LLT Journal, e-ISSN 2579-9533, p-ISSN 1410-7201, Vol. 24, No. 1, April 2021

LLT Journal: A Journal on Language and Language Teaching
http://e-journal.usd.ac.id/index.php/LLT
Sanata Dharma University, Yogyakarta, Indonesia

RATER AGREEMENT AND DISAGREEMENT IN THE MEASUREMENT OF ENGLISH ARTICLE ACQUISITION SUPPLIANCE AND ACCURACY

Rose Acen Upor
University of Dar es Salaam
upor@udsm.ac.tz
correspondence: upor@udsm.ac.tz
https://doi.org/10.24071/llt.v24i1.2603
received 19 May 2020; accepted 25 February 2021

Abstract
This study combines language assessment processes and interlanguage analysis techniques to determine rater agreement and disagreement in assessing English article acquisition. Employing native English speaking and non-native English speaking raters, picture sequence narratives that were written by English as a Foreign Language (EFL) learners (n=97) were coded and scored for suppliance-in-obligatory context (SOC) and target-like utterance (TLU). Although the kappa statistic revealed a fair agreement between raters (0.17 – 0.33), content analysis methods revealed much higher agreement (88.29% - 94.07%). Furthermore, language background effects between the raters could not be substantiated however the results demonstrated a discernable disagreement pattern between them. Thus, the study recommends the inclusion of a foreign language teaching background as a factor for rater selection to minimize language background effects on rating language assessments.

Keywords: Article acquisition, Inter-rater agreement, Inter-rater disagreement, Language background effects

Introduction
Although the general relationship between language assessment and second language acquisition is relatively well established, the association with foreign language learning situations such as in Africa has not been clearly understood. Despite, the wide acknowledgment of the multidimensional research in language assessment studies, appraisal of foreign language learning situations has not been fully explored. Most studies of inter-rater reliability (IRR) on language assessment focus on tests of English proficiency and issues of rater assessment. Some of the issues identified include rater bias, rater background, rater severity/leniency and formats of testing. Other aspects include methodology, rater sample, and rater agreement, to mention a few. In some studies, rater bias has been shown to impact the results of proficiency tests in particular rater language background and rater severity (Caban, 2003; Johnson & Lim, 2009; Kim, 2009). In other studies, possible effects of rater training on levels of inter-rater agreement

and rater severity were noted (Elder, Barkhuizen, Knoch, & von Randow, 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, Read, & von Randow, 2007; O'Sullivan & Rignall, 2007). Inter-rater reliability measures have also been used in studies that are not necessarily dependent on samples from language proficiency testing (Stolarova, Wolf, Rinker & Brielmann, 2014). This paper intends to explore and bridge foreign language learning research and language assessment methods through measurement of suppliance and accuracy in article acquisition as part of a methodology in inter-rater agreement. The aim of the study is two-fold; first, it addresses the inter-rater reliability measures of the ability of learners to supply articles and determine the accuracy of these forms, second it determines inter-rater agreement and disagreement effects on article suppliance.

In addressing the two aims of the study, this article is divided into 2 major sections. First, it builds on the existing body of research on the acquisition of English articles by adopting the Bickerton/Huebner model in determining the constructs for the rating scale (Bickerton, 1981; Huebner, 1983) and interlanguage analysis techniques in the collection of performance data (Pica, 1983). On one hand, the Bickerton/Huebner model is built on a taxonomy in the study of article use and it considers semantic and discourse-pragmatic features of the noun phrase (NP). According to the model, English NPs are classified based on referentiality i.e. specific reference [±SR] and hearer knowledge [±HK]. This allows for a comprehensive study of article use in four contexts namely, general reference (type 1), referential definite (type 2), indefinite reference (type3) and non-referential (type 4) (Bickerton, 1981; Huebner, 1983). This framework made it possible to differentiate the underlying uses of the English article system in narratives and set a rating scale. On the other hand, the interlanguage analysis techniques adopted from Pica (1983) intend to provide statistical support in determining the instances of suppliance and accuracy of article use by EFL participants in the study. The Suppliance-in-Obligatory Contexts (SOC) and Target-Like-Utterance (TLU) measures provide a basis for the raters to determine the obligatory contexts for suppliance and accuracy of the English articles. Norris and Ortega (1983) indicate that these measures reveal differential patterns in learner types that would have gone undetected. They claim that naturalistic learners and instruction-only learners tend to have a smaller expressive vocabulary than instruction-plus-exposure learners. This illustrates that these measures have an increased sensitivity of analytical units and procedures that may contribute to a better understanding within a given theory. Second, the study also builds on the constructs of rater assessment so as to determine rater agreement and disagreement. To do so, the study uses the assessment data from the raters to perform statistical tests to determine the rate of agreement and disagreement. Through the findings, the paper shall explore minimally two constructs of language assessment, namely, rater language background influence and rater bias. These constructs are associated with the analysis based on the non-native and native English speaking raters involvement in the study.

Hence, to expound on the relationship between language assessment and foreign language learning, and in particular, assessment of article suppliance and accuracy in narratives, the present study measured rater agreement and disagreement with a set of measures that span SLA and language assessment procedures. The findings of the study shall contribute to both the body of

knowledge in language assessment and foreign language learning by providing insight into open-ended language assessment and the role of foreign language teaching experience in rater criteria selection.

Acquisition of articles

It is a commonly discovered fact that EFL/ESL learners face difficulties in acquiring the English article system. Different reasons cited for these difficulties include the complexities of the English articles themselves (Celce-Murcia & Larsen-Freeman, 1999), the lack of an equivalent article system in the learner's native language (Mizuno, 2000) and a lack of effective teaching methods in English education (Yamada, 1982). Studies in the acquisition of English articles have approached from various viewpoints; the viewpoints of grammar (Yamada 1982; Lyons 1999), of usage (Dilin & Gleason, 2002), of context (Huebner, 1985; Parrish, 1987; Ionin, Ko & Wexler, 2004) and a typology of nouns preceding articles (Chierchia, 1998; Ogawa, 2008).

