The Thesis Committee for Maria Blokh
Certifies that this is the approved version of the following thesis:

Validity of Self-Ratings for Determining Language Proficiency: Evidence from Russian-English Bilingual Adults

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: _____
Thomas P. Marquardt

_____
Lisa M. Bedore

Validity of Self-Ratings for Determining Language Proficiency: Evidence from Russian-English Bilingual adults

by

Maria Blokh, A. B.

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 2013

Acknowledgements

I would like to thank Dr. Thomas P. Marquardt for the guidance, encouragement and interesting discussions over the course of this project.

I would also like to thank Dr. Lisa M. Bedore for her help in navigating the transcription, coding and other daunting aspects of this thesis.

Abstract

Validity of Self-Ratings for Determining Language Proficiency: Evidence from Russian-English Bilingual Adults

Maria Blokh, M.A.

The University of Texas at Austin, 2013

Supervisor: Thomas P. Marquardt

Narrative measures derived from English and Russian tell and retell narrative language samples of 20 L1-Russian, L2-English bilingual adults were correlated with their overall, speaking and verbal proficiency self-ratings to verify the validity of the self-rating scale for both languages. In English, measures of fluency, productivity and grammaticality were moderately correlated with speaking proficiency self-ratings. Strength of correlations with tell versus retell narratives varied by category of narrative measure. For Russian, correlations were not significant due to ceiling effects in proficiency. The effects of modifications to narrative measures were considered, showing that correlations with temporal fluency and productivity increased as mazes and fillers were excluded, while correlations with grammaticality increased as article omission errors were excluded. Sources of variation in self-ratings and narrative measures are described, and recommendations are presented for an alternative narrative elicitation method.

Table of Contents

List of Tables

Chapter 1: Introduction

RATIONALE

A speech-language pathologist assessing and treating an individual with an acquired language impairment faces an additional hurdle if this individual speaks two or more languages. This hurdle is the clinician's uncertainty as to the individual's premorbid proficiency in each language. If the patient's language ability has been impaired as a result of a cerebrovascular accident, primary progressive aphasia, traumatic brain injury, or any other aetiology, determining proficiency using a standardized language proficiency assessment or a language sample is not possible. The patient's premorbid proficiency must, however, be known to accurately classify the acquired bilingual language impairment (Kiran & Iakupova, 2011). Pre-stroke proficiency is also necessary for determining the language or languages to be used in intervention and the targets in each language. Reasonable targets in Spanish, and even the decision to treat in Spanish at all, would be very different for a patient who studied Spanish for two years in middle school and had not used it for 20 years than for a native Spanish speaker who used the language daily in both home and work settings. Descriptions of varying levels of proficiency in neurotypical adults are needed as a basis for classification of proficiency.

Speech-language pathologists often rely solely on language history and language use questionnaires to determine premorbid proficiency in each language. However, the questionnaires need to be quite detailed in order to produce an accurate estimate of proficiency, and the patient must provide precise and accurate responses to questions including the age at which they began to learn each language, the number of hours they speak each language per day, and the years of schooling in each language (citation).

Such facts may be difficult to recall, especially in the presence of cognitive-linguistic deficits, and others may not be able to accurately report this information for the patient. Moreover, the validity and sensitivity of these scales have not been firmly established, despite reports that instruments such as the Language Use Questionnaire (LUQ) in development are sensitive to differences in bilinguals' proficiency skills (Kiran & Iakupova, 2011).

Another option for determining premorbid proficiency is to use a self-rating scale in which the patient is asked to rate proficiency before their ability became impaired. Such self-ratings can be shorter than language use and history questionnaires, while allowing patients to rate several modalities and contexts of language use. As with any self-rating scale used to measure an objective value, human subjectivity can potentially reduce validity and reliability values to the extent that the rating is not a reasonable choice. For example, Selezneva (2009) used the scores for current language use to determine participants' proficiency in English and Russian, making no use of the collected proficiency self-ratings due to insufficient evidence of their accuracy in determining proficiency, particularly in Russian-English bilinguals (Selezneva, 2009). Even more positive claims are qualified. For example, while bilinguals' proficiency self-ratings predicted the same degree of dominance as the results of the rather subjective American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI), they did not reliably predict performance on a naming task (Gollan, Weissberger, Runnqvist, Montoya & Cera, 2012).

More evidence in support of proficiency self-ratings is needed. Due to the shortage of such evidence, studies of the accuracy of proficiency ratings as used by family members can also be considered. Recently two studies found such evidence to support the use of parent and teacher ratings of children's proficiency. Gutierrez-Clellen

and Kreiter (2003) reported significant correlations between teachers' proficiency ratings and students' grammatical performance on a narrative task, although ratings of students' language use were more reliable than ratings of proficiency. The investigators hypothesized that teachers' varying expectations for the English proficiency of bilingual students may have lowered reliability of proficiency ratings. Bedore, Pena, Joyner, and Macken (2011) found moderate correlations between parents' ratings and their children's' performance on both the semantics and morphosyntax sections of the Bilingual English and Spanish Assessment (BESA; Peña, Gutiérrez-Clellen, Iglesias, Goldstein, & Bedore, in development) and between teachers' ratings and their students' performance on the morphosyntax section. Evidence of the accuracy of child proficiency ratings by parents is promising for the validity of adult proficiency ratings by spouses and other family members.

There is some support for the use of proficiency self-ratings. In a study of the level of proficiency corresponding particular responses such as *very well* on a self-rating scale, Kominsky (1989) reported significant correlation between participants' scores on a test of English receptive skills and their self-rating scores. Marian, Blumenfeld and Kaushanskaya (2007) also found significant correlations between proficiency self-ratings and measures including receptive vocabulary and reading fluency. The primary purpose of the current study is to verify a proficiency self-rating scale's correlation with objective measures of expressive language to determine whether proficiency in each language can be confidently estimated given a bilingual patient's self-rating score.

OBJECTIVE MEASURES OF PROFICIENCY

A maximally accurate calibration of a proficiency self-rating scale should use the most valid objective measures of proficiency available. The selection of measures

should be guided by a clearly defined conception of proficiency, “one of the most poorly defined concepts in the field of language testing,” Farhady (1982, p. 44). Proficiency is not the same as language ability, or the capacity to attain mastery of a language. Farhady (1982) mentioned early definitions of proficiency as a person's achieved competence in using a language. More precisely, proficiency has been defined as “the extent to which a bilingual’s skills in one or both of their languages meet age-based native speaker or monolingual expectations,” including knowledge of vocabulary and of grammatical systems (Bedore et al., 2012, p. 617). While children’s proficiency is expected to improve with age as they progress through developmental stages of language development, adult proficiency remains rather stable once a steady state is reached.

Given that language ability is similarly determined by comparing an individual’s performance in expressive and receptive language tasks to non-clinical expectations for monolingual peers, (unless the measure is specifically designed to measure ability in bilinguals), the objective indicators of language ability should overlap with objective indicators of proficiency. If a non-impaired bilingual individual scores below average on a measure intended to identify language impairment in a particular language, a logical conclusion is that this individual has not achieved the level expected at his or her age for monolingual speakers of that language. In the absence of language impairment, this may be due to insufficient exposure to the language, a late age of acquisition or any other factors that limit proficiency in nonclinical populations (Abrahamsson & Hyltenstam, 2009). As a result, both language ability and language proficiency tests have been used to determine proficiency (Bedore et al., 2012). Indeed, a variety of studies have used measures intended, or commonly used, to evaluate ability in order to determine proficiency. For example, Bedore et al. (2011) correlated scores on the BESA, an assessment instrument intended to identify language impairment in bilingual children,

with parent and teacher ratings of children's proficiency in order to investigate the validity of these ratings.

While standardized assessments are used to evaluate proficiency, there is evidence that analysis of a language sample is in fact a more valid measure, due largely to its ecological validity (McMaster & Tilstra, 2007). For example, MacSwan (2001) found that a large proportion of children who made few grammatical errors in spontaneous language samples performed poorly on standardized tests, suggesting that such tests underestimate proficiency. MacSwan and Rolstad (2006) also found natural language sample analysis to be more empirically valid than standardized test scores.

