail and clarity in anticipation of examiners with basic knowledge of Spanish. Results from this study revealed insignificant differences in scoring among groups of higher and lower Spanish proficiency levels, leading to the potential for evaluations in the absence of ASHA-qualified bilingual service providers.

These data confirm the utility of high quality scoring procedures and instructions. Clear, detailed, and specific information about linguistic elements present in language disorders in Spanish was provided for each task, resulting in consistent scoring among

examiners with varying proficiency levels. For example, a list of acceptable and unacceptable answers was provided for each narrative comprehension question in the manual and specific targets were indicated on the scoring sheet, facilitating the examiner's ability to hear and determine whether the child answered the question correctly or incorrectly. For oral narration tasks, specific information was provided in order to aid scorer's abilities to make correct judgments about the child's narrative generation. For example, for tasks assessing a child's ability to indicate a problem, conflict, or event that motivates the character, explanations and examples were included for each potential score (i.e. the child receives a score of "1" if the child depicts a "vague or incomplete description of a problem, conflict, or event. The author says that there is an ogre or unicorn, but does not say what the specific problem is or that the children would have to react" [e.g. "Los niños vieron al unicornio. Marcos corrió hasta ella"]).

This level of detail contributes to examiners' abilities to score each task of the assessment, even with relatively low levels of Spanish knowledge. While there was a wide variety of Spanish proficiency among the participant group, there was not a significant difference among scorers. This provides valuable information to test developers, especially those creating assessments designed for speakers of languages other than English. Without the level of detail provided for subjects in this study, it is unknown if these results would generalize. In other words, examiners with varying levels of Spanish proficiency may not demonstrate similar scores when using other types of assessments that do not include such detailed scoring procedures.

Comfort Level

It was predicted that scores would also be similar among groups stratified for comfort level resulting from previous knowledge and experiences with speech and language assessments. This hypothesis was projected on the basis of clear and detailed instructions given to scorers, facilitating examiner's abilities to understand scoring procedures even with limited experience and/or comfort with testing. Although most participants had never scored an assessment, subjects were able to consistently score samples using only the given manual and record form. Participants were not trained on the scoring procedures and conducted the scoring without assistance. However, the data show consistent scores across participants, regardless of comfort level, age, experience level, and proficiency. As this was the first assessment experience for most subjects, results clearly reveal the power of detailed scoring procedures and explanations within both the manual and record form. This result is even greater given the format of the test. While many subjects indicated lower comfort levels, many subjects were not comfortable with Spanish either. Additionally, narrative scoring is often subjective, and it is difficult to maintain inter-rater reliability, which makes consistency a difficult task. However, these factors did not result in drastically different scores among participants. Therefore, the resulting data can be directly attributed to thorough scoring instructions. Data resulting from this study provide valuable information for test developers of assessments intended for use with languages other than English. By including explanations and examples of specific linguistic elements that indicate the presence of a language disorder within a narrative, examiners with varying levels of comfort can reliably score tests, regardless of prior experience or linguistic knowledge.

Limitations

Several caveats must be observed when interpreting the results of this study. First, the number of participants is small. A larger, more representative group would have given more accurate data. Furthermore, only one graduate student participated, resulting in a disproportional number of undergraduate students with little to no experience with assessments. However, this inexperienced group yielded important information on the value of quality manuals and scoring instructions, perhaps more so than a participant pool consisting of mainly graduate students and experienced SLPs.

Additional Considerations for Future Research

While results from this study show scoring consistency among raters given the level of detail in the manual and record form, future research should investigate the accuracy of scoring for each sample, with attention to any differences on samples with larger or fewer amounts of errors. It would also be of interest to examine accuracy for each subtest to determine whether certain portions of the test are easier to score more accurately by those with limited Spanish proficiency and/or experience. Along this line, an analysis of specific items would be warranted in order to identify what types of questions are more reliably scored across raters. From this information, modifications to the manual and record form may be made to ensure that scoring procedures for certain items are more clearly described.

The information resulting from this study, as well as future directions for research, would be stronger with a greater number of participants. Therefore, further research should incorporate a larger number of scorers who are more representative, including both undergraduate and graduate students, as well as practicing clinicians with a wide variety of Spanish proficiency levels.