Evidence has shown that second language (L2) learners of English often have persistent difficulty in the use of articles until very late stages of acquisition or do not ever reach native-like levels of performance (Zdorenko & Paradis, 2008), even when there is increased time in instruction (Master, 1987; Ogawa, 2008). Some studies that have included comparisons of L2 learners from first language (L1) backgrounds with and without article systems suggest that L1 transfer most likely plays a role in the L2 learners' acquisition of English articles (Master, 1987; Murphy, 1997; Wakabayashi, 1997; Trademan, 2002; Hawkins, Al-Eid, Almahboob, Athanasopoulos, Chaengchenkit, Hu, Rezai, Jaensch, Jeon, Leung, Matsunaga, Ortega, Sarko, Snape, & Velasco-Zarate, 2006). Findings by Master (1987) indicate that there are variations that are considered in cases where L1s differ among subjects. However, the zero article (henceforth referred to as zero, Ø) dominates, which indicates that it is acquired first. Although the definite article, the, emerges early, there was evidence to indicate the-flooding in all environments. It is also noted that [-ART] learners delay in the acquisition of a when compared with the. With the acknowledgment of variation in learners from different L1 backgrounds, the argument in the case was whether there was a role played by the L1 transfer and whether the learners fluctuated in article parameter setting. Zdorenko and Paradis (2008) in their study of 17 ESL children discovered that the children substituted the definite article for the indefinite a in indefinite specific contexts regardless of the L1 background. Moreover, the children were more accurate in the use of the definite article in definite-specific contexts. The opposite was discovered by Jaensch (2008) who found that learners did not fluctuate between definiteness and specificity, although group comparisons proved that learners with higher proficiency outperformed learners with lower proficiency. Kaku (2006) brings forth an impelling perspective to article use. In his study of Japanese learner's use of the, he discovered that the definite article is associated referentiality and with Japanese being a [-ART] language, he noticed that learners were reassembling the newly acquired feature in relation with their current use of the Japanese demonstratives for specificity. In terms of using SOC and TLU measures, Lu (2001) investigated the accuracy rate and the order of acquisition and observed a different order of emergence of the articles the>a>zero. Differentiation of orders could be attributed to the instruction,

length of exposure, the participants themselves and/or the nature of the research tasks. Even where there were varied tasks performed by a group of learners, the results still yielded a systematic order of acquisition; however, the accuracy rate of the results was in question. The SOC measure is considered the most reliable index for accuracy levels (Lu, 2001).

Inter-rater reliability tests in article acquisition

Several studies have explored rater variability in both oral and written ESL performance assessment. Some of these studies focused on different rater backgrounds (Barnwell, 1989; Brown, 1995; Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Fayer & Krasinski, 1987; Galloway, 1980; Hadden, 1991), others studied rater severity (Barnwell, 1989; Caban, 2003; Fayer & Kransinski, 1987; Johnson & Lim, 2009; Kim, 2009), while others focused on rater decision-making strategies (Barkaoui, 2010; Crisp, 2008; Cumming, 1990; Cumming, Kantor, & Powers, 2002; Huot, 1993; Lumley, 2005; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Vaughan, 1991), and others on the interaction between rater and criteria (Knoch et al., 2007; McNamara, 1996; Schaefer, 2008; Wigglesworth, 1993). A common thread among all these studies was the use of standardized language performance assessment as the basis of their investigation. A study by Richard Nickalls at the University of Birmingham employed four raters in determining the inter-rater reliability testing of article error tags by checking the extent raters would reliably classify article use as 'correct' or 'incorrect' and if the correctness is consistently classified over time. The study used the Bickerton/Huebner Model and the raters received identical training. First, the raters tagged noun phrases for correctness using the online interface and three weeks later, the researchers tagged the same noun phrases again for correctness using the Bickerton/Heubner framework. The findings indicated that human raters were more reliable than automated computer methods. However, in terms of the Bickerton/Heubner framework, the findings showed that the raters could not use the framework consistently. Nickalls (2013) argues that raters cannot apply classification frameworks, in which the decision goes beyond a rater's dichotomous intuition especially in this case where they could not make reliable choices between generic, indefinite, non-referential and idiomatic contexts.

It also needs to be pointed out that rater background has been shown to impact the results of language proficiency in test-takers. Studies of raters with diverse backgrounds, both linguistic and professional have been conducted. Some studies focused on rater severity based on rater background (Brown, 1995; Chalhoub-Deville, 1995), others on raters' professional background (Hadden, 1991) and linguistic background (Fayer & Kransinski, 1987; Kim, 2009). Findings from these various studies indicate that teachers and non-native speakers tend to be more severe in their assessments (Brown, 1995; Chalhoub-Deville, 1995), teachers tend to be more severe than non-teachers (Hadden, 1991) and non-native raters tend to be more severe (Fayer & Kransinski, 1987). Discrepant findings from Chalhoub-Deville (1995) and Brown (1995) indicate that teachers who participated in their studies were attendant to creativity and adequacy of information in a narration task and, there was no significant difference between the rating done by NS and NNS, respectively. Johnson and Lim (2009) have

identified variables that could attribute to rater language background effects and intervene with the analysis when it comes to issues of NS and NNS raters. These issues included language distance affecting language performance (Elder & Davies, 1998); NS taking a more intuitive approach in rating (Brown, 1995), use of trained/untrained raters and different rating scales. These discrepancies call for further research into the area.

Method

Research questions

This present study will use data collected from Tanzanian EFL learners who were enrolled in 3different levels of education. The data were scored by 2 raters who possessed different language backgrounds. The study addressed the following research questions:

a. Is there variability in the suppliance and accuracy of the English article acquisition among the EFL learners?
b. To what extent will the raters agree in rating the article suppliance and accuracy?
c. Is there an identifiable pattern to rater disagreement?

If there is an identifiable pattern to rater disagreement, can an argument be made regarding the language background of the raters?