Test items that indicate proficiency often coincide with those that indicate language ability. Naturalistic assessments such as analysis of a language sample are likely to show an even stronger relationship between proficiency and ability measures, since these have not been specifically developed to identify language impairment. Such analysis typically includes several measures of productivity, fluency, syntactic complexity and g rammaticality. The considerations that apply to developmental standardized tests also apply to narrative measures, and these measures should not be assumed to be a valid indicator of proficiency in adults simply because they increase developmentally in children. Due to variability in the definition and use of these measures, each will be defined as used in the current study, along with a justification of the measure's usefulness.

Productivity, or the amount of language produced by one speaker in a language sample, indicates the speaker's knowledge of the language, with standardization values available for monolingual and bilingual English speakers through the Systematic Analysis of Language Transcripts (Bedore, Pena, Gillam, & Ho, 2010). Language productivity composites can be used to determine language dominance, or relative proficiency

(Solorio et al., 2011). Measures of productivity include total number of utterances, total number of words and number of different words, with mazes optionally excluded from these counts (McMaster & Tilstra, 2007; Scott, Roberts & Krakow, 2008; Solorio et al., 2011).

Fluency measures are divided into two categories: measures of the frequency with which speech is disrupted by fillers and by mazes, which include repetitions, revisions or false starts, and measures of productivity over time (McMaster & Tilstra, 2007; Solorio et al., 2011). Speech disruptions reflect a speaker's semantic and morphosyntactic abilities (Scott et al., 2008; Solorio et al., 2011). Miller et al. (2006) found that temporal measures are related to proficiency in the non-native language, while Nicholas and Brookshire (1993) reported that such measures show greater stability in analyses of the language of aphasic adults than a simple count of words or utterances produced.

Syntactic complexity is most often measured as the average length of a unit of speech such as a T-unit (sentence) or C-unit (independent communicative unit), referred to as the mean length of utterance (MLU). Halleck (1995) found significantly higher syntactic complexity in Chinese students of English rated Superior than in students rated Advanced-Intermediate on the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI). Speech disruptions optionally can be excluded from these measures. Solorio et al. (2011) compared MLU results with maze words included and excluded.

Grammaticality measures are used widely to assess language knowledge, and a number of studies have shown that these measures correlate positively with language proficiency judgements (Bedore et al., 2010, p501). An investigation of validity of proficiency tests by MacSwan (2001) used grammaticality measures to "empirically confirm that children are competent in their native language" (Pray, 2005, p404).

Gutierrez-Clellen and Kreiter (2003) used grammaticality to evaluate the validity of parent and teacher ratings of children's proficiency.

While some studies measure grammaticality as the proportion of grammatically correct utterances (Gutierrez-Clellen & Kreiter, 2003; Halleck, 1995; Fiestas & Pena, 2004; Scott et al., 2008), others code each error individually (Rice & Wexler, 1996; Bedore & Leonard, 1998; Restrepo, 1998; McMaster & Tilstra, 2007). The latter approach may have higher validity by differentiating between utterances with only one or more than one error, as well as allowing for a count of particular kinds of errors, e.g. morphological errors only (Bedore et al., 2010). In fact, composites of select error types increase specificity and sensitivity when identifying language impairment in children (Bedore et al., 2010). For example, the contribution of morphosyntactic and semantic errors can be considered separately, following the finding by Bedore et al. (2011) that teacher proficiency ratings correlate with children's performance on the morphosyntax but not the semantics section of the BESA.

A combination of measures of productivity, fluency, sentence complexity and grammaticality is likely to be more valid that any one measure alone. T-unit analysis alone is not useful for determining children's second language proficiency, as it does not give sufficient consideration to morphology or vocabulary, and may excessively reward circumlocutions (Barnwell, 1988). Similarly, Ginther (2010) found that measures of fluency were correlated to scores on an oral proficiency test, but that additional measures were needed to accurately classify adult speakers at various levels of proficiency. Correspondingly, the current investigation considered evidence from each category of narrative measures – productivity, fluency, sentence complexity and grammaticality. Measures were correlated with ordinal proficiency self-rating scores in order to determine the validity of a proficiency self-rating scale for Russian-English bilinguals.

Chapter 2: Method

PARTICIPANTS

Data were collected from 22 Russian-English bilingual adult participants. Participants were born in Russian speaking countries, began learning English after the age of 6, and had no reported history of cognitive or language disorders. Table 1 presents a summary of participant demographic, language history and language use information. The percent Russian and English proficiency was determined by the number of hours each language was spoken during a typical week. The percent of time each language was spoken determined the bilingual type. *TRUE* referred to participants who used English and Russian each at least 20% of the time, and *Pred English* referred to participants who used Russian less than 20% of the time. Two participants' data were not included in the current analysis due to incomplete or highly unintelligible audio recordings of the narrative language samples.

Table 1: Participant demographics and language history and use, adapted from Selezneva (2009)

ID	Age	Gender	Country of Origin	Russian Proficiency (%)	English Proficiency (%)	More than 10 yrs Exposure to English	Bilingual Type	Years of Education
1	32	F	Russia	54	46	Yes	TRUE	20
2	34	F	Russia	13	88	Yes	Pred English	27
3	28	F	Ukraine	62	38	Yes	TRUE	18
4	22	M	Azerbaijan	26	74	Yes	TRUE	14
5	26	F	Russia	55	45	No	TRUE	20
6	34	F	Russia	36	64	Yes	TRUE	21
7	37	M	Russia	16	84	Yes	Pred English	12
8	38	F	Russia	44	56	Yes	TRUE	16
9	35	M	Russia	58	42	No	TRUE	18
10	36	F	Ukraine	30	70	Yes	TRUE	16
11	33	M	Russia	19	81	Yes	Pred English	12
12	34	M	Russia	54	46	Yes	TRUE	14
13	37	F	Russia	17	83	Yes	Pred English	22
14	46	F	Belarus	61	39	Yes	TRUE	15
15	26	F	Russia	40	60	No	TRUE	16
16	34	F	Russia	71	29	Yes	TRUE	15
17	24	F	Kazakhstan	64	36	Yes	TRUE	16
18	34	M	Russia	62	38	Yes	TRUE	15
19	37	F	Russia	59	41	Yes	TRUE	11
20	33	F	Russia	45	55	No	TRUE	19

MATERIALS

This study took the form of a secondary data analysis of a subset of the data collected by Selezneva (2009) using the Computerized Language Analysis (CLAN; used to analyze transcripts for the Child Language Data Exchange System [CHILDES]), Microsoft Word and Microsoft Excel programs. This subset consisted of two sets of

deidentified information: scores from self-ratings of proficiency in English and Russian and audio recordings of narrative language samples in English and Russian. Selezneva (2009) focused on the relationship between proficiency as determined by language use as shown in Table 1, and verbal fluency as determined by generative naming scores in both languages. Neither set of data was used in the current study.

PROCEDURES

Selezneva (2009) determined proficiency using participants' current language use scores, however the current study considered only proficiency self-ratings. In the original study, self-rated proficiency scores were collected for each language by asking participants to rate their ability to communicate in four speaking and listening situations: speaking in casual conversations, listening in casual conversations, speaking in formal situations, and listening in formal situations. In addition, participants rated their ability to communicate through reading and writing, and their overall communication ability in each language. They were informed that these scores would help the researcher "understand how comfortable you are in English and [Russian]." Scores ranged from 1 ("non-fluent, only know several words or a few simple sentences") to 5 ("fluent, completely comfortable with skills like a native speaker") (Selezneva, 2009).

Three values derived from these self-ratings were considered as possible indicators of proficiency: the rating for overall communication ability, or *overall proficiency*; the average of the two speaking communication ability ratings, or *speaking proficiency*; and the average of all four speaking and listening communication ability ratings, or *verbal proficiency*. The averages of *formal* and *casual* measures were taken as more thorough measures, as well as more representative of the current narrative task, which was neither formal nor completely informal. The interaction could be considered a

formal situation as it took place not among friends but rather as part of a systematic research investigation, but could also be considered an informal situation as the task was to retell a children's story in a low-pressure setting, and many participants incorporated humor into their story.

