A Comparison of Methods for Measuring Writing Accuracy

Aimee Clayton, Jenna Smith, Joclyn Farrales, Lidija Vasica, Merrie Kay Ames, Marylin Oliphant, Nathan Burgess, and Sarah Freeman

Department of Linguistics, Brigham Young University

LING 620: Research in TESOL

Dr. James Hartshorn

December 20, 2023

Research Questions

1) Is the General Accuracy Ratio (GAR) method and the Error-Free Clause Ratio (EFCR) equally practical?
2) Regarding validity, do the respective methods measure what they are intended to measure?
    a) Do the respective methods measure accuracy equivalently?
    b) Can one method be a valid substitute for the other?
    c) Is one method better than the other?
    d) Are there benefits and drawbacks to both methods and under what circumstances?

Methods

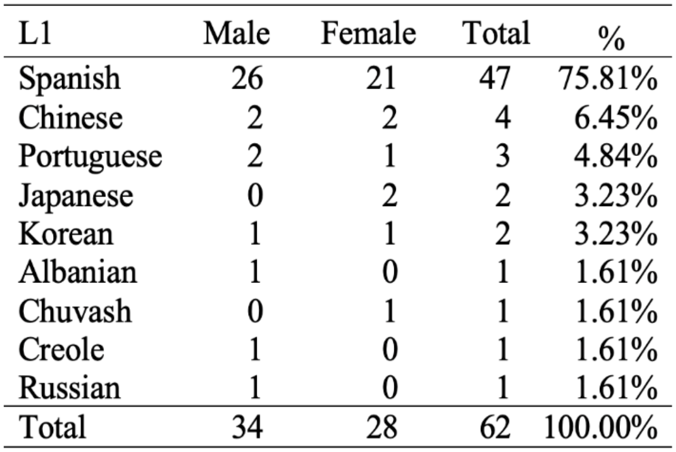
Context

Participants included learners of English and raters who rated the student samples. The samples gathered for rating were taken from 62 students of the English Language Center (ELC) at Brigham Young University (BYU). The criterion for inclusion in the study was students who attended the ELC over the course of 45 consecutive weeks (three semesters). Table 1 displays the students' language backgrounds, sex, and age range.

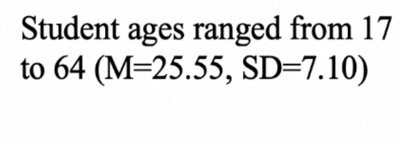
The ELC is a lab school and all students sign a form that grants access to their data for research purposes. Thus, the data on student demographics from the ELC was available and no recruitment methods were needed. The English proficiency levels of the students ranged from Novice-High to Intermediate-High, with the majority of students being at the Intermediate-Mid level as shown in Table 2.

Table 1

Participant Language Background



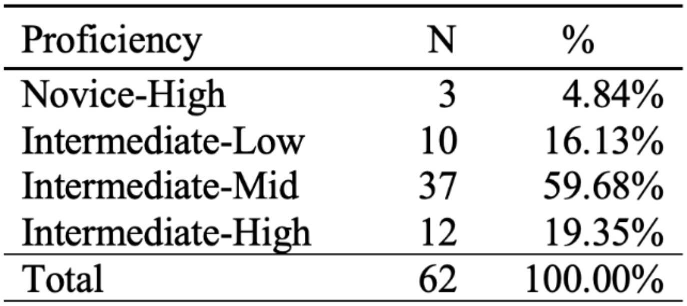
L1	Male	Female	Total	%
Spanish	26	21	47	75.81%
Chinese	2	2	4	6.45%
Portuguese	2	1	3	4.84%
Japanese	0	2	2	3.23%
Korean	1	1	2	3.23%
Albanian	1	0	1	1.61%
Chuvash	0	1	1	1.61%
Creole	1	0	1	1.61%
Russian	1	0	1	1.61%
Total	34	28	62	100.00%



Student ages ranged from 17 to 64 (M=25.55, SD=7.10)

Table 2

Participant Proficiency Level



Proficiency	N	%
Novice-High	3	4.84%
Intermediate-Low	10	16.13%
Intermediate-Mid	37	59.68%
Intermediate-High	12	19.35%
Total	62	100.00%

Data Collection and Instruments

The ELC administers a placement exam and a proficiency exam (Language Acquisition Test or LAT). These two tests are equivalent in form and have been shown to be valid and reliable. The pretest samples were taken from the students' initial placement test for the ELC. The posttest was taken from a portion of the students' LAT, which was administered at the end of the three semesters.

The portion of the exams used for this research were 30-minute and 10-minute essays for both the pretest and the posttest (see Appendix A). They were administered in a computer laboratory at the ELC. Proprietary software was used, which does not allow for any editing except deleting, cutting, and pasting text. The computer program timed the essays such that the screen shut down when the time expired.

Raters

The participant raters for this study were first and second-year TESOL graduate students who completed the ratings as part of their coursework in a research class. We had a group of eight raters ranging in age from 23 to 59. Seven of the raters were female with one male rater, and only one had an L1 other than English (Serbian). At the time of rating, most of these students were also practicum teachers at the same institution where the writing samples had been collected.