Participants

A total of 97 Tanzanian EFL learners participated in this study, 30 primary (elementary) school pupils (hereafter referred to as children), 30 secondary (high) school students (hereafter referred to as teenagers) and 19 students in their first year at University and 18 in their final year of university education. The elementary level students were enrolled in a public primary school in the outskirts of the city of Dar es Salaam. These are children who had at least 5 – 7 years of learning English as a subject, with all other subjects being taught in Swahili. The secondary school students were also enrolled in a public school; however, it is at this level of education that the medium of instruction shifts to all subjects being taught in English with Swahili as a subject. All university courses are taught in English with an exception for the Swahili language courses.

Table 1. Descriptive characteristics of the study sample

Characteristics	N	%
Participants		
Children	30	*30.9*
Teenagers	30	*30.9*
First year	19	*19.5*
Final Year	18	*18.5*
Gender		
Total	97	*100*
Male	50	*51.5*
Female	47	*48.5*
Mean Years of learning English		
Children	8.67	n.a.

Characteristics	N	%
Teenagers	9.14	n.a.
First Year	11.82	n.a.
Final Year	13.95	n.a.
Number of languages spoken		
Two	67	*69.1*
Three	27	*27.8*
Four +	3	*3.1*
First language		
Swahili	83	*85.6*
Other	14	*14.4*

The raters

The participants' narratives were scored by two raters. Both raters were trained in using SOC and TLU scoring methods. The rating scale was determined by the researchers following the Bickerton/Huebner model. Both raters were experienced instructors of English as a Foreign Language and had taught English to NNS through formal classroom instruction in environments where learners had limited language resources from which they could do language practice. Below is a profile of the raters:

Table 2. Descriptive Characteristics of the Raters

Characteristics	Rater 1	Rater 2
Language experience		
L1	Swahili	English
L2	English	Vietnamese
Other languages spoken	Luo and Jita (rudimentary)	Russian
English language proficiency		
	NNS	NS
	Native-like proficiency	Native speaker
Gender		
	Female	Female
Professional experience		
Teaching	21 years	26
Research	17 years	20

Methodology

Most studies on the acquisition articles have made use of language proficiency ascription for groups (Huebner, 1983; Jaensch, 2008; Kaku, 2006; Lu 2001; Ogawa, 2008; Tarone 1985; Zdorenko & Paradis 2008;); however, in this study levels of proficiency were not considered instead the groups were identified and ascribed based on the level of schooling. Due to distinct characteristics in the larger adult group (university students), this group was split into two smaller groups; first year students and seniors. All of the participants were asked to write out a narrative from a text with picture sequences (See Appendix A). Different picture sequences for data collection were used in the study, however, it should be noted that variation in narratives does not affect the results or findings of a study

(Ayoun & Salaberry, 2008). Each group of respondents was given different picture sequences for narration based on content, the number of years spent learning English and the difference in levels of education.

Rating scale and data analysis procedures

The picture sequences were designed to elicit narrative passages from the study participants. First, the researchers agreed on a protocol of their analysis before coding the data. They made use of suppliance-in-obligatory context (SOC) and target-like-use (TLU) measures. The first procedure, SOC is a method used to determine accurate suppliance of morphemes in linguistic environments in which the morphemes are required in Standard English. The basis for this analysis is that, if a participant produces an utterance such as 'I have few books', this speaker creates an obligatory context for use of the plural –s inflection. The reason behind this being that the participants appear to have acquired the rule of production of the morpheme, but have simply applied this rule to an exception (Pica 1983, Gass and Selinker 2001). This quantification method is represented in the following formula:

$$\text{SOC} = \frac{\text{number of correct suppliance} \times 2 + \text{number of misformations}}{\text{Total obligatory contexts} \times 2}$$

In the second procedure, TLU is used to determine accurate use and distributional patterns for morphemes. This analysis was developed in light of the criticism that SOC analysis does not account for the over suppliance of a particular morpheme in inappropriate contexts (Pica 1983, Gass and Selinker 2001). The method is represented as follows;

$$\text{TLU} = \frac{\text{number of correct suppliance in obligatory contexts}}{\text{Number of obligatory contexts} + \text{number of suppliance in nonobligatory contexts}}$$

Analysis by SOC reveals how well participants had learned to produce a morpheme where it is required while analysis by TLU reveals how well participants have learned to control the production of that morpheme about where it is and is not required (Pica 1983). The results from the SOC and TLU were computed into percentages. To determine the interactions between the factors as well as individual factors, statistical procedures were performed on the data. These methods of morpheme quantification were adopted to demonstrate the ability of EFL learners in using articles as they write narratives. The following definitions of constituents in the measures were as follows;

Correct suppliance: When the participants provide the correct form of the item in such a way that it does not make a construction ungrammatical
Obligatory context: When the participants create a context of the use of an item in such a way that without it the construction is deemed ungrammatical and with it, the construction is deemed grammatical
Misformation: When the participants provide an incorrect item in the context of a correct item in such a way that it deems the construction ungrammatical

Non-Obligatory Context: When the participants provide an item in a context in which it was not required or not created for its inclusion

After the defining constituents in the SOC and TLU, a rating scale was established for articles based on the types of forms and their functions in Standard English. The rating scale is as follows:

Rating for Articles
Step 1: General or Specific to Specific
Does the narrative make use of articles in a general way?
If yes → the beginning of the narrative will use 'a/an' and then move towards specific 'the'.
If no → the narrative will maintain the specific form 'the' from start to end, using the narratives to provide prior context for a specific reference.

Step 2: Naming
Do any of the narratives use the naming of characters?
If yes → No article should appear before the noun form referring to the characters, which should be capitalized.
If no → refer back to step 1.

The scale was to be used as the researchers identified the SOC and TLU scores of the narratives. The analysis was conducted as follows: 1) the researchers independently reviewed and coded the written narratives to identify articles produced in each context as either correct suppliance, misformation, non-obligatory context, and obligatory context, and; 2) the scores that the researchers awarded the SOC and the TLU were then entered into SPSS for further analysis

Findings and Discussion

Suppliance and accuracy of articles

A one-way analysis of variance (ANOVA) was conducted on the scores of the groups' SOC and TLU to evaluate the relationship between the ability to supply the forms in the study and the accuracy of this suppliance within the different groups. A statistically significant difference was found among the four levels of EFL learner groups on the average SOC for articles ($F(3, 93) = 18.80$, $p = .000$) and on the average TLU for articles ($F(3, 93) = 15.72$, $p = .000$).