The speaking self-rating was selected due to its ostensibly direct relationship to performance on a narrative task. The verbal self-rating also was considered because listening and speaking ability are closely related, so that a participant who rated his speaking proficiency as equal to another's, but rated his listening proficiency as higher, may perform better on a narrative task. By the same token, reading and writing self-ratings, also collected in the questionnaire, could also have been included in the analysis, however these skills were judged to be too distantly related to verbal language skills to be useful for predicting spoken proficiency. Finally, the validity of the overall self-rating was investigated due to its utility as a simple self-rating value that may be easier to use for patients with deficits caused by a cerebrovascular accident.

The narrative language samples were collected through a tell and retell task in the original study. Participants first read aloud the text to one of two of the Frog Stories by Mercer Meyer – *Frog All Alone* or *Frog Goes to Dinner*. Next, participants created their own story based on wordless illustrations to one of two other Frog Stories: *Frog, Where are You?* or *One Frog too Many*. This invented story is subsequently referred to as the tell task. Participants took as much time as they wanted to first browse through the pictures. The researcher also encouraged them to look at the picture book while producing the narrative. Immediately following the tell task, the researcher inquired whether each participant remembered the story they had read, then asked them to do the best they can to reproduce this story. This narrative will subsequently be referred to as the retell task.

Transcription

A native speaker of Russian transcribed each Russian narrative produced during the tell and retell tasks. The narratives were transcribed following the Codes for the Human Analysis of Transcripts (CHAT) transcription conventions as used for the CHILDES database, described in MacWhinney (2000). A graduate student who had been trained in transcription and was a native speaker of Russian independently transcribed three randomly selected participants' tell and retell narratives to determine the reliability of the transcription. Independent inter-rater reliability was 98%, calculated as the total number of words agreed upon by both transcribers, divided by the total number of words in the original transcriptions and multiplied by 100. The two native Russian speakers agreed to use modified C-units when segmenting the narrative into utterances. The SALT transcription conventions for Spanish language samples justify the use of such modified C-units based on the optional use of a subject in Spanish. While subjects are less frequently omitted in Russian than in Spanish, they are not obligatory. For example, the following is a possible utterance in Russian, given that the subject is known from the context:

a) Sto^jali dolgo.
Stood long
"They stood for a long time."

Three undergraduate students, native speakers of English, transcribed each English tell and retell narrative. Four tell and four retell narratives also were transcribed by the researcher to determine the reliability of the transcriptions, with at least one tell and retell narrative chosen from each undergraduate student. Average reliability was 96%, calculated as the total number of words agreed upon by both transcribers, divided by the total number of words in the original transcriptions and multiplied by 100. The

majority of inconsistencies resulted from untranscribed fillers and from words and phrases deemed unintelligible by the undergraduate transcribers. After determining reliability, the researcher listened to each English narrative in order to correct inaccurate or incomplete transcriptions. Having grown up in a Russian-speaking family in America, the researcher can be assumed to be more familiar with both the grammatical and phonetic features of Russian-influenced English, allowing her to more accurately transcribe utterances with reduced intelligibility to non-speakers of Russian. Most corrections involved the addition of missing fillers and of words originally omitted due to poor intelligibility.

All fillers were transcribed as *uh* in order to facilitate analysis. The word *oh* was counted as a filler when it was used at the beginning of an utterance and indicated that the speaker had suddenly recalled a part of the narrative, e.g. “Oh, then the frog jumped out of the jar.” However, when the word *oh* was used in a different context, for example in a quote, e.g. “The boy yelled 'oh you bad frog!'” as it had greater semantic content than a filler in these cases.

Some participants expressed periodic uncertainty regarding the narrative, either due to difficulty remembering the story they had previously read during the retell, or difficulty interpreting the pictures during the tell. Participants also made other comments or expressed opinions and ideas about the story, which sometimes took the form of entire utterances and at other times were included in an utterance which related events in the story. Such meta-narrative elements were counted the same way as utterances in which the participant related events in the story. While these elements could be omitted as they were not technically part of the narrative, this would lead to an artificially lowered rate of speech, and would require separating them from the rest of the narrative. However, there

was not always a clear division, and both narrative and meta-narrative elements often were contained within a single utterance.

Coding

The narratives were coded using modified CHAT coding conventions as described in MacWhinney (2000). Modifications to the CHAT conventions served to allow for analysis of certain measures of fluency, productivity and grammaticality, as described in the literature review and the analysis sections.

Maze Coding

Table 2 presents a summary of the maze coding employed in the study. Grammatical mazes were differentiated to facilitate their analysis separate from other revisions, although this is not part of the CHAT coding conventions. In all mazes except false starts, the part of the utterance that was repeated or revised was enclosed in < >, allowing for a count of the number of words that were repeated or revised. False starts are different from the other mazes in that the utterance was not revised but rather discontinued.

Table 2: Maze coding used in narrative transcriptions

Sign	Name	Explanation of Use	Examples of Use
[/]	Repetition	Exact repetition	the boy was [/] the boy was
[//]	Minor Revision	Revision minimally changes semantics but does not change syntax of utterance.	<he> [//] the big frog <he leaped> [//] he jumped
[//g]	Grammatical Revision (included in minor rev)	English: self-correction of morphosyntax. Russian: possible self-correction of morphosyntax	<two dog> [//g] two dogs <catched> [//g] caught
[///]	Semantic Revision	Revision changes semantics but not syntax of utterance.	<walked in> [///] jumped up <the frog> [///] the boy
[/-]	Syntactic Break	Revision changes syntax of utterance, causes syntactic break with preceding phrase, without a pause.	<the boy left> [/-] it was time to go <and she> [/-] next to her sat a cat
[+...]	False start	Utterance is discontinued, new utterance produced after a pause.	the dog didn't see any [+...] they continued walking

There were very few clear cases of morphosyntactic self-correction in the Russian narratives due partly to the overall much lower number of morphosyntactic errors. Additionally, due to the complex nature of morphological agreement, most revisions could have been either morphosyntactic self-corrections or lexical revisions, as in the following case:

b) tam byl byla ^zhen'^scina.
there was (masc) was (fem) woman
"A woman was there"

In an utterance such this there are two possible interpretations: a) the participant intended

to say “a man was there,” used correct masculine agreement on the first “was,” and then made a lexical revision to “woman” requiring feminine agreement; b) the participant intended to say “a woman was there,” erroneously used masculine agreement on “was,” then made a grammatical revision. Therefore, the [//g] symbol was used in all such possible cases of morphosyntactic self-correction in order to determine if the occurrence of such revisions was correlated with participant proficiency self-ratings.

Error Coding

The coding of errors was modified. The CHAT coding conventions differentiate between the following five error categories, among others: semantic errors such as use of an incorrect word, morphological errors such as use of an incorrect suffix, formal lexical device errors such as use of wrong article or part of speech, omission errors such as omission of a verb or article, and utterance-level grammatical errors such as several missing prepositions or use of a double negative in English.

These five categories were used as a guide for error coding used in the current study, and the semantic and utterance-level errors were used following CHAT conventions, labeled as *lexical* and *phrase-level* errors. However several changes were made to facilitate analyses. For example, the function word category was used rather than the formal lexical device category to group prepositions, copula, articles and other function words together. A tense/aspect error category was included to differentiate between errors which could reflect a lack of understanding of tense or aspect categories, and errors of overregularization, which were coded as morphological errors because they indicated lack of knowledge of the form rather than the category. Also, omissions were included in either the lexical or function word category. Table 3 details the kinds of errors included in each category and provides examples in English and Russian phrases.