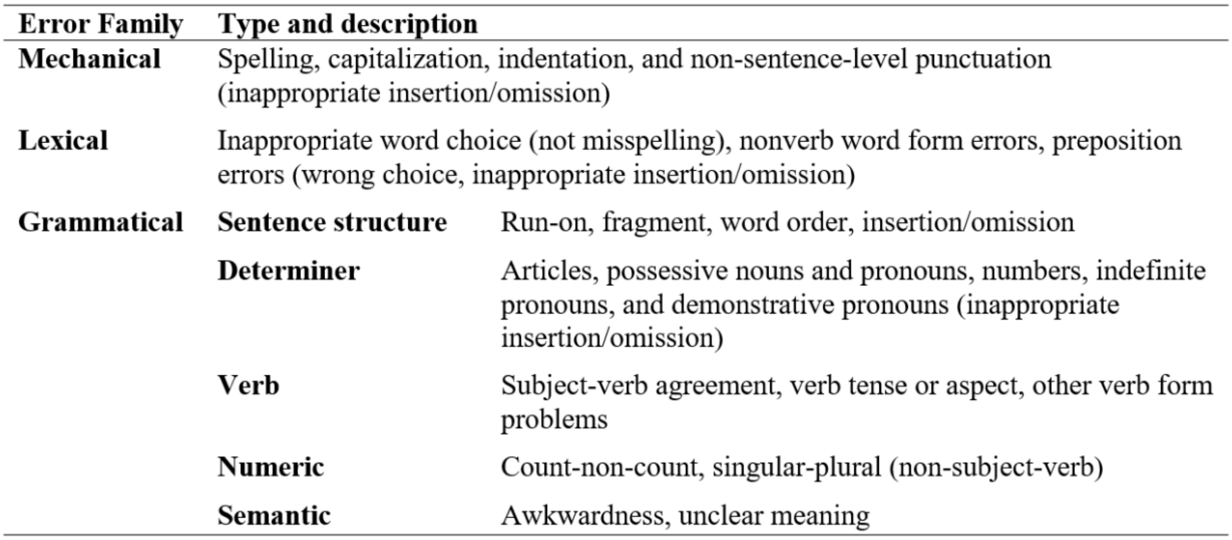
Procedures

The methods used to rate student samples were the General Accuracy Ratio (GAR) and the Error-Free Clause Ratio (EFCR). The GAR method measures student accuracy by identifying the total number of errors that a student makes calculated by the formula $1 - \left(\frac{errors}{words}\right)$. The EFCR method measures accuracy by identifying the number of error-free clauses over the number of total clauses represented by the formula $\frac{error\ free\ clauses}{total\ clauses}$.

The raters were intentionally given little training in order to mirror conditions of TESOL teachers in the field. They were provided two documents, one to help identify what constituted an error (Table 3) and another to understand where to break clauses when using the EFCR method (see Appendix B). Each rater was assigned 10 students who were made anonymous by number, and they rated the students’ pre and posttests twice, once using the GAR method and once using the EFCR method.

Table 3

Types of Errors



Error Family	Type and description	
Mechanical	Spelling, capitalization, indentation, and non-sentence-level punctuation (inappropriate insertion/omission)	
Lexical	Inappropriate word choice (not misspelling), nonverb word form errors, preposition errors (wrong choice, inappropriate insertion/omission)	
Grammatical	**Sentence structure**	Run-on, fragment, word order, insertion/omission
	Determiner	Articles, possessive nouns and pronouns, numbers, indefinite pronouns, and demonstrative pronouns (inappropriate insertion/omission)
	Verb	Subject-verb agreement, verb tense or aspect, other verb form problems
	Numeric	Count-non-count, singular-plural (non-subject-verb)
	Semantic	Awkwardness, unclear meaning

To ensure that raters would not allow previous ratings with one method to influence new ratings with the other method on the same sample, raters worked in a stepwise motion down their assigned ratings, which were recorded in an Excel spreadsheet (see Appendix C). Some assigned samples overlapped to ensure inter-rater reliability. Raters timed themselves rating each sample and did not collaborate with one another. They rated at home and in the classroom, and when raters had questions about the samples, errors, or methods, they asked their professor.

Data Analysis

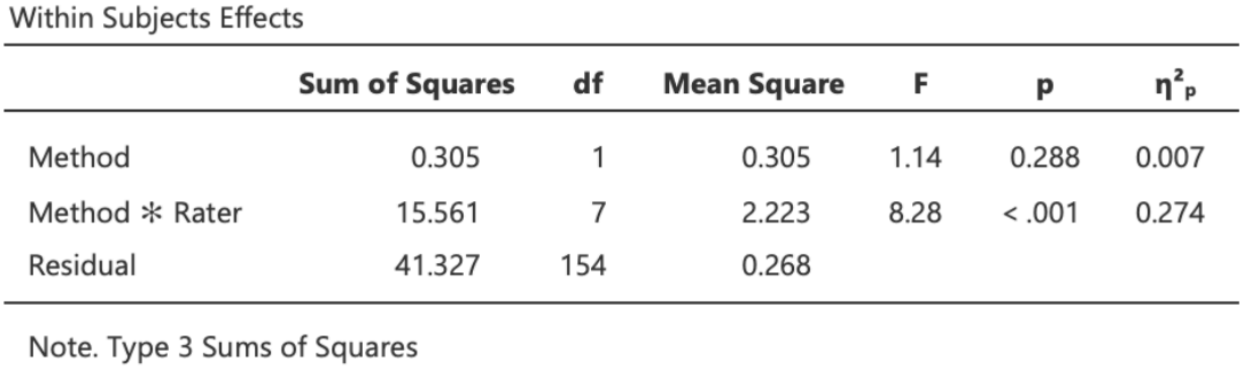
The rating process took 6 weeks, at the end of which every rater re-rated their first four samples for continuity since they were now more accustomed to the methods. Results were then sent to their professor who ran Repeated Measures Analysis of Variance (ANOVA) tests.

Results

To answer our research questions, we used an ANOVA test. We compared results on time for each rater and between raters according to both the GAR and the EFCR methods and analyzed student improvement over time. These results are presented in Table 4. The tests showed no significant finding for seconds per word by method, F(1, 154) = 1.14, p = 0.288, η^2_p = 0.007. Neither was there a significant difference between the GAR method time (*M* = 1.54, *SD* = 0.767) and the EFCR method time (*M* = 1.47, *SD* = 0.642). However, the method by rater calculation did yield significant results, F(1, 154) = 8.28, = p<0.001, η^2_p = 0.274.