Table 3. ANOVA Table for the SOC and TLU for Articles

	Items	*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
SOC	Between Groups	16371.643	3	5457.214	18.798	.000*
	Within Groups	26998.401	93	290.305		
	Total	43370.044	96			
TLU	Between Groups	17888.655	3	5962.885	15.719	.000*
	Within Groups	35277.842	93	379.332		
	Total	53166.497	96			

Due to the number of groups, a posthoc test was performed to uncover specific differences between the group means using the average SOC and TLU scores. The Games Howell test reveals that the four groups differed significantly in their ability in their suppliance and accuracy of articles. There was a significant difference in the suppliance of articles between the children group (p = <0.5) and the teenage group however there was no significant difference between the children group and the adult groupings. This limited variability between the children and adult groupings could be attributed to the length of the narratives and the number of correct formations. Although the children's narratives were shorter, the magnitude of correct formations, misformation, and obligatory contexts was much similar to the adult groupings. Likewise, there were also significant differences between 1st-year students, teenagers, and final year students. In the accuracy of the articles, the test results indicated that the only group that was statistically significant from the rest of the groups was the teenage group (p=<0.5). This significance is important because it was within this group that both raters experienced very short narratives, high instances of naming and inconsistent use of capitalization compared to the other groups, therefore, proving a challenge to the raters. Furthermore, it is the same group that was consistently outperformed by the other groups in terms of both suppliance and target-like use of articles. The other group that has also shown to be significantly different based on this test is the final year adult group (p=<0.5). This group has illustrated a significant difference from the other groups in terms of the average identifying of contexts of use of articles. Table 4 illustrates the results of the Games-Howell tests on the groups' average TLU and SOC.

Table 4. Games-Howell Test of the Average SOC and TLU of Articles

Dependent Variable	*(I) Age Groups*	*(J) Age Groups*	*Mean Difference (I-J)*	*Std. Error*	*Sig.*	*95% Confidence Interval*	
						Lower Bound	Upper Bound
Average SOC for Articles	Children	Teens	27.29166*	4.99206	.000*	13.9603	40.6230
		1st Year	7.35311	4.31444	.338	-4.3316	19.0378
		Final Year	-5.63480	2.81565	.203	-13.1424	1.8728
	Teens	Children	-27.29166*	4.99206	.000*	-40.6230	-13.9603
		1st Year	-19.93855*	5.72807	.006*	-35.1957	-4.6814
		Final Year	-32.92646*	4.70365	.000*	-45.5936	-20.2593
	1st Year	Children	-7.35311	4.31444	.338	-19.0378	4.3316
		Teens	19.93855*	5.72807	.006*	4.6814	35.1957
		Final Year	-12.98791*	3.97718	.016*	-23.9380	-2.0378
	Final Year	Children	5.63480	2.81565	.203	-1.8728	13.1424

Dependent Variable	*(I) Age Groups*	*(J) Age Groups*	*Mean Difference (I-J)*	*Std. Error*	*Sig.*	*95% Confidence Interval*	
						Lower Bound	Upper Bound
		Teens	32.92646*	4.70365	.000*	20.2593	45.5936
		1st Year	12.98791*	3.97718	.016*	2.0378	23.9380
Average TLU for Articles	Children	Teens	24.81160*	5.43016	.000*	10.4023	39.2209
		1st Year	6.52689	5.45956	.634	-8.1981	21.2518
		Final Year	-12.59477*	4.36680	.030*	-24.2578	-.9317
	Teens	Children	-24.81160*	5.43016	.000*	-39.2209	-10.4023
		1st Year	-18.28471*	6.28700	.028*	-35.0673	-1.5022
		4th Year	-37.40637*	5.36548	.000*	-51.7139	-23.0989
	1st Year	Children	-6.52689	5.45956	.634	-21.2518	8.1981
		Teens	18.28471*	6.28700	.028*	1.5022	35.0673
		Final Year	-19.12166*	5.39524	.007*	-33.7553	-4.4880
	Final Year s	Children	12.59477*	4.36680	.030*	.9317	24.2578
		Teens	37.40637*	5.36548	.000*	23.0989	51.7139
		1st Year	19.12166*	5.39524	.007*	4.4880	33.7553

* The mean difference is significant at the .05 level.

Inter-rater Agreement

Three separate tests were involved in determining the rate of agreement and disagreement between the two raters i.e. Cohen's kappa, Holsti's content analysis, and Scott's pi. Cohen's kappa statistic is frequently used to measure the agreement between two raters. The cross-tabulation between the rating of suppliance and accuracy of articles shows that there is an agreement between the two raters. The symmetric measures table shows that Kappa for each level of rating between the raters indicates fair agreement for correct formations (.29), misformations (.30) and non-obligatory contexts (.33) and slight agreement (0.17) for obligatory contexts as shown in Table 6.

Table 6: Symmetric Measures of Cohen's Kappa between the two raters

Item	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Correct formations	.293	.048	15.067	.000*
Misformations	.300	.061	6.403	.000*
Obligatory contexts	.170	.041	9.503	.000*
Non-obligatory contexts	.330	.086	4.320	.000*
N of Valid cases	97			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

These results indicate a large amount of disagreement than expected between the raters. In as much as the kappa is used to measure inter-rater agreement, its strength lies in the fact a study has collected correct representations of the variables measured (McHugh, 2012). A probable explanation for this low agreement could be a symmetrical imbalance between the two raters. However, the kappa statistic is also known to have its limitations. The terms symmetrical, asymmetrical, imbalance, prevalence, and bias have been used to describe the limitations associated with the statistic (Flight & Julious, 2015). The most probable explanation for low kappa in the context of the study would be the problem of oversuppliance errors as predicted by Pica (1983) which point towards prevalence in this case. Moreover, Feinsten and Cicchetti (1990) highlight what they refer to as 'paradoxes' of the kappa. They indicated that asymmetric, imperfectly imbalanced tables have higher kappa than perfectly imbalanced symmetric tables. Also where there were high values of agreement, lower values of kappa were recorded. Based on this observation, we could predict that because of the low kappa recorded, probable high values shall be recorded in through other indices. Most of the studies that have recorded limitations in the kappa statistic are health-related studies (Flight & Julious, 2015; McHugh, 2012; Tang, Hu, Zhang, Wu, & He, 2015).