28

Conclusion

A preliminary study of the use of a specific scoring system yielded evidence suggesting that assessments intended for narrative production in Spanish can be scored consistently, resulting from data that yielded similar scores across several factors, including presentation method, Spanish proficiency, and comfort level, when examiners were given information about specific linguistic elements that characterize language impairment in Spanish. Although there was a wide variety of variability among participants, scores were consistent across groups. Because of this, it can be derived that results are directly attributed to well established and thorough instructions included in the manual and record forms provided to raters.

This study highlights the importance of clear scoring instructions and detailed descriptions of appropriate responses and expansions of certain areas of Spanish, such as tense, grammar, and clausal phrases, for examiners with limited knowledge. In order to accurately identify the presence of a language disorder within a narrative sample, examiners must have a certain degree of knowledge about the language and specific errors that distinguish typically developing children from those with language disorders. When provided with this information within the manual, along with thorough scoring instructions, examiners of all levels have the potential to accurately and consistently classify Spanish-speaking students with language disorders. Equipped with this information, assessments created for speakers of languages other than English, such as the TNL-Spanish, will be targeted towards the student and examiner alike.

References

Ambert, A. (1986). Identifying language disorders in Spanish-speakers. In A. C. Willig & H. F. Greenberg (Eds.), *Bilingualism and learning disabilities: Policy and practice for teachers and administrators* (pp. 15–33). New York: American Library Publishing.

American Speech-Language-Hearing Association (2012). *Demographic profile of ASHA members providing bilingual services.* Retrieved from http://www.asha.org/uploadedFiles/Demographic-Profile-Bilingual-Spanish-Service-Members.pdf

American Speech-Language-Hearing Association (2012). *Professional issues: Caseload and workload.* Retrieved from http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589934681&section=Key_Issues

American Speech-Language-Hearing Association. (2001, December 26). Code of ethics (revised). *The ASHA Leader, 6*(23), 2. Retrieved from http://www.asha.org/policy/KS2004-00215/#sthash.DGpnNhub.dpuf

American Speech-Language-Hearing Association. (2004). *Knowledge and skills needed by speech-language pathologists and audiologists to provide culturally and linguistically appropriate services*. Retrieved from http://www.asha.org/policy/KS2004-00215/#sthash.DGpnNhub.dpuf

American Speech-Language-Hearing Association. (2012). *2012 schools survey report: SLP caseload characteristics.* Retrieved from www.asha.org/research/memberdata/schoolssurvey/.

Anderson, Raquel (1995). Spanish Morphological and Syntactic Development. In H. Kayser (Ed.), *Bilingual Speech-Language Pathology: A Hispanic Focus* (pp. 41-74). San Diego: Singular Publishing Group.

Bedore, L., & Peña, E. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism. 11,* 1–29.

Beitchman, J., Wilson, B., Johnson, C., Atkinson, L., Young, A.,...Adlaf, E. (2001). Fourteen-year follow-up of speech/language impaired and control children: Psychiatric outcome. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 75–82.

Birdsong, D., Gertken, L.M., & Amengual, M. *Bilingual language profile: An easy-to-use instrument to assess bilingualism.* COERLL, University of Texas at Austin. Web. 20 Jan. 2012. <https://sites.la.utexas.edu/bilingual/>.

Bishop, D., & Donlan, C. (2005). The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology, 23,* 25-46.

Bishop, D., & Edmundson, A. (1987). Language-impaired four-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52,* 156–173.

Dromi, E., Leonard, L., & Shteiman, M. (1993). The grammatical morphology of Hebrew-speaking children with specific language impairment: Some competing hypotheses. *Journal of Speech and Hearing Research, 36,* 760–771.

Gillam, R. B., Peña, E. D., Bedore, L. M., & Pearson, N. (in development). *Test of Narrative Language, Spanish Adaptation*.

Gillam and Pearson, (2004). Test of Narrative Language. Austin, TX: Pro-Ed.

Gillam, R. & Johnston, J. (1992). Spoken and written language relationships in language/learning-impaired and normally achieving school-age children. *Journal of Speech and Hearing Research, 25,* 1303-1315.