Table 4

Repeated Measures ANOVA (Seconds per Word)



Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Method	0.305	1	0.305	1.14	0.288	0.007
Method ✻ Rater	15.561	7	2.223	8.28	< .001	0.274
Residual	41.327	154	0.268			

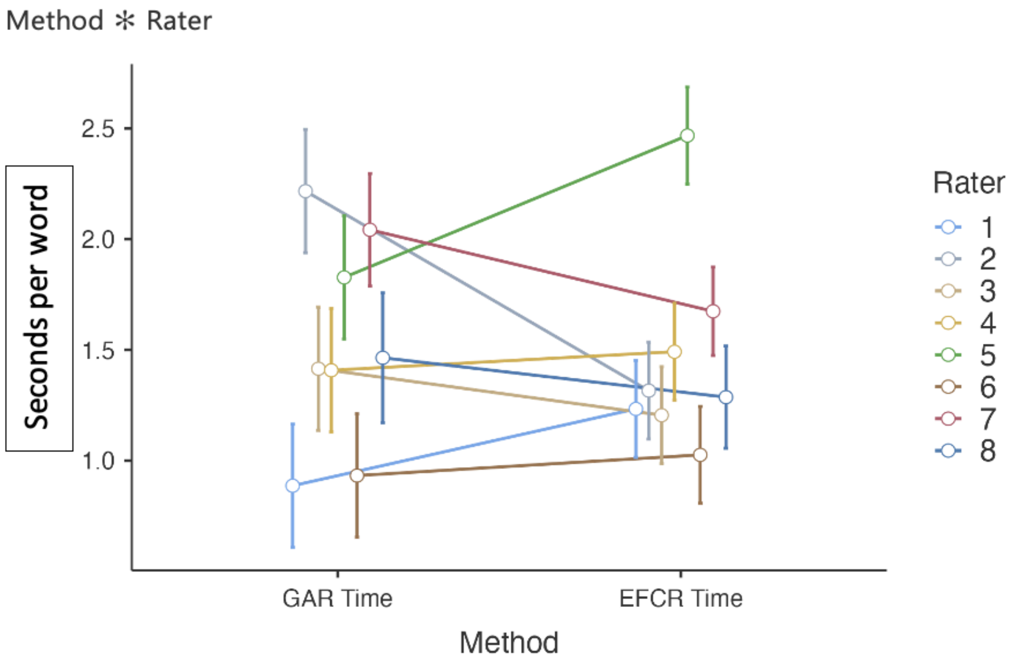
Note. Type 3 Sums of Squares

This result is further detailed in Figure 1, which illustrates the individual raters' seconds per word according to the GAR and EFCR methods. Raters 1, 4, 5, and 6 increased in time from the GAR to the EFCR methods, with rater 5 having the greatest mean gain of 0.64. However, raters 2, 3, 7, and 8 all showed a decrease in time from the GAR to the EFCR method, with rater 2 having the greatest mean loss of 0.9. Because half of the raters had a lower mean time for the GAR method, and half had a lower mean time for the EFCR method, there was no significant difference between the two methods in terms of time.

Regarding validity, both the EFCR pretests (M = 0.165, SD = 0.118) and posttests (M = 0.252, SD = 0.162), as well as the GAR pretests (M = 0.771, SD = 0.0676) and posttests (M = 0.845, SD = 0.137) showed near parallel gains in terms of accuracy. These results are represented in Table 5. The ANOVA test showed there was significant improvement in student accuracy

Figure 1

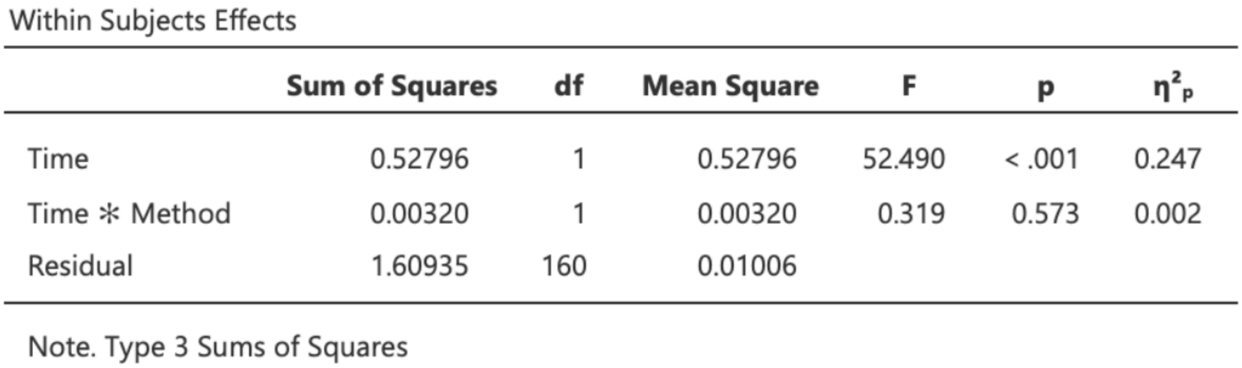
Rater Seconds per Word



scores in both methods over time, F(1, 160), = 52.49, p < 0.001, η^2_p = 0.247. However, when measuring time by method, no significance was shown, F(1, 160), = 0.319, p = 0.573, η^2_p = 0.002.

Table 5

Repeated Measures ANOVA (Accuracy)