Although a Prevalence and Bias Adjusted Kappa (PABAK) is proposed to overcome the limitations of the kappa statistic (Byrt, 1993), this study chose to use the content analysis method proposed by Holsti (1969). The two-stage process was chosen: first, to determine the degree of token-based agreement among the raters and second, to determine the degree of agreement through traditional inferential statistics. The first part of the analysis contains a count of the tokens of articles between the two raters for the participants and use Holsti's method (1969) for determining the agreement. The method is a variation of percentage agreement, a measure that is popular and easy to understand and calculate, yet it can be applied to more than two coders (Lombard et al., 2002), unlike for Holsti's method that is limited to two coders as evidenced in its formula.

$$\textbf{Coefficient of Reliability} = \frac{2M}{N1 + N2}$$

M is the number of judgments on which both of the coders agree
N1 and N2 are the total number of judgments made by both coders

Source: Holsti, O. R. (1969). Content analysis for the social sciences and humanities, pp140

Table 7 presents the description of the results of the narratives, showing total use (number of tokens) and percentage usage by the group and by the rater. Table 6 is followed by Table 8 that summarizes the information from Table 7.

Table 7. Step by Step descriptives and coefficients of reliability by group and rater

Group	Rating items		Rater1	Rater2	Agreement	2M	N1 + N2	C.R. (%)
Children	Suppliance-in-obligatory context	*Corr*	197	215	194	388	412	94
		Mis	25	35	18	36	60	60
		Oblig	229	264	223	446	493	90
		Total	*451*	*514*	*435*	*870*	*965*	*90*
	Target-like use	*Corr*	197	215	194	388	412	94
		Oblig	229	264	223	446	493	90
		Non	11	17	9	18	28	64
		Total	*437*	*496*	*426*	*852*	*933*	*91*
Teens	Suppliance-in-obligatory context	*Corr*	245	270	236	472	515	92
		Mis	80	56	50	100	136	74
		Oblig	500	464	426	852	964	88
		Total	*825*	*790*	*712*	*1424*	*1615*	*88*
	Target-like use	*Corr*	245	270	236	472	515	92
		Oblig	500	464	426	852	964	88
		Non	9	17	2	4	26	15
		Total	*754*	*751*	*664*	*1328*	*1505*	*88*
First Year Students	Suppliance-in-obligatory context	*Corr*	392	415	378	756	807	94
		Mis	64	66	45	90	130	69
		Oblig	517	532	492	984	1049	94
		Total	*973*	*1013*	*915*	*1830*	*1986*	*92*
	Target-like use	*Corr*	392	415	378	756	807	94
		Oblig	517	532	492	984	1049	94
		Non	17	22	2	4	39	10
		Total	*926*	*969*	*872*	*1744*	*1895*	*92*
Final Year Students	Suppliance-in-obligatory context	*Corr*	607	643	602	1204	1250	96
		Mis	36	39	31	62	75	83
		Oblig	653	715	652	1304	1368	95
		Total	1296	1397	1285	2570	2693	95
	Target-like use	*Corr*	607	643	602	1204	1250	96
		Oblig	653	715	652	1304	1368	95
		Non	0	5	0	0	5	0
		Total	*1260*	*1363*	*1254*	*2508*	*2623*	*96%*

Table 8. Summary of Descriptives and Coefficients of Reliability

	Suppliance in Obligatory Context (SOC)				Target Like Utterance (TLU)			
	Correct	Mis	Oblig	Total	Correct	Oblig	Non	Total
Rater 1	1441	205	1899	3545	1441	1899	37	3377
Rater 2	1543	196	1937	3676	1543	1937	61	3541
Agreement	*1410*	*144*	*1831*	*3385*	*1410*	*1831*	*13*	*3254*
2M	2820	288	3662	6770	2820	3662	26	6508
N1 + N2	2984	401	3836	7221	2984	3836	98	6918
C.R. *(%)*	94.50	71.82	95.46	93.75	94.50	95.46	26.53	94.07

KEY:

N1 Count of instances by rater 1
N2 Count of instances by rater 2
2M Expected total IFF the raters agreed on all instances/twice the agreement count
C.R Coefficient of Reliability

In summation, the coefficients used to calculate inter-rater reliability were reported in most of the articles (94.07%, n=97). Rater agreement in the suppliance of articles in obligatory contexts and target-like use in obligatory contexts was reported at 95.46% as the most frequent coefficient. The area of disagreement between the researchers was the use of articles in non-obligatory contexts (26.53%) whereas there was a satisfactory agreement when it came to misformations. Overall, both raters agreed 2820 times out of 2984. A major drawback of Holsti's method reported is the lack of ability to calculate the agreement by chance (Wang, 2011). Due to this weakness, we adopted a third index, Scott's pi (π), which not only improves on simple percent agreement but also takes into consideration category values and accounts for chance agreement (Wang, 2011). Scott's pi (π) was used to determine inter-rater reliability and its results were used to check rater bias and language background effects.

Inter-rater reliability and language background effects

The coding for the reliability sample included identification of all instances of correct suppliances, misformations, obligatory contexts and non-obligatory contexts in all 97 narratives. In as much as the raters worked independently in coding the samples, the researchers used Scott's pi (π) for verification of the reliability and inter-rater agreement. The equation for Scott's pi is:

$$\pi = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where: Pr(a) = observed agreement between coders
Pr (e) = expected agreement between the coders

To obtain coefficients of reliability for Scott's pi scores, the raters compared each instance of agreement in each narrative for articles SOC and TLU categories. The results indicated consistency in inter-rater reliability. However, it was anticipated that issues would arise from the teen group since it was the only group that had a completely different perspective towards the narrative exercise. This group chose to name the characters rather than objectifying them as they would

have appeared in the text. This necessitated revision of the rating scale to include naming since there were significant differences in how the raters chose to address the issue. The coefficient of reliability for all cases was 88.52% (Articles SOC) and 88.29% (Articles TLU). Table 9 illustrates the inter-rater scores using Scott's pi (π).