Huang, R., Hopkins, J., & Nippold, M. (1997). Satisfaction with standardized testing: A survey of speech-language pathologists. *Language, Speech, and Hearing Services in Schools*, *28*(January), 12-29.

Jacobson, P. and Schwartz, R. (2002) Morphology in incipient bilingual Spanish-speaking preschool children with specific language impairment. *Applied Psycholinguistics, 23* (1), 23-41.

Jacobson, P.F. and Schwartz, R.G. (2005). English past tense use in bilingual children with language impairment. *American Journal of Speech Language Pathology, 14*(4), 313-323.

Johnson, C., Beitchman, J., Young, A., Escobar, M., Atkinson, L.,…Wilson, B. (1999). Fourteen-year follow-up of children with and without speech/language impairments: Speech/language stability and outcomes. *Journal of Speech, Language, and Hearing Research, 42*(3), 744–760.

Juarez, M. (1983). *A comparison of the oral language skills of language-impaired and normally-developing Spanish-English bilingual children.* Unpublished doctoral dissertation, University of Texas, Austin.

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performance. *American Journal of Speech-Language Pathology*, *15*, 177–191.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43,* 465–473.

Lahey, M. (1988). *Language disorders and language development*. New York: Macmillian.

Langdon, H. W., & Weddington, G. T. (2007). Bilingual program receives federal grant. *The ASHA Leader, 12*(12), 27.

Leonard, B., Sabbadini, L., Leonard, J., & Volterra, V. (1987). Specific language impairment in children: A cross-linguistic study. *Brain and Language, 32,* 233–252.

Leonard, L. (1992). Specific language impairment in three languages: Some cross-linguistic evidence. In P. Fletcher & D. Hall (Eds.), *Specific speech and language disorders in children* (pp. 118–126). San Diego, CA: Singular.

Liles, B. (1985). Narrative abilities in normal and language disordered children. *Journal of Speech and Hearing Research, 28,* 123–133.

López, M., Barrueco, S., Feinauer, E., & Miles, J. (2007). Young Latino infants and families: Parental involvement implications from a recent national study. *Family involvement research digests,* Retrieved from http://www.hfrp.org/publications-resources/browse-our-publications/young-latino-infants-and-families-parental-involvement-implications-from-a-recent-national-study

Merino, B. J. (1983). Language development in normal and language handicapped Spanish-speaking children. *Hispanic Journal of the Behavioral Sciences, 5,* 379–400.

Merino, B. (1992). Acquisition of syntactic and phonological features in Spanish. In H. W. Langdon and Li-Rong L. Cheng (Eds.) *Hispanic children and adults with communication disorders: Assessment and intervention* (pp. 57-98)*.* Gaithersburg, MD: Aspen Publishers.

Merritt, D., & Liles, B. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling and story comprehension. *Journal of Speech and Hearing Research*, *30*, 539–552.

Miller, J. and Chapman, R. (1998). *Systematic analysis of language transcripts.* Madison: WI: Language Analysis Laboratory, Waisman Research Center.

Norbury, C. & Bishop, D. (2003). Narrative skills of children with communication impairments. *International Journal of Language and Communication Disorders, 38,* 287-313.

Oxford University Language Centre. (2010). *Oxford placement test: Spanish.* Retrieved from http://www.lang.ox.ac.uk/courses/tst_placement_spanish.html

Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B. (2004). "Frog where are you?": Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88,* 229-247.

Restrepo, M. and Gutierrez-Clellen, V. (2001). Article use in Spanish-speaking children with specific language impairment. *Journal of Child Language 28*(2), 433-452.

Roseberry-McKibbin, C., Brice, A., & O'Hanlon (2005). Serving English language learners in public school settings: A national survey. *Language, Speech, and Hearing Services in Schools, 36,* 48-61.

Scott, C., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43,* 324–339.

Steckol, K., & Leonard, L. (1979). The use of grammatical morphemes by normal and language- impaired children. *Journal of Communication Disorders, 12,* 291–301.

Stein, N., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 53-120). Norwood, NJ: Ablex.

Thomas, W., & Collier, V. (2002). A national study of school effectiveness for language minority students' long-term academic achievement. Santa Cruz, CA: Center for Research on Education, Diversity & Excellence.

 "UTBox;" http://www.utexas.edu/its/cloudstorage/.