Within Subjects Effects

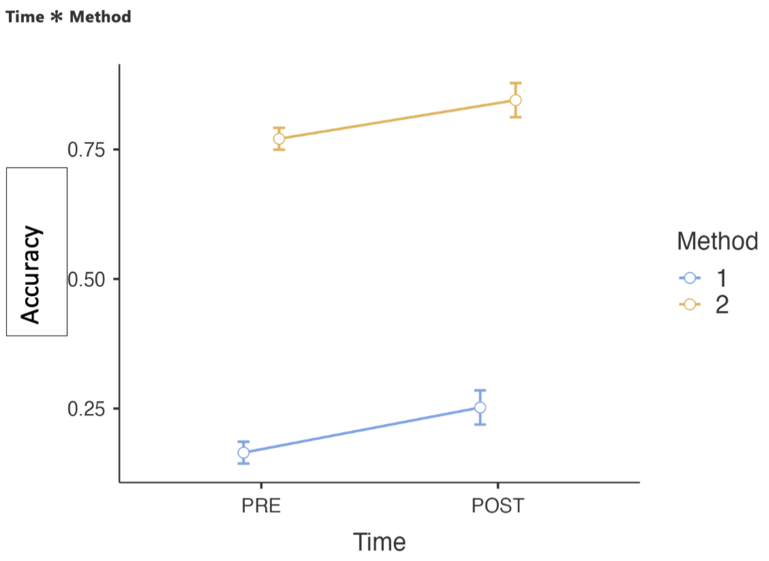
	Sum of Squares	df	Mean Square	F	p	η^2_p
Time	0.52796	1	0.52796	52.490	< .001	0.247
Time ✻ Method	0.00320	1	0.00320	0.319	0.573	0.002
Residual	1.60935	160	0.01006			

Note. Type 3 Sums of Squares

The interaction of accuracy by time is displayed in Figure 2. The GAR method resulted in a pretest mean of 0.771 and a posttest mean of 0.845 with a gain of 0.074 (7.5%). Additionally, the EFCR method resulted in a pretest mean of 0.165 and a posttest mean of 0.252 with a gain of 0.087 (8.7%).Both methods on the graph show an improvement in accuracy in students' writing between the pre and posttests.

Figure 2

Time by Method (Accuracy)



Discussion

The data collected from the research indicates that the methods were comparable in terms of practicality. The mean number of seconds per word for the GAR method was 1.54 seconds while for the EFCR method, it was 1.47 seconds, which is not a statistically significant difference. However, the time between raters was significant, meaning that some raters took much longer using the GAR method while others took much longer using the EFCR method. Although the reasons for this are not entirely known, the identification of clauses with the EFCR method was likely easier for some, while perhaps being overly concerned with precise accounting of each error with the GAR method cost others a considerable amount of time. This indicates that the rater matters in determining which method is more practical.

Another criterion for comparing the two methods is validity, including predictive validity. Figure 2 shows that there was an increase in student accuracy in their writing between pre and posttests for both methods. Interestingly, Figure 2 also shows that the gains in accuracy over time were nearly parallel, indicating predictive validity between these two methods.

Concurrent validity was also evaluated based on the results. Referencing Figure 2, one may easily see that these methods report vastly different measurements of accuracy. The GAR method shows an approximate 75-80% accuracy while the EFCR method shows an approximate 20-25% accuracy. This is because the two ratings measured the samples in distinct manners. The GAR method measures errors per word count, whereas the EFCR method measures the number of error-free clauses per total number of clauses. Figure 3 depicts that if a student has more distributed errors, one in each clause, then the GAR method result would be .90, but the EFCR method result would be .00. On the other hand, if the student made all their errors in one clause, the GAR method result would be .90 and the EFCR method result would be .90 as well. Thus, the more concentrated the errors are in using the EFCR method, the more accurate the writing will appear. Conversely, using the GAR method as a measurement, the errors will be counted equally whether found in the same clause or not.

This is also illustrated by Figures 4 and 5, which show a rater’s calculations of the same passage using the EFCR and GAR methods. As demonstrated by Figure 4, the rater used the EFCR method to identify the number of *correct* clauses. In contrast, the GAR method, featured in Figure 5, measured the total number of *errors* made compared to the word count.

Figure 3

Theoretical Musings on Differences

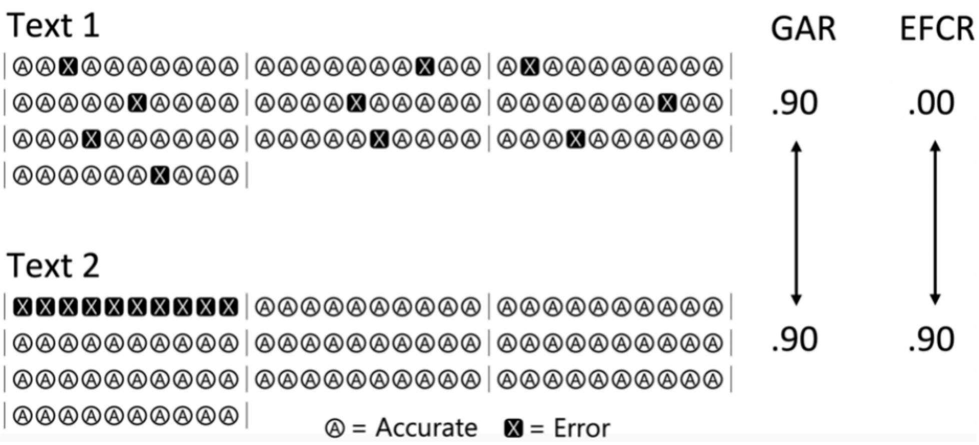
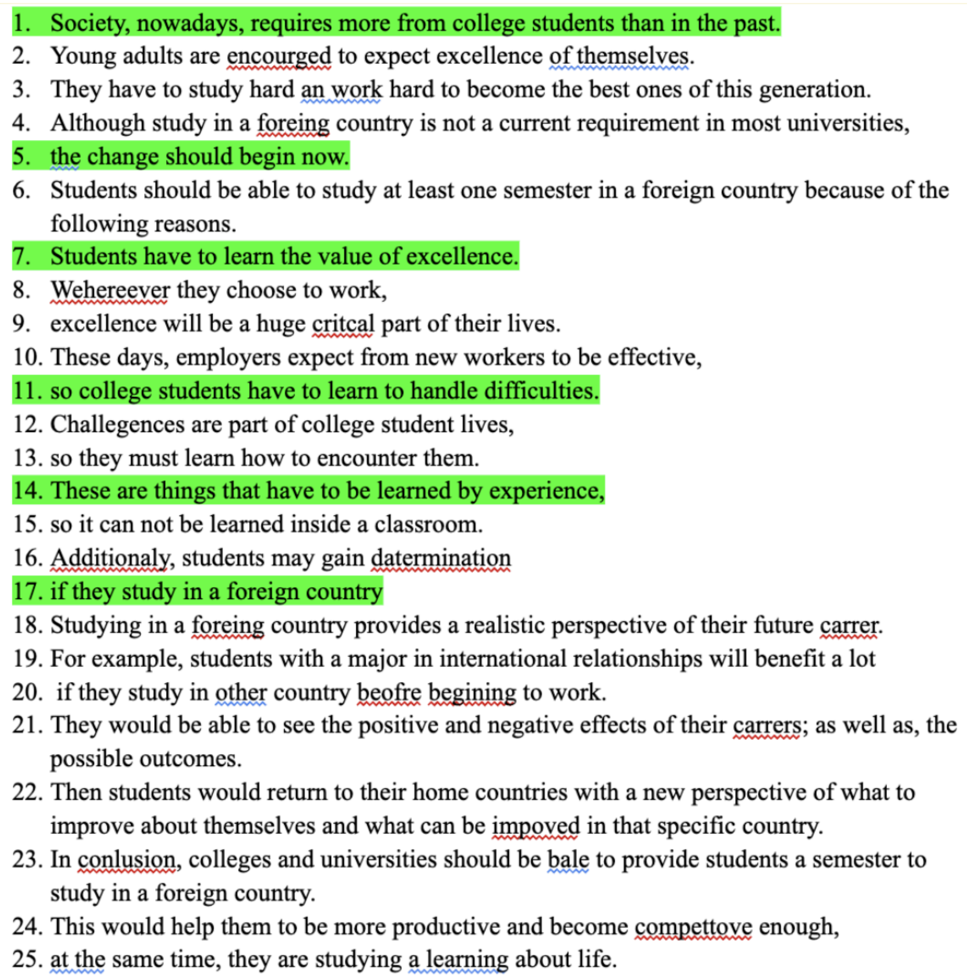