Table 3: Error coding used in narrative transcriptions

Category	Codes Included Within Category	Examples
Function word	Missing, extra or incorrect articles, auxiliary, copula (excluding agreement errors), prepositions, relative pronouns	*-The house big.* *-The frog what lived in the jar.* *-On voshol k komnatu.* *(He entered to room.)*
Tense/Aspect	Tense and aspect errors excluding overregularization	*-She will going.* *-Kazhdyj den' poshla tuda.* (*Every day went-perf there.)*
Morphology	Errors of overregularization, agreement (gender/number/case/...)	*-He catched the frog.* *-Ona prishol.* *(She came-masculine)*
Lexical choice	Incorrect (including code-switching, misworded expressions), non-standard choice	*-Out of a sudden.* *-On quickly ubezhal.* *(He quickly [code-switch] ran away.)*
Phrase structure	Word order, omitted and extra parts of speech, non-standard phrasing	*He felt himself guilty* *He so much disliked the little frog.*

Non-standard lexical choice and phrase structure were differentiated so that these could optionally be excluded from error counts, and were marked in utterances which were grammatically correct however noticeably non-nativelike. For example, in "They looked on the earth," *ground* should have been used instead of *earth*. Examples of non-standard lexical choice Other CHAT coding conventions were used without modification, such as denoting partial words by including the omitted portion of the word in parentheses.

ANALYSIS

The measures of fluency, productivity, sentence complexity and grammaticality were obtained from the transcribed and coded language samples using the CLAN, Microsoft Word and Microsoft Excel programs. All measures were calculated for each tell and retell individually, as the varying cognitive demands of creating a novel narrative and recalling a previously read narrative could lead to differences in performance in the tell versus the retell tasks.

The total number of words and number of different words were obtained using the "FREQ" function in CLAN. The occurrences of mazes and errors, total number of utterances and number of fillers were obtained through the *find* function in Microsoft Word. Other measures were derived by simple arithmetic applied to these obtained measures using Microsoft Excel. The number of words without fillers was used instead of, or in addition to, the total number of words in the calculations of rates and percentages, following the observation that participants who appeared equally proficient in a language varied widely in their use of fillers. This observation suggested that use of fillers may not be strongly correlated to language proficiency, and that inclusion of fillers in word counts may skew the results. The number of words excluding fillers in a narrative subsequently will be referred to as WEF.

Fluency

The fluency measures used are listed in Tables 4 and 5. Examples of each kind of maze are provided in Table 2.

Table 4: Measures of fluency as speech disruption rates in bilingual narrative samples

Measure	**Definition**
Mazes per utterance	Number of mazes divided by number of utterances
Mazes per WEF	Number of mazes divided by the number of words excluding fillers
Maze words per WEF	Number of maze words, or words within < >, divided by the number of words excluding fillers
Ratio of fillers to words	Number of fillers divided by the number of words
Revisions per WEF	Number of minor and semantic revisions divided by number of words excluding fillers
Repetitions per WEF	Number of occurrences of repetition, not counting multiple iterations, divided by number of words excluding fillers

Table 5: Measures of fluency as productivity over time in bilingual narrative samples

Measure	**Definition**
Words per second	Number of words divided by narrative length in seconds
WEF per second	Number of words excluding fillers divided by narrative length in seconds
Productive words per second	Number of words excluding fillers and maze words divided by narrative length in seconds

Productivity

The measures of productivity used in the current study are listed in Table 6.

Table 6: Measures of productivity in bilingual narrative samples.

Measure	**Definition**
WEF	Number of words excluding fillers
Utterances	Number of utterances
Different words	Number of different words
Productive words	Number of words excluding fillers and maze words, or words within < >

The FREQ function in CLAN produced a count of total number of words and number of different words in each narrative, along with a list of each word and the number of occurrences in the sample. However, the CLAN program considers two words with any difference between them to be different words, so, for example, the singular *bee* and the plural *bees* both contribute a word, or a token, to the count of different words, whereas the singular and plural forms of a word would typically contribute a single token to a count of lexical diversity. In order to obtain an accurate count, the list of words produced for each narrative was examined for instances where two or more words differed only in inflectional morphology or in one of several other features described

below, (e.g. *dog* and *dog's, want,* and *wanted*) and the number of such instances was subtracted from the given number of different words to produce an accurate value for the number of different words. Instances of different inflectional morphology in Russian and English included case, gender, tense, aspect, plurality, diminutives, comparative and superlative forms of adjectives, agreement for any of these and reflexivity when it did not change the meaning. Instances of phonological variants of a single word, such as *a* and *an* in English and *s* (with) and *so* in Russian counted as a single token. Two words that differed in derivational morphology counted as two separate tokens, even when the morphological rule was largely regular, such as the use of *-ly* to form adverbs from adjectives in English. This was decided firstly because even regular derivational rules are less widely applicable; for example *youngly* cannot be derived from the adjective *young.* Secondly, a noun and verb which share a common root are more "different," both syntactically and semantically, than two verbs that are of different tense and can therefore appear in the same position in a sentence, or even two nouns which are of different case and can appear only as subject or only as object of a sentence.

Two complicating factors in Russian are the use of prepositions as verb prefixes, and the use of emphatic particles as postfixes. Two cases of such prefixes are shown in the following examples:

c) L^jagu^shka provela odin ^chas v prudu.
Frog through+led one hour in pond
"The frog spent one hour in the pond."

d) P^chela vyletela iz uli^ja.
Bee out+flew out hive.
"The bee flew out of the hive."

In the above sentences, the preposition indicating direction of motion becomes a prefix in Russian, even when it is used in addition to an actual preposition as in example d). However not all meanings are transparent, since "most of [the prefix and verb combinations] have one or more abstract meanings whose connection with the primary

sense...may vary from fairly obvious to remote or unestablishable," as in example c) (Townsend, 1975, p.20). One option is to count all verbs that differ only in prefix as a single token of this verb. Another option is to count the prefix separately from the verb, and to consider each prefixed verb as contributing two tokens. In the current study, each novel prefixed verb counted as a single, unique token, largely due to the non-transparent meaning of many such prefixed verbs. To further complicate matters, some prefixes also are used to make an imperfective verb perfective without changing any other aspect of the verb semantics. The most common such prefix is "po-." Since aspect is an obligatory feature of Russian verbs, the imperfective and perfective form of a verb counted as a single token.

A similar dilemma was posed by the presence of emphatic post-positive particles. These particles, including -zhe and -to, are added to question words and certain other words in order to convey various shades of emphasis or indicate the subjunctive mood. However, similarly to the directional prefixes above, these particles have a non-transparent meaning when used with certain words, making it unreasonable to count all uses of one particle as one token. Therefore, for consistency, each novel word-particle combination counted as a unique token.

Sentence Complexity

In the current study, a single measure of syntactic complexity was used: the number of productive words per utterance, or the productive mean length of utterance (MLU). Productive words include all words that are not fillers and are not part of a maze. MLU in words has been shown to be a more reliable measure of productivity cross-linguistically than MLU in morphemes, therefore the former was used (Devescovi et al., 2005). Post-positive particles were not treated as separate words, and were

connected to the preceding word using the + symbol following the CHAT transcription conventions for compound words, e.g. *ty+li*, (you, subjunctive).

Grammaticality

The measures of grammaticality used in the current study are listed in Table 7.

Table 7: Measures of grammaticality as error rates in bilingual narrative samples

Measure	**Definition**
Errors per WEF	Number of errors divided by number of words excluding fillers
EEOA per WEF	Number of errors excluding article omissions divided by number of words excluding fillers
EEOA excluding non-standard phrasing/ lexical choice per WEF	Number of errors excluding article omissions and non-standard phrasing and lexical choice divided by number of words excluding fillers
Lexical errors per WEF	Number of lexical errors divided by number of words excluding fillers
Phrase-level errors per WEF	Number of phrase-level errors divided by number of words excluding fillers
EEOA per utterance	Number of errors excluding article omissions divided by number of utterances
Morphosyntactic errors and revisions per WEF	Number of function word, tense/aspect, morphological and phrase-level errors and grammatical revisions divided by number of words excluding fillers

Rather than identifying each utterance that contained one or more errors, each error was identified individually following Rice and Wexler (1996), Bedore and Leonard (1998) and Restrepo (1998). During the transcription and error coding process, a significant proportion of the participants did not use the obligatory definite article in English when referring to the characters in the narrative. For example, many participants referred to the boy as *boy* and referred to the turtle as *turtle* omitting the expected article.