Table 9. Scott's pi (π) Inter-rater Reliability

Items	*Children (%)*	*Teenagers (%)*	*First Year (%)*	*Fourth Year (%)*	*Overall (%)*
Articles SOC	82.19	77.89	85.76	91.31	88.52
Articles TLU	83.46	75.10	84.43	91.25	88.29

Apart from reaching the inter-rater reliability for raters, the need for determining patterns of disagreement was important with regards to the rater profile, i.e. NS and NNS. Out of 97 participants, it was noted that Swahili was the L1 for 83 participants and L2 for 14 participants, English was L2 and L3 respectively. Rater 1's L1 is Swahili and it may be inferred from the research on [-ART] languages as to background effects on their rating unlike for rater 2, whose was L1 was English. Bias terms were measured for each of the raters despite the absence of an English L1 participant. The bias terms followed the SOC and TLU scores of each rater per participant where a total of 86 participant scores fell within the Z score range of -1.96 and +1.96 using a 95% confidence level. Only 11 participants' scores fell out of range. This indicates that disagreement effects were not significant as expected because the magnitude of bias was not substantive and both raters contributed to the bias. Where bias was exhibited, it was discovered that most of the cases were found in one particular group of participants. Table 10 illustrates the bias terms by participant.

Table 10. Bias terms by participant

SOC ≤ -1.96		TLU ≤ -1.96		Participant #	SOC ≥ 1.96	TLU ≥1.96
			R2	32		
R1			R2	35*		
R1			R2	43*		
R1	R2	R1	R2	49*		
R1	R2		R2	51*		
			R2	56*		
R1	R2	R1		57*		
R1	R2	R1		58*		
R1				66*		
R1		R1	R2	67*		
	R2		R2	72		

Key: * teenage group

Table 10 indicates that rater 1, as an NS of Swahili, was biased when participants supplied articles in the obligatory contexts (production) than rater 2 who was more inclined towards the accuracy of the use of the articles (performance) by the participants. Using the notion of the directionality of severity even though bias, in this case, does not entail severity (Johnson & Lim, 2009), it is noted that both raters' biases were negative numbers and were

clustered between -3.656 and -1.988. Although Johnson and Lim (2009) made use of a different analysis index from the one adopted in this study, their analysis claimed that positive numbers indicate harshness and negative numbers indicate leniency. This could be loosely interpreted that the raters had a similar inclination towards leniency and were consistent in their observations of the data. This observation supports the findings of the study by Kim (2009) that indicates the NS and NNS raters showing consistency. However, the results do not support studies (Chalhoub-Deville, 1995; Brown, 1995; Chalhoub-Deville & Wigglesworth, 2005; Hadden, 1991) that noted significant differences in how NS and NNS raters behave, and with NNS and teachers being more severe in their assessments. One possible cause for the consistency found in this study could be the experience that both raters had with foreign language teaching. Moreover, the issue of NS and NNS is fluid in this study because there is a rater who happens to be an NS of an L1 that is shared by over 85.5% of the study participants as well as being an NNS with near-native fluency to the language of study. This raises the question of the application of intuitive knowledge by the raters. Despite the use of a rating scale, suppliance, and accuracy judgments, it became evident that some judgments were also made based on each rater's intuition and perception of student intent. Inconsistent student use of capitalization, inconsistent use of the definite article, and spelling mistakes further complicated the rating process. Although Scott's pi places the inter-rater reliability at an average of 88.52% (SOC) and 88.29% (TLU), subjective impressions from initial agreement analyses revealed that there may be patterns to the non-agreement (11.59%), with misformation and non-obligatory context as frequent areas of non-agreement. Despite the perception of systematic non-agreement between raters, the disagreement was not statistically significant. Disagreement occurred primarily in narratives that used capitalization variably, which was perceived by one rater as naming (no article required), but by the other as misformation. Because of this limited effect, we believe that rater language background effects were not significant.

Conclusion

This study was guided by three research questions; i) is there variability in the suppliance and accuracy of the English article acquisition among the EFL learners?; ii) to what extent will the raters agree in rating the article suppliance and accuracy? and; iii) is there an identifiable pattern to rater disagreement? If there is an identifiable pattern to rater disagreement, can an argument be made regarding the language background of the raters?

Regarding the performance of the learners on the narrative task, variability was found to be significant among the four groups that participated in the study. Further analysis revealed that the results on the suppliance and accuracy of articles confirm that native-like performance for the more advanced participants has not been reached despite the increased time of instruction compared to other participants of the study (Zdorenko & Paradis, 2008; Masters, 1987; Ogawa, 2008). Even though for 11 out of 18 of the advanced participants English was an L3, there is no indication of any substantial effect on the overall results. Higher proficiency in article suppliance and accuracy was found in the advanced participants which support findings by Jaensch (2008) and can be attributed to the increased time of instruction (mean years of learning = 13.95). A methodological

choice was made to leave the Ø article out of the analysis and focus on the definite and indefinite articles according to the specifications of the rating scale. The issue of the-flooding was not an area of focus and where it occurred it was considered as a misformation. Evidence of fluctuation can be implied by the performance of the teens' group (mean years of learning = 9.14 years). Also, the findings are indicative of U-shaped learning and it can be assumed that the learners are at the stage of parameter setting (Zdorenko & Paradis, 2008). This particular group also exhibited the use of the distal demonstrative 'that' to substitute the referential function of the definite article. Similar sentiments are expressed by Kaku (2006) who found Japanese learners of English using demonstratives for specificity. In terms of the learner performance and the coding decisions between the raters, consistency in articles was relative and when it occurred, it was seemingly governed by the learners' perception of the semantic function of the characters in the narratives and character-character interaction. In regards to how well the four groups of English language learners used articles, the study revealed there was a significant difference between the four groups in SOC and TLU measures. Follow-up discussion of the perception of student intent and exploration of disagreement between the raters discovered that there were systematic shifts in anaphoric use of articles in the narratives. This could be explained as an L1 effect in the learners. \