Figure 4

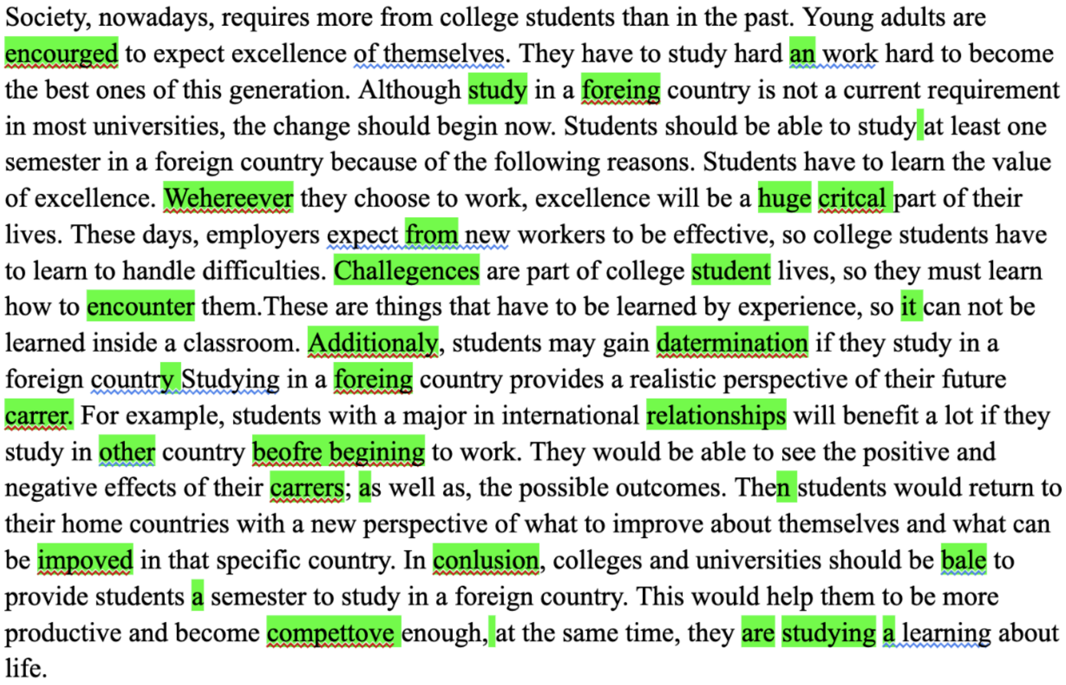
EFCR Method: Correct Clauses Identified



1. Society, nowadays, requires more from college students than in the past.
2. Young adults are encourged to expect excellence of themselves.
3. They have to study hard an work hard to become the best ones of this generation.
4. Although study in a foreing country is not a current requirement in most universities,
5. the change should begin now.
6. Students should be able to study at least one semester in a foreign country because of the following reasons.
7. Students have to learn the value of excellence.
8. Wehereever they choose to work,
9. excellence will be a huge critcal part of their lives.
10. These days, employers expect from new workers to be effective,
11. so college students have to learn to handle difficulties.
12. Challegences are part of college student lives,
13. so they must learn how to encounter them.
14. These are things that have to be learned by experience,
15. so it can not be learned inside a classroom.
16. Additionaly, students may gain datermination
17. if they study in a foreign country
18. Studying in a foreing country provides a realistic perspective of their future carrer.
19. For example, students with a major in international relationships will benefit a lot
20. if they study in other country beofre begining to work.
21. They would be able to see the positive and negative effects of their carrers; as well as, the possible outcomes.
22. Then students would return to their home countries with a new perspective of what to improve about themselves and what can be impoved in that specific country.
23. In conlusion, colleges and universities should be bale to provide students a semester to study in a foreign country.
24. This would help them to be more productive and become compettove enough,
25. at the same time, they are studying a learning about life.