Omission of both definite and indefinite articles is influenced by the lack of articles in Russian, however it is unclear why definite articles preceding references to characters were especially likely to be omitted. A possible explanation is that some participants treated these nouns as proper nouns, e.g. *Boy* and *Turtle*. Whatever the reason, participants who omitted definite articles preceding characters tended to be significantly more accurate in their use of articles with other nouns. Following this observation, the number of omitted articles was excluded from several error counts, for example the total number of errors excluding omitted articles (EEOA). Words in mazes were not excluded from error counts as a large proportion of errors occurred in such mazes, whether or not they were subsequently corrected in a revision.

Chapter 3: Results

The data collected from each participant's narrative, and from which all the measures described in the analysis section can be derived, are presented in Appendix A along with group means for each variable. Each participant's proficiency self-ratings are listed in Appendix B. Tables 8 through 12 contain the Pearson's r correlation values for all measures detailed above with verbal and speaking self-ratings for English and Russian tell and retell separately. No corrections were used for multiple statistical comparisons. Correlations with overall proficiency self-ratings were not reported.

Table 8. Pearson's r correlations of speech disruption fluency measures with proficiency self-ratings in English-Russian bilinguals, WEF = words excluding fillers.

	English				Russian			
	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking
Mazes per utterance	-.26	-.42*	-.33	-.46*	.03	.01	.16	.13
Mazes per WEF	-.33	-.46*	-.34	-.48*	-.04	-.06	.12	.10
Maze words per WEF	-.35	-.45*	-.31	-.43*	.14	.11	.10	.06
Fillers per word	-.24	-.28	-.16	-.28	.17	.20	.12	.17
Revisions per WEF	-.31	-.39*	-.46*	-.56**	-.12	-.14	.02	-.01
Repetitions per WEF	-.29	-.39*	-.07	-.21	-.02	-.03	.26	.24

* = p <. 05, ** = p < .01

Table 9. Pearson's r correlations of temporal fluency measures with proficiency self-ratings in English-Russian bilinguals, WEF = words excluding fillers.

	English				Russian			
	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking
Words per second	.30	.40*	.11	.33	.09	.15	.01	.04
WEF per second	.31	.41*	.11	.33	.06	.12	0	.01
Productive words per second	.33	.43*	.15	.37	.04	.10	-.01	.01

* = p <. 05

Table 10. Pearson's r correlations of productivity measures with proficiency self-ratings in English-Russian bilinguals, WEF = words excluding fillers.

	English				Russian			
	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking
WEF	.14	.24	.29	.38*	.05	.10	0	-.07
Utterances	.08	.19	.28	.32	-.16	-.13	-.21	-.12
Different words	.18	.24	.39*	.47*	.01	.06	-.03	-.08
Productive words	.17	.27	.32	.42*	.04	.10	-.01	-.07

* = p <. 05

Table 11. Pearson's r correlations of sentence complexity measure with proficiency self-ratings in English-Russian bilinguals.

	English				Russian			
	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking
Productive MLU	.21	.17	.03	.11	.26	.29	.25	.28

Table 12. Pearson's r correlations of grammaticality measures with proficiency self-ratings in English-Russian bilinguals, WEF = words excluding fillers, EEOA = errors excluding article omission, x = undefined correlation.

	English				Russian			
	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking	Tell Verbal	Tell Speaking	Retell Verbal	Retell Speaking
Errors per WEF	-.07	-.11	.11	.13	.06	.05	-.22	-.23
EEOA per WEF	-.41*	-.45*	-.30	-.29	.06	.05	-.22	-.23
EEOA excluding non-standard phrasing/ lexical choice per WEF	-.42*	-.46*	-.30	-.29	-.04	-.06	-.22	-.23
Lexical errors per WEF	-.55**	-.53**	-.08	-.10	.16	.18	.16	.12
Phrase-level errors per WEF	-.37	-.44*	-.06	-.06	x	x	-.09	-.19
EEOA per utterance	-.42*	-.46*	-.31	-.30	.08	.08	-.20	-.20
Morphosyntactic errors and revisions per WEF	-.30	-.36	-.31	-.29	.14	.10	.17	.15

* = p <. 05, ** = p < .01,

CORRELATIONS IN RUSSIAN

The participants' proficiency in Russian as indicated both by self-ratings and performance on narrative tasks was quite uniformly high, reducing the potential for meaningful correlations between these two measures. In fact, none of the correlations between Russian narrative measures and self-ratings reached significance, and many were opposite of the expected direction. For example, while higher rates of errors at various levels of speech would be expected to correspond to lower proficiency self-ratings, the inverse pattern was shown by 5 out of the 7 correlations between grammaticality and proficiency self-ratings in Russian, with positive correlations including r = .05 for total errors and r = .18 for lexical errors. The correlation of speaking proficiency self-ratings with sentence complexity was the closest to reaching significance in Russian at r = 0.26, suggesting that MLU in words may have been

sensitive to variation in proficiency even among a highly uniform sample. However the correlation with sentence complexity also failed to reach $p < .05$, so no decisive conclusions can be drawn.

CORRELATIONS IN ENGLISH

English Correlations by Proficiency Rating and Task

Seventeen of 20 participants rated their overall English proficiency as 4. As was the case with all Russian self-ratings, this lack of variance reduced the meaningfulness of correlations with English overall proficiency self-ratings, therefore these correlations are not reported. Self-ratings of verbal proficiency, the average of casual and formal listening and speaking self-ratings, and speaking proficiency, the average of casual and formal speaking ratings, were sufficiently varied, however, and significantly correlated with performance in 3 out of 4 categories of narrative measures. All but two English narrative measures were more highly correlated with the speaking than the verbal self-rating, which averaged speaking and understanding. While 11 English narrative measures were significantly correlated with speaking but not verbal self-ratings in tell and/or retell, none of the measures were significantly correlated with verbal but not speaking self-ratings. The speaking self-rating was used in the correlations reported below unless otherwise noted.

A comparison of correlations of proficiency self-ratings with tell versus retell narrative measures shows inconsistent results. Both tell and retell correlations were comparable for speech disruption rates and sentence complexity; correlations were significant for tell only for temporal fluency and grammaticality measures, and significant for retell only for productivity measures.

Correlations with Fluency

All tell and/or retell speech disruption rates except the use of fillers per word were significantly negatively correlated to speaking proficiency self-ratings. These rates included both revisions and repetitions, with r = -.39 for each, and both the number of occurrences of mazes and the total number of words in all such occurrences, with r = -.46 and r = -.45

All temporal measures of fluency from the tell task, but not retell task, were significantly negatively correlated with speaking proficiency self-ratings. Correlations increased from -0.40 to -0.43 as fillers and maze words were excluded from the number of words, suggesting that a temporal measure of a participant's productive words is a better indicator of proficiency.

Correlations with Productivity

Three measures of productivity from the retell task, but not tell task, were significantly positively correlated with speaking proficiency ratings. The number of different words was most indicative of proficiency, at r = 0.47. As with temporal fluency measures, correlations increased when maze words were excluded from the count of words, from r = .38 to 0.42.

Correlations with Sentence Complexity

The positive correlation of sentence complexity with proficiency self-ratings failed to reach significance, with $p > .05$. Interestingly, this was the only correlation which was higher for Russian than for English, with r = .29 for Russian compared with r = .17 for English. Sentence complexity was also the only category of narrative measures that did not prove to be at least moderately correlated with speaking proficiency.

Correlations with Grammaticality

Grammaticality measures were significantly negatively correlated with speaking proficiency self-ratings in the tell task but not the retell task. An especially large discrepancy was found between tell and retell in the correlation of lexical choice errors and phrase-level errors with speaking proficiency self-ratings. While both of these measures were significantly correlated with proficiency self-ratings in the tell task, reaching r = -0.44 and r = -0.55, correlations were at or below -.10 in the retell task. Discrepancies were smaller for the composite error rates. Among these, the exclusion of article omission from the error count greatly increased the correlations with proficiency self-ratings, from r = -0.11 to r = -0.45. These errors were excluded following the observation that some participants systematically omitted articles before character names such as *boy* and *dog* apparently treating these as proper names. The exclusion of non-standard choices in lexicon and phrasing from error counts had a small effect on correlations, likely due to their low overall occurrence. The rate of all morphosyntactic errors and grammatical revisions in the tell task was short of producing a moderate correlation with proficiency self-ratings, at r = -.36.