With regards to the preceding research questions on rater agreement, the researchers used inter-rater reliability and inter-rater agreement measures in what may be considered traditional SLA tests of learner ability to produce articles by measuring SOC and TLU scores. In using these tests, we find that it is constructive and it bridges language testing methods to SLA research. Through the combination of SOC and TLU measures, inter-rater agreement and inter-rater reliability and SOC and TLU methods employed, the findings of the study have revealed through two inter-rater agreement indices that there is a very high level of agreement whereas in one index there seems to be fair to slight level of agreement. Feinsten and Cicchetti (1990) confirm that there is a tendency of a low kappa statistic recorded with high agreement levels as we have found in this study. It is important to note that the study did not make use of final scores of the narratives as would in most IRR studies but rather the scores of the raters' judgments of production and accuracy of English articles as interpreted in the narratives. This method contributes to the body of knowledge on rater agreement studies in that teasing apart the aspects of measurements may provide insight into levels of agreement. Furthermore, the analysis indicates that the language background of the raters does not influence agreement between them. The evidence of support is found in the bias terms as indicated in Table 10 which indicates consistency between the raters. It further signifies that the raters shared challenges in rating the same narratives of the participants. Additionally, it points out that experience in foreign language teaching had a role to play in how the raters viewed these same narratives even more so the language proficiency of the NNS rater. The study has proven that where studies involving NNS with above intermediate proficiency, the likelihood for them to rate at almost the same level of the NS is very high. Johnson and Lim (2009) hypothesize that NNS raters could rate performance assessments differently because they possess a language background from places with well-developed varieties of English thus causing

them to overlook or accept features that are unacceptable in a standard dialect. This has not been the case in this study. Still, the major question also lies in how much of the rater's intuitive knowledge of the language matter is being used, which cannot be measured or observed as part of the rating scale that has been agreed upon. A major conclusion of this study is that training of the rating scale and probably the experience of the raters minimizes the language background effects and other possible biases. However, it does not eliminate the possibility of rater focus on particular areas of rating that emanate from their intuitive knowledge and use of the language of assessment.

This study acknowledges and addresses some methodological limitations faced in the analysis processes. First, the study employed labor-intensive procedures in the coding and analysis of the data. This intensity is evident in the rating scale, SOC and TLU measures, narrative method and the Holsti method. The SOC and TLU measures are not common methods in the collection of data for IRR studies but through this study, it has proven to be a means through which individuality and freedom of rater judgments can be achieved. Second and closely related to the first limitation is the design of the rating scale. The rating scale not only allows for individuality and freedom of the rater judgments but it can also allow for intuitive methods that rely mostly on the interpretation of the raters about the learner narratives. The Holsti method allowed the raters to revisit each instance they coded painstakingly and determine the level of agreement and disagreement. Both raters, however, had previous experience of using the SOC and TLU measures, therefore, limiting the training time of the adopted scale in the study. Third, the number of raters involved in the study does not strongly provide a basis for rater language background influence argument in comparison to most studies on rater language background effects. The study had only two raters of varying English language background, as a result, it only amplifies issues that could arise from rating systems of language tests that may have not been standardized; consider the SOC and TLU measures as well as the use of narratives. Methodological choices of this nature may sometimes permit unreliable conclusions where rating lacks a systematic procedure and as a result, it inadequately expresses the proficiency of a learner but it can also provide grounds for developing systematic procedures for analyzing learner compositions. Based on these three limitations, it is prudent to argue that generalizability of the results would require some amount of caution.

In conclusion, this study suggests that the kappa coefficient may not be sufficient in expressing inter-rater agreement as also indicated in other studies (Flight & Julious, 2015; McHugh, 2012, Tang, et.al. 2015). It proposes the use of other indices that may support the results acquired through Cohen's kappa. Evidence from the study also supports that training in the rating scale rubric (Johnson & Lim, 2009) is an important factor in the scoring of the assessments, however, the study also emphasizes the importance of the experience of the raters in foreign language teaching as an important factor in minimizing language background effects in cases where NS and NNS raters are used. Due to this observation, the study could not provide a concrete argument as there being any language background effects in the assessment of the narratives.

References

Barnwell, D. (1989). 'Naïve' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*, 152–163.

Bickerton, D. (1981) *Roots of language*. Ann Arbor, MI: Karoma.

Brown, A. (1995). The effect of rater variables in the development of an occupation- specific language performance test. *Language Testing*, *12*, 1–15.

Byrt, T., Bishop, J. & Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Epidemiology*, *46*(5): 423-429

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*, 1–44.

Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd Ed.), Boston: Heinle & Heinle Publishers

Chalhoub-Deville, M. & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*, 383–391.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*, 16–33.

Chierchia, G. (1998). Plurality of mass nouns and the notion of 'semantic parameter'. In S. Rothstein (Ed.), *Events and Grammar* (pp 53-103). Kluwer: Dordrecht.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, *38*, 247–264.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*, 31–51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*, 67–96.

Dilin, L. & Gleason, J.L. (2002). Acquisition of the article the by non-native speakers of English: An analysis of four non-generic uses, *Studies in Second Language Acquisition*, *24*(1), 1-26.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*, 37–64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175–196.

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning*, *37*, 313–326.

Feinsten, A. R. & Chicchetti, D.V. (1990). High agreement but low kappa: The problems of two paradoxes, *Journal of Clinical Epidemiology*, *43*, 543-548.

Flight, L., & Julious, S. A. (2015). The disagreeable behavior of the kappa statistic. *Pharmaceutical Statistics*, *14*(1), 74-78. https://doi.org/10.1002/pst.1659

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, *64*, 428–433.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, *41*, 1–24.