Figure 5

GAR Method: Errors in Isolation



Society, nowadays, requires more from college students than in the past. Young adults are encourged to expect excellence of themselves. They have to study hard an work hard to become the best ones of this generation. Although study in a foreing country is not a current requirement in most universities, the change should begin now. Students should be able to study at least one semester in a foreign country because of the following reasons. Students have to learn the value of excellence. Wehereever they choose to work, excellence will be a huge critcal part of their lives. These days, employers expect from new workers to be effective, so college students have to learn to handle difficulties. Challegences are part of college student lives, so they must learn how to encounter them.These are things that have to be learned by experience, so it can not be learned inside a classroom. Additionaly, students may gain datermination if they study in a foreign country Studying in a foreing country provides a realistic perspective of their future carrer. For example, students with a major in international relationships will benefit a lot if they study in other country beofre begining to work. They would be able to see the positive and negative effects of their carrers; as well as, the possible outcomes. Then students would return to their home countries with a new perspective of what to improve about themselves and what can be impoved in that specific country. In conlusion, colleges and universities should be bale to provide students a semester to study in a foreign country. This would help them to be more productive and become compettove enough, at the same time, they are studying a learning about life.

This writing sample portrays a distinction between the two methods. The disparity is partly because the EFCR method separates the sample into clauses, while the GAR method keeps it intact. In the EFCR method, once one error is detected in a clause, the clause is considered inaccurate. One example is in clause 24 of Figure 4. This clause contains 12 words. However, because word 11 is spelled incorrectly, ("compettove" as opposed to "competitive"), the entire clause is considered inaccurate. This automatically marks 12 words as inaccurate. Alternatively, Figure 5 shows that in using the GAR method, each error is considered in isolation. For instance, this same spelling error ("compettove") can only be marked as one error.

Irrespective of the disparity in inaccuracy types between the EFCR and the GAR methods, the gravity of errors using the EFCR method may prove useful for certain contexts, such as when rating high-proficiency students' writings. Using the EFCR method, these high-proficiency students can be held to a much higher standard of accuracy because they will be expected to produce accurate thought groups or clauses. This might keep these students from "hitting a ceiling". This ceiling effect may occur when high-proficiency students consistently perform well. For example, if most high-proficiency students consistently have a high accuracy rating according to the GAR method measurement, then they can only improve by a small amount. In contrast, a lower accuracy shown by the EFCR method encourages these high-proficiency students to work hard to improve by producing accurate clauses.

An example of this concept can be observed in Figure 4, which demonstrates the rating of a high-proficiency student's writing. The rater recorded 34 errors using the GAR method. Because there were 287 total words, the accuracy ratio is 0.88. This leaves little room for accuracy improvement for this writer. On the other hand, Figure 5 shows the rating of the same writing sample using the EFCR method. The rater counted 6 error-free clauses out of 25 total clauses, which renders an accuracy of 0.24. Some of the errors found in the sample include spelling errors, word omissions, word insertions, punctuation, verb conjugations, etc. While using the EFCR method for identifying small errors in clauses such as spelling errors may be considered severe, it allows for a higher standard of accuracy when the context demands this.

Contexts where a higher accuracy standard using the EFCR method may be beneficial could include high-proficiency ESL and EFL classes, grammar-focused writing courses, courses for university-bound or university-matriculated students, and courses for students preparing for high-stakes assessments where grammatical accuracy plays a large role in the scoring. In contrast, in lower-proficiency ESL and EFL classes where producing accurate thought groups is neither feasible nor expected, using the GAR method may be beneficial as it allows for a more lenient view of accuracy.

Limitations and Future Research

One of the limitations of this study is that students were given different writing prompts. Even though the results focused on measuring accuracy and time, it is possible that the variation of the prompt affected student performance in terms of accuracy. Another limitation could be the potential lack of consistency in how raters segmented the provided texts into clauses. In the future, researchers could compare the number of clauses segmented per passage by raters by creating a correlation coefficient, which would contribute to the validity of the results.

Likewise, another possible approach in the future could be to establish better inter-rater reliability. Raters were given minimal training. This was not necessarily a disadvantage since it contributed to the ecological validity of the research, but there are possible advantages to improved rater training. If raters are to be trained, a clearer rubric could be established for rating written papers, which would leave less room for mistakes in identifying clauses and errors.

Another limitation was that all mistakes were considered with the same weight. Although it would change an aspect of the methods, the gravity of different mistakes could be evaluated. Classifying the mistakes based on their type (lexical, spelling, grammatical, punctuation, etc.) may increase the time needed to rate papers. Nonetheless, having a list that clearly points out which mistakes are most common would likely serve as a formative guide for teachers who could then focus their teaching on correcting the most common mistakes.

The lack of qualitative data due to time constraints is also a limitation. An area where a qualitative analysis might prove useful is in examining the differences in rater efficiency between the two methods. In the future, researchers might consider conducting rater interviews or focus groups to determine the differences in the raters' approaches for each measurement. Such data may explain the variance in rater timing observed in this study and whether raters in the future need more training to use a method efficiently.

Additionally, as mentioned previously, some raters might have found it more difficult to identify clauses with the EFCR method, whereas other raters labored over identifying individual errors with the GAR method. It is possible that the same pattern of rater efficiency, based on strengths, could be found in various contexts. In this study, all of the raters were TESOL graduate students at an American university. Even though the implications of this research are most valuable to teachers of English as a second language, results could be valuable to anyone who grades written papers. These contexts may include public and private school teachers in child, adolescent, and adult education. Therefore, research with raters in various contexts should be conducted in order to determine if the pattern observed within this study is consistent.

Conclusion

The GAR and EFCR methods are two distinct approaches to measuring the grammatical accuracy of student writing samples. Thus, they cannot be used interchangeably, and one method is not better than the other in regard to practicality and validity. Rather, we recommend that researchers and practitioners evaluate which method best fulfills the purposes of their assessment. Such an evaluation might consider program and course objectives, student goals, and proficiency level, as each of these could play a role in determining which method would best suit practitioners and students.