Chapter 4: Discussion

Correlations between participants' proficiency self-ratings and objective narrative measures in English and Russian were calculated to examine the validity of the proficiency self-rating scale. No correlations were found to be significant in Russian, and three out of four categories of narrative measures produced moderate correlations in English. The weak correlations between Russian proficiency self-ratings and narrative measures are likely due to participants' consistently high performance on narrative tasks and equally consistently high proficiency self-ratings. In other words, the participants spoke Russian too well to identify a relationship between their reported and actual proficiency. As a result, evidence of the validity of proficiency self-ratings was found only for participants' second language, English, so it not possible to make any conclusions about use of self-ratings for determining first-language proficiency. This highlights one of the limitations of the current study: the use of only L1-Russian, L2-English participants, all of whom were highly proficient in Russian. The inclusion of participants who were L1-English, L2-Russian, or whose Russian proficiency was more variable, would have allowed separate analysis of language history, language and proficiency variables. Only proficiency in L2 English could be examined, and the following discussion focuses on this variable alone.

Measures from three of the four categories of narrative measures - productivity, fluency and grammaticality, but not sentence complexity - were found to be moderately correlated with participants' speaking proficiency self-ratings. While narrative measures derived from natural language samples are among the most valid measures of language proficiency available, they are necessarily only estimates of proficiency. Studies continue to explore the effects of manipulating the particular counts and rates used in

order to determine the most valid among them, for example, by "allowing mazes to count towards…MLU and TNW, versus the standard practice of ignoring them" (Solorio et al., 2011). The current study, while assuming a generally acceptable level of validity of narrative measures, aimed to contribute to these explorations by considering a small set of measures in the categories of productivity, fluency and grammaticality. For example, speech disruptions were measured per utterance and per word; fillers or both fillers and mazes were excluded from measures of productivity; lexical and phrase error rates alone were considered as measures of grammaticality along with composite error rates. Despite such variations in the measures, moderate-to-high correlations between narrative measures and proficiency self-ratings were taken to indicate the validity of the rating scale, as in Gutierrez-Clellen and Kreiter (2003) and Bedore et al. (2011), where narrative measures or standardized tests were used to determine the validity of parent and teacher proficiency ratings of children. Further justifying such conclusions, categories of narrative measures produced comparable correlations with proficiency self-ratings; for example all English tell temporal fluency measures produced correlations between 0.41 and 0.43, and all but one English tell grammaticality measure produced correlations between -0.36 and -0.53.

Speaking proficiency self-ratings were most representative of participants' performance on the narrative tasks. Overall proficiency self-ratings were not variable enough to produce meaningful correlations, and correlations with verbal proficiency were almost uniformly lower than correlations with speaking proficiency. This pattern could be expected because performance on both tell and retell tasks relied primarily on expressive rather than receptive language skills, despite the element of reading comprehension in the retell task, and only the speaking proficiency self-rating measured expressive language exclusively.

The value of the overall proficiency self-rating was particularly low due to the format used in the current study: an integer between 1 and 5 inclusive. Speaking and verbal proficiency self-ratings were calculated by averaging either two or four responses, allowing for greater variability of responses and therefore higher correlations (e.g. [3+4]/2 = 3.5, [4+4+4+5]/4 = 4.75). A measure allowing greater precision and variability of responses, such as a visual analog scale, may increase the strength of correlations between the overall proficiency self-rating and performance on narrative tasks to a point where this single self-rating would be sufficient to determine a person's expressive ability. However, even if such a rating were available, expressive and receptive skills may not be equivalent, and information about each is valuable to a clinician, if the participant is able to rate them separately.

While measures derived from performance on all narratives produced significant correlations with proficiency self-ratings, the specific correlations that reached significance were different for tell and retell tasks. For example, correlations with grammaticality measures were significant for the tell task only, while correlations with productivity measures were significant for the retell task only. While the reasons for this particular pattern are unclear, several differences between tell and retell tasks may have contributed to differences between correlations. First, while the tell task places greater demands on participants' ability to interpret images and structure a novel narrative, the retell task places greater demands on participants' memory. Participants frequently paused during the retell task and reported difficulty remembering the story, sometimes proceeding to change or omit a substantial part of the narrative. Participants also had the option of repeating vocabulary or even entire phrases that they had read, and with which they were otherwise not very familiar, potentially increasing the number of different words and other narrative measures. Indeed the conclusion, much of the dialogue, and

certain other memorable phrases were repeated literally or almost literally by many participants in both English and Russian retells. How demands on memory versus demands on interpretational and organizational abilities would affect narrative measures to produce the pattern of results for tell and retell is not determined. Since a narrative tell is the more commonly used method to elicit a natural language sample in clinical settings, results for the tell task should be given priority.

The decision to exclude fillers from word counts when calculating certain fluency, productivity and grammaticality measures appears justified by the weak correlation between rate of the use of fillers and proficiency self-ratings. Use of fillers may have been affected by factors unrelated to proficiency, such as difficulty recalling the narrative during the retell task or difficulty inventing a novel narrative during the tell task. While these factors likely affected rates of maze production as well, these rates appear to relate more closely to proficiency than the use of fillers. Given that the number of maze words decreased as proficiency self-ratings increased, it follows that the total number or rate of productive words alone would be a more robust measure of proficiency than the total number or rate of words including mazes. Indeed, correlations with measures of both productivity and temporal fluency were strongest when both fillers and mazes were excluded.

Variation in self-ratings or in narrative performance not directly related to language proficiency likely weakened the relationship between the two. Sources of such variation in narrative performance potentially included the following: varying levels of effort and anxiety due to the importance ascribed to the task by each participant; varying ability to recall the story during the retell task; varying length of time taken to formulate a story before beginning the tell task. While all participants were told to take the time to look through the pictures and then begin telling the story, some participants began talking

almost as soon as they were given the picture book, while others spent 3-4 minutes flipping through the pictures. Another source of variation is the number and length of pauses, which were not counted and may have significantly increased the narratives' length in seconds.

There are several potential sources of variation in proficiency self-ratings. Participants may have varying expectations of their own language proficiency, or the proficiency of any non-native English speaker. These expectations may be affected by work setting. For example a participant who frequently attends academic presentations at a University may have set higher standards for English proficiency than a participant who works as a child care provider. Although participants were asked to rate their proficiency in formal and casual settings separately, a participant who prioritizes performance in a formal work setting may be led to have overall higher expectations even when considering casual speaking situations. Other situational factors that could affect participants' perception of proficiency include whether they frequently speak English with other non-native speakers and whether they are regularly made aware of their grammatical errors or non-standard lexical choices by critical listeners such as their children.

CONCLUSION

This intention of this study was to determine whether a bilingual proficiency self-rating scale can be reliably used to estimate a bilingual's level of expressive language proficiency in each language. This goal was addressed by correlating proficiency self-rating scores with a series of narrative language measures. The self-rating scores included three alternatives: overall proficiency ratings, the average of formal and informal speaking proficiency ratings (speaking), and the average of formal and informal

speaking and understanding proficiency ratings (verbal). The narrative measures were taken from tell and retell narratives produced by L1-Russian, L2-English speakers, and included measures of fluency as rate of speech disruption and rate of speech production, productivity, sentence complexity and grammaticality. Several alternative measures of fluency, productivity and grammaticality were considered in order to explore the effects of excluding vs. including fillers from word counts, considering different kinds of error separately vs. together, measuring fluency and grammaticality rates per word vs. per utterance, and other modifications of measures typically used in clinical and research settings.

Limitations of the study included the lack of meaningful correlations between Russian self-ratings and narrative measures due to ceiling effects in Russian proficiency. Correlations with English overall proficiency ratings were also discarded due to lack of variation in ratings; however both speaking and verbal self-ratings produced moderate correlations with narrative measures of fluency, productivity and grammaticality. All participants were L1-Russian, L2-English, limiting the possible analyses of language, language history and proficiency variables. Validity of results was further reduced by the effects of factors other than language proficiency on both self-ratings and narrative measures.