Hawkins, R., Al-Eid, S., Almahboob, I., Athanasopoulos, P., Chaengchenkit, R., Hu, J., Rezai, M., Jaensch, C., Jeon, Y., Leung, Y-K.I., Matsunaga, K., Ortega, M., Sarko, G., Snape, N. & Velasco-Zarate, K. (2006) Accounting for English article interpretation by L2 speakers. In Foster-Cohen, S.H., Medved Krajnovic, M. and Mihaljevic Djigunovic, J. (eds) *EUROSLA Yearbook, Volume* 6, 7-25

Holsti, O. R. (1969). Content analysis for the social sciences and humanities, reading. MA: Addison-Wesley.

Huebner, T. (1985). System and variability in interlanguage syntax. *Language Learning*, *35*, 141-163

Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor, MI: Karoma.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Ionin, T., Ko, H. & Wexler, K. (2004) Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, *12*(1), 3-69

Jaensch, C. (2008). L3 acquisition of articles in German by native Japanese speakers. In *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007). Somerville, MA: Cascadilla Proceedings Project* (Vol. 8189, No. 2009, p. L3).

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*, 485–505.

Kaku, K. (2006). Second language learners' use of English articles: A case of native speakers of Japanese. *Cahiers Linguistiques d'Ottawa/Ottawa Papers in Linguistics*, *34*, 63-74.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing Writing*, *12*, 26–43.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

Lu, C.F-C. (2001) The acquisition of English articles by Chinese learners, *Second Language Studies*, *20*, 43-78.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Lang.

Lyons, C. (1999). *Definiteness.* Cambridge: Cambridge University Press.

Master, P. A. (1987). *A cross-linguistic interlanguage analysis of the acquisition of the English article system* (Doctoral dissertation, UCLA).

McHugh, M. L. (2012). Interrater reliability: The kappa statistic, *Biochem Med (Zagreb)*, *22*(3), 276-282.

McNamara, T. (1996). *Measuring second language performance*. New York, NY: Addison Wesley Longman Limited.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In *M. Milanovic & N. Saville (Eds.), Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 92–114). Cambridge, UK: Cambridge University Press.

Murphy, S. (1997) *Knowledge and production of English articles by advanced second language learners*, Unpublished doctoral dissertation, University of Texas at Austin.

Nickalls, R. (2013). *Inter-rater reliability testing of article error tags: an argument for framework simplicity.* Poster session presented at the Learner Corpus Research Conference, Bergen, Norway, Retrieved from https://lcr2013.w.uib.no/files/2013/09/Nickalls-poster.pdf

Norris, J. & Ortega, L. (2003). Defining and Measuring SLA. In *C. J. Doughty & M.H. Long (Eds.) The Handbook of Second Language Acquisition* (pp 717 – 760). https://doi.org/10.1002/9780470756492.ch21

Ogawa, M. (2008) The acquisition of English articles by advanced EFL Japanese learners: Analysis based on noun types, *Journal of Language and Culture Language and Information 3*, 133-151

Parrish, B. (1987) A new look at methodologies in the study of article acquisition for learners of ESL, *Language Learning 37*, 361-83

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*(1), 69-78.

Sakyi, A. A. (2000, October). Validation of holistic scoring for ESL writing assessment: How raters evaluate. In *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9, p. 129). Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465–493.

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in psychology*, *5*, 509.

Tang, W., Hu, J., Zhang, H., Wu, P., & He, H. (2015). Kappa coefficient: A popular measure of rater agreement. *Shanghai Archives of Psychiatry*, *27*(1), 62-67. https://dx.doi.org/10.11919%2Fj.issn.1002-0829.215010

Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax, *Language Learning, 35*, 373-404

Trademan, J. (2002). *The acquisition of English article system by native speakers of Spanish and Japanese: a cross-linguistic comparison* (Unpublished PhD dissertation, University of New Mexico).

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In *L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Wakabayashi, S. (1997). *The acquisition of functional categories by learners of English* (Unpublished doctoral dissertation, University of Cambridge).

Wang, W. (2011). A content analysis of reliability in advertising content analysis studies. Electronic Theses and Dissertations, p.1375. http://dc.etsu.edu/etd/1375

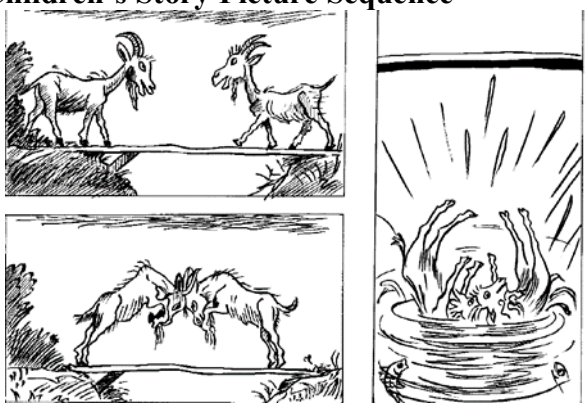
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305–335.

Yamada, J. (1982). The use of the English articles among Japanese students. *RELC Journal*, *13*(1), 50-63.

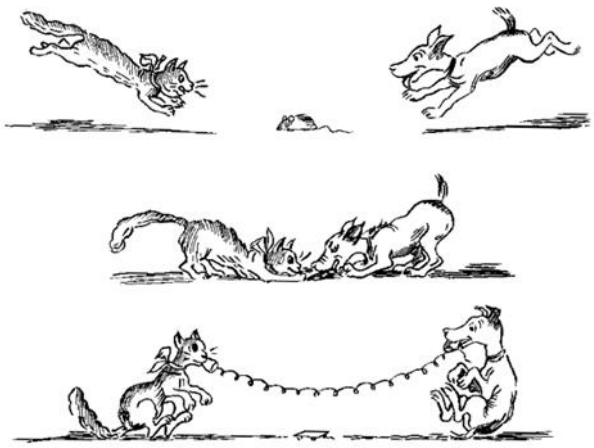
Zdorenko, T. & Paradis, J. (2008). The acquisition of articles in child second language English: fluctuation, transfer or both?, *Second Language Research*, *24*(2), 227-250.

Appendix A

A. Children's Story Picture Sequence



B. Teenager's Story Picture Sequence



C. Adults Story Picture Sequence