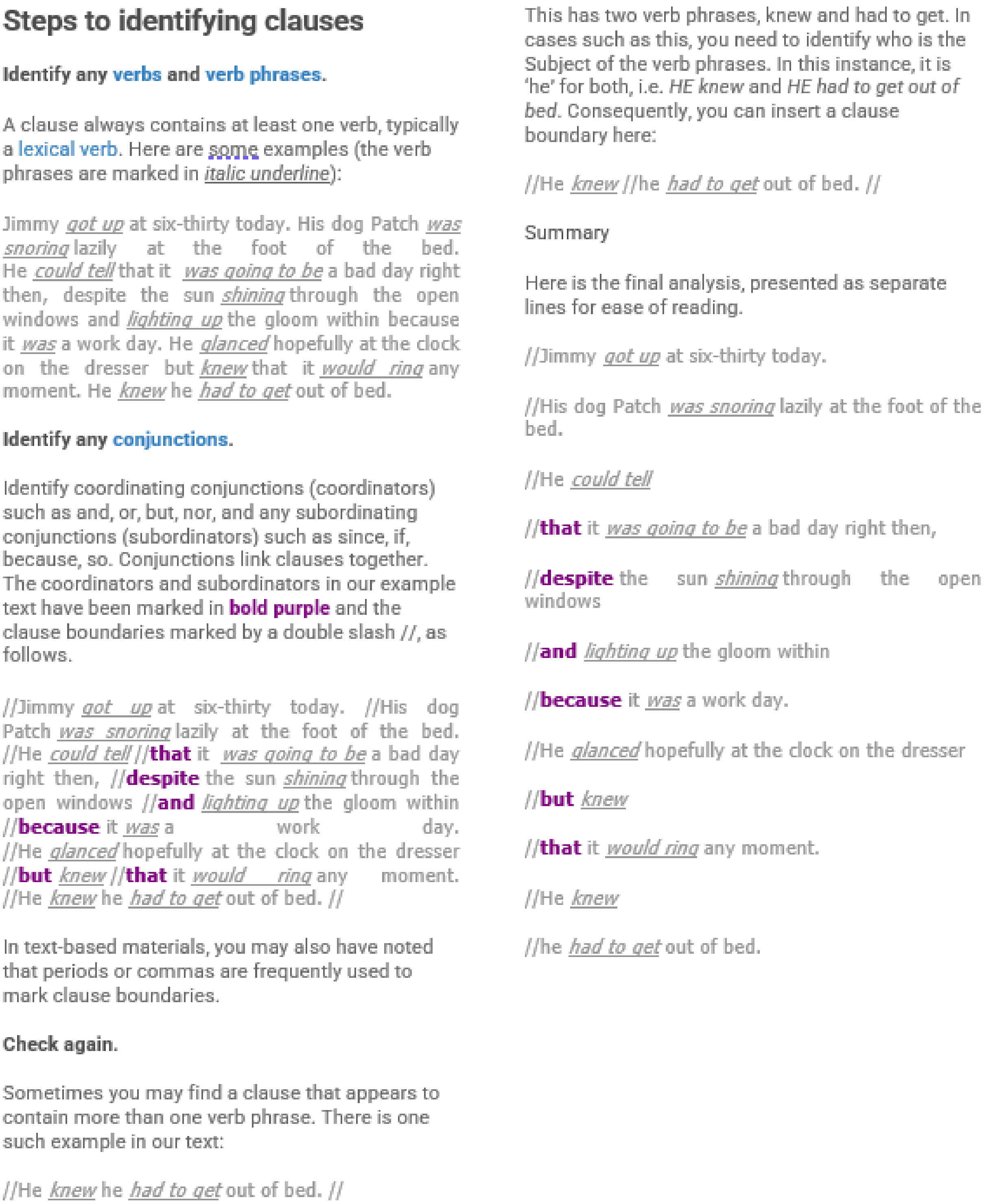
Appendix A

Number of students who answered each prompt for pre and posttests

Prompt	Pretest	Posttest	Total
1: What is the most important animal in your country? Why is the animal important? In the future will this animal's importance increase, decline or remain the same? Use reasons and specific details to explain your answer. You have 30 minutes to write your response.	14	4	18
2: Some people say that physical exercise should be a required part of every school day. Other people believe that students should spend the whole school day on academic studies. Which opinion do you agree with? Use specific reasons and details to support your opinion and describe the potential immediate and long-term consequences of this opinion. You have 30 minutes to write your response.	14	21	35
3: Do you agree or disagree with this statement? “Colleges and universities should require their students to spend at least one semester studying in a foreign country.” Use specific reasons and examples to support your opinion and describe the potential immediate and long-term consequences of this opinion	16	19	35
4: Identify one improvement that would make your city a better place to live for people your age and explain why people your age would benefit from this change. Use specific reasons and examples to support your opinion and describe the potential immediate and long-term consequences of this improvement.	16	17	33
5: Write at least a paragraph about a school that you attended in the past. This could be an elementary, high school, or college. What did this place look like inside and outside? Was it a good or poor place to study? Why do you think that? You have 10 minutes to write your response.	1	0	1
Total	**61**	**61**	**122**

Appendix B

Rater training document: Step to identifying clauses



Steps to identifying clauses

Identify any verbs and verb phrases.

A clause always contains at least one verb, typically a lexical verb. Here are some examples (the verb phrases are marked in *italic underline*):

Jimmy *got up* at six-thirty today. His dog Patch *was snoring* lazily at the foot of the bed. He *could tell* that it *was going to be* a bad day right then, despite the sun *shining* through the open windows and *lighting up* the gloom within because it *was* a work day. He *glanced* hopefully at the clock on the dresser but *knew* that it *would ring* any moment. He *knew* he *had to get* out of bed.

Identify any conjunctions.

Identify coordinating conjunctions (coordinators) such as and, or, but, nor, and any subordinating conjunctions (subordinators) such as since, if, because, so. Conjunctions link clauses together. The coordinators and subordinators in our example text have been marked in **bold purple** and the clause boundaries marked by a double slash //, as follows.