Factors which may have impacted narrative measures include the effort required to recall the retell, the time and effort spent formulating a tell and the number of pauses during both tell and retell. Narrative elicitation protocols with less susceptibility to variation due to factors other than proficiency can improve the value of findings in subsequent studies of self-rating scales. The effects of demands on memory during the retell task were apparent even in the non-disordered participants of this study, suggesting that performance on a retell may not reliably indicate proficiency. Validity of the tell

task should be further increased by designating a minimum of one minute for considering the sequence of images, given the observation that some speakers began narrating very soon, without first formulating a complete narrative. The tell task should be more relevant to the adult population; for example the story could involve adults interacting with colleagues at work or with their children at home, rather than pets and other animals. The images should convey a narrative that is transparent to avoid high demands on participants' organizational abilities while providing many elements for elaboration so that participants who are less creative can produce a sufficiently long language sample. Alternatively, a single detailed picture can be used to elicit a shorter description, such as the cookie theft picture in the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). Both insufficiently long and unnecessarily long language samples can be avoided by setting a suggested time range of two to four minutes for each narrative. Most of the narratives in the current study were within that time range, and those that lasted less than two or more than four minutes tended to be excessively short or long. The use of such a narrative elicitation protocol should result in greater validity of correlations between self-ratings and objective measures.

Participants' proficiency self-ratings were based on their expectations of proficiency, which may have been affected by their conceptualization of formal versus casual settings and their linguistic environment and requirements at work and at home. Furthermore, ratings were restricted to the numbers 1 through 5 inclusive, allowing little variability within low, medium and high proficiency levels. Greater precision may be achieved by using a visual analog scale rather than an ordinal scale, potentially resulting in more meaningful correlations between objective and self-rated proficiency measures in future studies.

Appendix A: Non-derived Data Collected From Narratives

Table 13: Non-derived narrative measures for English tell task

Participant	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	19	20	21	22	ave	stdev
Utterances	59	51	61	36	56	71	58	49	42	25	34	36	41	21	54	50	64	62	43	43	48	13.3
Words	525	462	519	467	495	635	683	387	405	206	348	403	438	236	580	542	807	606	545	384	484	144
Fillers	13	18	13	8	12	5	47	2	6	6	24	31	17	1	73	23	0	36	35	1	19	18.5
Different words	143	128	143	154	149	153	175	134	119	75	118	120	139	105	145	162	214	184	144	122	141	29.7
Time in seconds	234	199	208	240	213	248	366	203	180	134	149	215	263	156	447	347	279	210	320	178	240	79.2
Repetitions	0	4	8	13	3	3	5	3	4	4	2	5	6	3	7	15	1	6	6	1	5	3.7
Words in all repetitions	0	5	10	20	5	3	5	8		18	2	7	7	6	8	19	1	10	7	1	7.5	5.9
Minor semantic revisions	0	4	11	5	8	7	6	6	3	1	6	3	4	0	15	10	5	0	6	3	5.2	3.9
Words in all minor semantic revisions	0	6	22	10	21	14	18	13	3	1	14	7	11	0	30	15	16	0	11	8	11	8.2
Major semantic revisions	1	0	0	1	0	0	0	1	2	0	1	1	0	0	2	2	2	0	0	0	0.7	0.8
Words in all major semantic revisions	1	0	0	1	0	0	0	1	2	0	2	1	0	0	2	2	4	0	0	0	0.8	1.1
Syntactic breaks	0	0	0	0	0	1	1	1	0	0	2	0	0	1	0	0	0	0	0	0	0.3	0.6
Words in all syntactic breaks	0	0	0	0	0	4	1	4	0	0	3	0	0	3	0	0	0	0	0	0	0.8	1.4
Trail-offs	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.27
All errors	10	44	5	4	16	21	8	33	55	7	20	29	28	7	52	6	9	17	23	24	21	15.5
Omitted article errors	3	31	0	0	1	11	3	9	29	2	12	16	17	1	20	1	2	7	2	3	8.5	9.6
Lexical choice errors	0	2	1	0	1	4	3	4	1	0	0	0	2	1	8	1	2	3	13	5	2.6	3.2
Non-standard lexical choice	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0.3	0.6
Function word errors	7	39	1	2	6	14	4	15	41	3	14	24	20	2	34	4	4	12	7	12	13.3	12.4
Tense/aspect errors	1	3	1	2	3	3	0	11	10	2	3	4	4	3	6	1	1	1	1	6	3.3	3
Morphology errors	1	0	1	0	6	0	1	4	2	2	2	0	2	1	2	0	1	0	1	0	1.3	1.5
Phrase-level errors[3]	1	0	1	0	0	2	0	0	0	0	0	1	0	0	4	0	1	1	1	1	0.7	1
Non-standard phrasing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.1	0.2

Table 14: Non-derived narrative measures for English retell task

Participant	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	19	20	21	22	ave	stdev
Utterances	54	40	38	17	36	64	29	44	30	15	20	32	36	14	21	42	29	11	22	28	31.1	13.7
Words	535	312	366	197	335	537	313	360	266	172	222	330	397	197	258	491	425	130	289	254	319.3	115.5
Fillers	18	13	5	12	6	13	24	5	10	4	17	21	27	8	46	34	16	7	22	8	15.8	10.9
Different words	135	117	116	77	110	162	99	126	86	71	88	115	135	93	83	144	151	64	102	95	108.5	27.5
Time in seconds	236	127	129	148	133	275	156	147	129	90	135	151	196	114	211	270	149	41	166	123	156.3	57.2
Repetitions	2	2	4	4	2	5	0	4	8	1	7	5	4	1	4	7	2	2	2	2	3.4	2.2
Words in all repetitions	6	2	5	5	4	7	0	6	5	1	9	7	4	2	4	19	3	3	3	2	4.9	4.0
Minor semantic revisions	1	0	10	1	3	6	0	6	0	3	6	5	5	2	8	12	4	1	4	6	4.2	3.4
Words in all minor semantic revisions	2	0	26	3	4	19	0	12	0	4	11	11	9	4	12	31	5	1	9	10	8.7	8.5
Major semantic revisions	0	0	0	1	0	1	0	0	0	1	0	2	0	1	3	0	0	0	0	1	0.5	0.8
Words in all major semantic revisions	0	0	0	2	0	2	0	0	0	1	0	2	0	1	4	0	0	0	0	1	0.7	1.1
Syntactic breaks	0	0	1	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0.3	0.6
Words in all syntactic breaks	0	0	2	0	0	5	0	0	0	0	0	15	0	0	0	0	0	0	0	0	1.1	3.5
Trail-offs	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0.2	0.5
All errors	10	26	6	0	4	6	1	31	22	12	16	21	23	5	19	8	6	4	7	11	11.9	8.9
Omitted article errors	2	19	1	0	1	3	1	6	4	2	6	10	7	0	9	2	2	2	0	2	4.0	4.6
Lexical choice errors	0	0	1	0	0	1	0	0	1	0	0	0	2	2	1	3	0	0	1	1	0.7	0.9
Non-standard lexical choice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0
Function word errors	6	26	3	0	3	3	1	13	9	5	7	16	13	3	12	4	3	2	3	6	6.9	6.3
Tense/aspect errors	1	0	0	0	0	1	0	12	9	4	6	3	1	0	2	2	0	1	0	1	2.2	3.3
Morphology errors	0	0	0	0	1	0	0	5	2	3	3	0	4	0	3	0	1	1	2	0	1.3	1.6
Phrase-level errors[3]	3	0	2	0	0	1	0	1	1	0	0	2	3	0	1	0	2	0	0	3	1.0	1.1
Non-standard phrasing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0