//Jimmy *got up* at six-thirty today. //His dog Patch *was snoring* lazily at the foot of the bed. //He *could tell* //**that** it *was going to be* a bad day right then, //**despite** the sun *shining* through the open windows //**and** *lighting up* the gloom within //**because** it *was* a work day. //He *glanced* hopefully at the clock on the dresser //**but** *knew* //**that** it *would ring* any moment. //He *knew* he *had to get* out of bed. //

In text-based materials, you may also have noted that periods or commas are frequently used to mark clause boundaries.

Check again.

Sometimes you may find a clause that appears to contain more than one verb phrase. There is one such example in our text:

//He *knew* he *had to get* out of bed. //

This has two verb phrases, knew and had to get. In cases such as this, you need to identify who is the Subject of the verb phrases. In this instance, it is 'he' for both, i.e. *HE knew* and *HE had to get out of bed*. Consequently, you can insert a clause boundary here:

//He *knew* //he *had to get* out of bed. //

Summary

Here is the final analysis, presented as separate lines for ease of reading.

//Jimmy *got up* at six-thirty today.

//His dog Patch *was snoring* lazily at the foot of the bed.

//He *could tell*

//**that** it *was going to be* a bad day right then,

//**despite** the sun *shining* through the open windows

//**and** *lighting up* the gloom within

//**because** it *was* a work day.

//He *glanced* hopefully at the clock on the dresser

//**but** *knew*

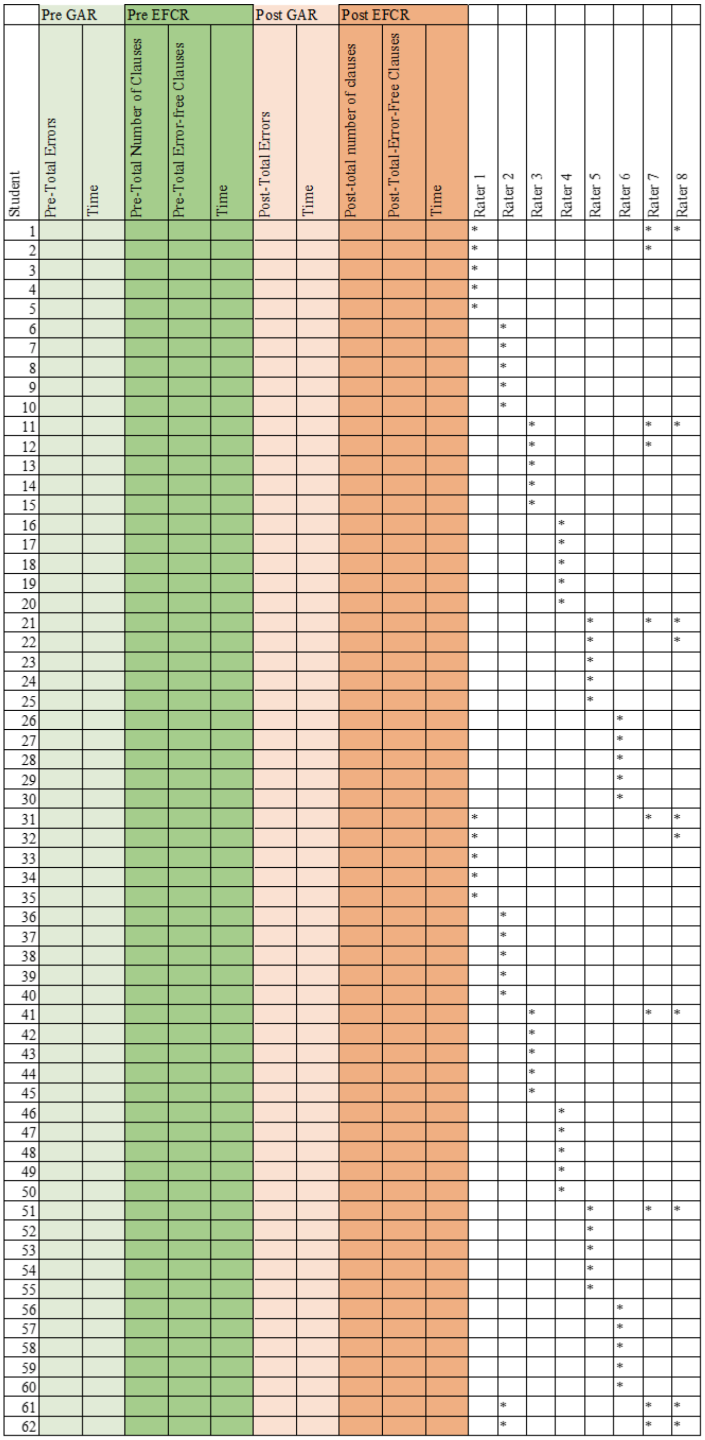
//**that** it *would ring* any moment.

//He *knew*

//he *had to get* out of bed.

Appendix C

Excel spreadsheet for individual raters to record ratings



	Pre GAR		Pre EFCR			Post GAR		Post EFCR										
Student	Pre-Total Errors	Time	Pre-Total Number of Clauses	Pre-Total Error-free Clauses	Time	Post-Total Errors	Time	Post-total number of clauses	Post-Total-Error-Free Clauses	Time	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8
1											*						*	*
2											*						*	
3											*							
4											*							
5											*							
6												*						
7												*						
8												*						
9												*						
10												*						
11													*				*	*
12													*				*	
13													*					
14													*					
15													*					
16														*				
17														*				
18														*				
19														*				
20														*				
21															*		*	*
22															*			*
23															*			
24															*			
25															*			
26																*		
27																*		
28																*		
29																*		
30																*		
31											*						*	*
32											*							*
33											*							
34											*							
35											*							
36												*						
37												*						
38												*						
39												*						
40												*						
41													*				*	*
42													*					
43													*					
44													*					
45													*					
46														*				
47														*				
48														*				
49														*				
50														*				
51															*		*	*
52															*			
53															*			
54															*			
55															*			
56																*		
57																*		
58																*		
59																*		
60																*		
61												*					*	*
62												*					*	*