Table 15: Non-derived narrative measures for Russian tell task

Participant	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	19	20	21	22	ave	stdev
Utterances	62	63	75	23	41	85	74	49	57	30	42	87	61	38	82	82	81	73	40	64	60.5	19.7
Words	303	358	411	204	291	501	609	436	264	224	234	571	482	323	475	629	826	534	303	367	417.3	161.9
Fillers	1	16	1	1	3	0	87	1	2	19	13	14	12	0	20	6	3	28	1	4	11.6	19.6
Different words	137	143	163	105	118	202	194	126	127	90	108	226	180	167	206	239	333	228	138	156	169.3	58.4
Time in seconds	169	158	230	132	144	241	469	213	153	182	113	281	288	254	351	408	370	223	137	210	236.3	99.0
Repetitions	1	0	0	3	0	3	5	5	0	6	1	5	2	0	5	7	2	2	1	2	2.5	2.3
Words in all repetitions	1	0	0	4	0	4	7	5	0	11	1	7	2	0	5	10	7	3	1	2	3.5	3.4
Minor semantic revisions	1	4	3	4	2	5	8	4	4	4	5	7	3	1	10	10	4	2	3	3	4.4	2.6
Words in all minor semantic revisions	3	8	5	10	5	10	18	7	12	6	10	12	6	1	16	20	8	2	6	4	8.5	5.2
Major semantic revisions	0	0	0	0	0	1	0	0	0	0	0	1	2	0	3	0	1	0	1	2	0.6	0.9
Words in all major semantic revisions	0	0	0	0	0	2	0	0	0	0	0	1	9	0	3	0	1	0	1	2	1.0	2.1
Syntactic breaks	0	1	0	0	0	1	0	3	1	0	0	1	1	2	2	0	0	0	0	0	0.6	0.9
Words in all syntactic breaks	0	1	0	0	0	1	0	9	2	0	0	1	5	9	11	0	0	0	0	0	2.0	3.5
Trail-offs	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0.1	0.3
All errors	1	0	0	0	3	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0.4	0.8
Lexical choice errors	0	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	3	0	0	0.4	1.1
Non-standard lexical choice	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0.2	0.7
Function word errors	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2
Tense/aspect errors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0
Morphology errors	0	0	0	0	2	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0.3	0.6
Phrase-level errors[3]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0
Non-standard phrasing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0

Table 16: Non-derived narrative measures for Russian retell task

Participant	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	19	20	21	22	ave	stdev
Utterances	60	59	81	30	46	61	30	46	30	23	42	30	28	13	51	64	51	53	23	49	43.5	17.2
Words	312	362	400	170	259	394	238	255	173	147	256	158	219	136	373	354	398	457	138	286	274.3	102.1
Fillers	2	17	1	6	6	6	32	6	1	8	18	6	16	1	23	18	7	30	10	11	11.3	9.3
Different words	143	130	175	83	117	162	87	113	83	76	107	88	126	79	157	170	178	201	68	130	123.7	40.1
Time in seconds	161	161	232	104	138	192	160	139	118	165	121	88	149	86	277	248	155	242	79	162	158.9	55.8
Repetitions	1	2	2	3	2	0	1	3	6	4	3	0	3	0	3	3	2	5	1	1	2.3	1.6
Words in all repetitions	1	2	3	5	3	0	1	3	7	5	3	0	4	0	3	3	2	5	1	1	2.6	1.9
Minor semantic revisions	1	4	1	3	2	2	0	7	3	2	4	2	4	0	13	6	5	0	2	2	3.2	3.0
Words in all minor semantic revisions	1	6	1	6	2	5	0	8	6	3	4	3	6	0	24	14	12	0	2	4	5.4	5.8
Major semantic revisions	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	2	0	0	0	1	0.4	0.6
Words in all major semantic revisions	0	0	0	0	1	2	0	1	0	0	0	0	0	0	2	2	0	0	0	1	0.5	0.8
Syntactic breaks	0	1	0	2	2	0	2	1	0	1	0	2	1	1	3	3	0	0	1	0	1.0	1.0
Words in all syntactic breaks	0	5	0	5	6	0	4	1	0	2	0	4	1	3	6	7	0	0	1	0	2.3	2.5
Trail-offs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.1	0.2
All errors	2	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0.4	0.6
Lexical choice errors	1	1	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0.3	0.5
Non-standard lexical choice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0
Function word errors	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2
Tense/aspect errors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0
Morphology errors	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2
Phrase-level errors[3]	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2
Non-standard phrasing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0

Appendix B: Participant Proficiency Self-ratings

Table 17: English and Russian proficiency self-rating results, speaking = average of formal speaking and casual speaking ratings, verbal = average of formal speaking, casual speaking, formal listening and casual listening ratings.

Participant	English Overall	English Speaking	English Verbal		Participant	Russian Overall	Russian Speaking	Russian Verbal
1	4	4.5	4.5		1	5	4.5	4.75
2	5	5	5		2	5	5	5
3	4	3.5	3.5		3	5	4.5	4.75
4	4	3.5	4.25		4	5	4.5	4.75
5	4	4.5	4.5		5	5	5	5
6	4	4.5	4.75		6	5	5	5
7	4	4.5	4.5		7	5	5	5
9	4	3.5	3.75		9	5	5	5
10	4	4	4		10	5	5	5
11	4	3.5	3.5		11	5	5	5
12	4	4.5	4.75		12	5	5	5
13	4	4	4.25		13	4	4.5	4.5
14	4	5	5		14	5	5	5
15	4	4.5	4.75		15	5	5	5
16	3	3	3.5		16	4	4.5	4.75
17	4	4	4		17	5	5	5
19	4	5	4.75		19	5	5	5
20	4	4	4		20	5	5	5
21	3	3	3		21	5	5	5
22	4	3.5	3.75		22	4	3.5	3.75

References

Abrahamsson, N., Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249-306.

Barnwell, D. (1988). Some comments on T-Unit research. *System, 16*(2), 187-92.

Bedore, L. M., Pena, E. D., Gillam, R. B., Ho, T.H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498-510.

Bedore, L. M., Pena, E. D., Joyner, D., Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14*(5), 489-511.

Bedore, L. M., Pena, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition, 15*(3), 616-29.

Bedore, L. M., Leonard, L. B. (1998). Specific language impairment and grammatical morphology: a discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185-92.

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., Bates, E. (2005). A crosslinguistic study of the relationship between grammar and lexical development. *Journal of Child Language, 32*(4), 759-86.

Farhady, H. (1982). Measures of Language Proficiency from the Learner's Perspective. *TESOL Quarterly, 16*(1), 43-59.

Fiestas, C. E., Pena, E. D. (2004). Narrative discourse in bilingual children: language and task effects. *Language, Speech, and Hearing Services in Schools, 35*, 155-68.

Ginther, A. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*, 379-399.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., Cera, C. M. (2012).

Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition, 15*(3), 594-615.

Gutierrez-Clellen, V. F., Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24*(2), 267-88.

Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal, 79*(2), 223-34.

Kiran, S., Iakupova, R. (2011). Understanding the relationship between language proficiency, language impairment and rehabilitation: Evidence from a case study. *Clinical Linguistics and Phonetics, 25*(6-7), 565-83.

Kominsky, R. (1989). *How good is "how well"? An examination of the Census English-speaking ability question*. Paper resented at the annual meeting of the American Statistical Association, Washington, D.C.

MacSwan, J. (2001). *The non-non crisis: Knowledge of language and problems of construct validity in native language assessment*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle, WA.

MacSwan, J., Rolstad, K. (2006). How language proficiency tests mislead us about ability: Implications for English language learner placement in special education. *Teachers College Record, 108*(11), 2304-28.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Marian, V., Blumenfeld, H. K., Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940-67.

McMaster, K., Tilstra, J. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly, 29*(1), 43-53.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research and Practice, 21*(1), 30-43.

Nicholas, L. E., Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338-50.

Pray, L. (2005). How well do commonly used language instruments measure English oral-language proficiency? *Bilingual Research Journal, 29*(2), 387-409.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1398-1411.

Rice, M. L., Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*(6), 1239-57.

Selezneva, S. S. (2009). *Generative Naming Performance in Russian-English Bilingual Speakers: The Influence of Category and Language*. (Unpublished master's thesis). University of Texas, Austin, TX.

Scott, K. A., Roberts, J. A., Krakow, R. (2008). Oral and Written Language Development of Children Adopted From China. *American Journal of Speech-Language Pathology, 17(2)*, 150-60.

Solorio, T., Sherman, M., Liu, Y., Bedore, L. M., Pena, E. D., Iglesias, A. (2011). Analyzing language samples of Spanish–English bilingual children for the automated prediction of language dominance. *Natural Language Engineering, 17*(3), 367-95.

Townsend, C. E. (1975). Russian Word-Formation. Corrected Reprint. Cambridge, MA. Slavica Publishers, Inc.