Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2012-06-11

A Study in Computerized Translation Testing (CTT) for the Arabic Language

Amanda J. Kuhn
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Linguistics Commons

BYU ScholarsArchive Citation

Kuhn, Amanda J., "A Study in Computerized Translation Testing (CTT) for the Arabic Language" (2012). *All Theses and Dissertations*. 3108.
https://scholarsarchive.byu.edu/etd/3108

This Selected Project is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.

A Study in Computerized Translation Testing (CTT)

for the Arabic Language

Amanda J. Kuhn

A selected project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Alan K. Melby, Chair
R. Kirk Belnap
Dan P. Dewey

Department of Linguistics and English Language

Brigham Young University

June 2012

Copyright © 2012 Amanda J. Kuhn

All Rights Reserved

ABSTRACT

A Study in Computerized Translation Testing (CTT)
for the Arabic Language

Amanda J. Kuhn
Department of Linguistics and English Language
Master of Arts

Translation quality assessment remains pertinent in both translation theory and in the industry. Specifically, the process of assessing a target document's quality or a person's translation competence involves a lot of time and money on the part of various governments, organizations and individuals. In response to this issue, this project builds on the ongoing research of Hague et al. (2012), who seek to determine the capabilities of a computerized translation test for the French-to-English and Spanish-to-English language pairs. Specifically, Hague et al. (2012) question whether a good score on a detect-and-correct style computerized translation test that is calculated by a computer also indicates a good score on a traditional full translation test that is calculated by hand. This project seeks to further this research by seeking to answer the same question using an Arabic-to-English language pair.

The methods used in this research involve testing individuals using two different style translation tests and then comparing the results. The first style translation test involves a detect-and-correct format where a subject is given a list of project specifications in the form of a translation brief, a source text passage and a corresponding target text passage that has errors introduced throughout. The subject is expected to detect and fix the errors while leaving the rest of the text alone. A score is given for this test using an automated algorithm. The second style test is a traditional translation test where a subject is given the same translation brief and a source text. The subject is expected to produce an acceptable target text, which is subsequently scored by hand. Thereafter, various forms of analysis are used to determine the relationship between the scores of the two types of tests.

The results of this research do not strongly suggest that a high score on the detect-and-correct portion of the test indicates a high score on a hand-graded full translation test for the subject population used. However, this research still provides insight, especially concerning whether the detect-and-correct portion of the test actually measures translation competence and concerning second language acquisition (SLA) programs and their intentions. In addition, this research provides insight into logistical issues in testing such as the impact text difficulty and length may have on a detect-and-correct style test as well as the negative impact the American Translators Association (ATA) grading practices of weighting errors and capping errors can have on an experiment such as the one described in this research.

Keywords: Translation Quality Assessment, Computerized Translation Testing, Arabic Translation Assessment

ACKNOWLEDGEMENTS

The author wishes to express sincere appreciation to the many individuals who helped in completion of this project especially Dr. Alan Melby for his time, patience, expertise and extensive support throughout the entire process. Dr. Kirk Belnap and Dr. Dan Dewey must also be thanked for their support and expertise during many crucial parts of the journey. Another person of critical importance to this project was Dr. Paul Fields who offered his kindness, time and much needed expertise. Other individuals who deserve mention include Nettie Downs, Reem Naouri, and Khalil AbuManneh. This project would have never been completed without the support of these key individuals as well as family and friends.

Table of Contents

List of Figures

List of Charts

List of Images

INTRODUCTION

Because international communication encompasses a large portion of the daily interchange in this world, the translation field has become an actual industry. A point of concern in the translation industry and in translation theory is that of translation quality: quality of a translator and quality of a translation product. In an effort to promote the idea that translation quality is relative to specifications, Durban and Melby (2008) produced a pamphlet that the American Translators Association (ATA) circulates as one of its publications. This pamphlet explains that because of the variety of items that need to be specified for each translation project (specifications), translation cannot be a commodity (p.3). It further clarifies that the translation quality is the "degree to which it [the translation product] follows the agreed-upon specifications" (p.4). One can conjecture that if the largest organization of translators in America produces a pamphlet to campaign the idea of translation quality to society at large, then quality stands as an important concern to the industry. In addition, the field of translation theory has also produced much work concerning translation quality, specifically in the area of defining translation quality, which in turn helps define how to assess translation quality. Because methods of defining and assessing translation quality vary within and between both of these sectors, quality remains a pertinent topic of discussion in today's world.

Translation quality assessment remains particularly important not only in the field of translation theory, but also in the industry. Multiple scholars in the theoretical field have stressed translation quality assessment. Similarly, the translation industry stresses the importance of translation quality assessment as well in discussions of translators earning credentials and quality assurance. Because quality assurance can be defined as a part of the translation process, and the purpose of this research is to investigate translation quality as the

quality of a translator and the quality of a translation product, the former discussion of earning credentials will be discussed here.

Stejskal's (2003) report of his two year examination of the means of earning credentials in the field of translation and interpretation in over 30 countries and six continents clearly shows the importance of translation quality assessment in the translation industry. He outlines four methods of credentialing and mentions various types of assessment from the use of "rigorous assessments of knowledge and skills," to providing moral and academic credentials alone (p. 16). This clearly shows that non-assessment credentialing does exist. However, it also shows that translation quality assessment as defined by both the quality of a translator and the quality of a translation product is an important part of the credentialing process in the industry. In addition, Stejskal reviews four general arguments for the necessity of credentials, "to establish standards of professional practice; to elevate the status of the profession; to satisfy public demand for standards; and to extend the 'shelf life' of academic degrees through continuous professional development" (2003, p. 15). One can easily see how these arguments for the necessity of credentials can easily be used to argue the necessity of translation quality assessment.

However, just as the act of translation takes large amounts of time and money, many forms of translation quality assessment do as well. For example, the ATA offers a certification exam in which one's translation ability is determined by the quality of two sample translation products. This certification exam starts at $300 dollars, which pays solely for administration and grading costs. In addition a person must allow a minimum of 15 weeks before his or her score is received, and even more time during busier parts of the year ("ATA Certification," 2011). This problem is compounded by the fact that the percentage of people that pass the ATA Certification Exam has been below 20 percent for the past 15 years despite efforts to limit the number of

unqualified people taking the exam by raising the required prerequisites (Koby and Champe, in press). In addition, time is also an issue for many academic programs of translation when assessing the translation quality of students. Examples such as these lead one to consider the possibility of whether technology might assist the industry and theoretical field with these problems of time and money, just as it has in many other industries.

Attempting to solve problems of time and money is not new to most fields and is not new to the translation field. Many technologies have been developed that automatically assess essays (Dikli, 2006). In addition, metrics like BLEU (Papenini et al., 2002) and NIST (Doddington et al., 2002) have been developed to automatically assess machine translation. While these methods of assessment are justly criticized for their shortcomings, they are still highly useful in certain situations and for certain purposes. Therefore, for certain situations and purposes, a partially automated translation test might be better in terms of cost and time as well as in assessing human translation. However, this is certainly not the case for all situations or purposes.

Hague et al. (2012) address the possibility of computerized translation testing (CTT). These researchers conducted a study on Brigham Young University (BYU) campus using a CTT testing tool, which they developed. Their CTT testing tool includes, among other things, a two part test: a detect-and-correct style test and a full translation test. The former includes a source text and a faulty English target text in which students are expected to correct faulty chunks of text and leave the non-faulty chunks alone. The latter involves a different source text which the students are expected to translate. The language pairs they seek to test include French-English and Spanish-English. Specifically, Hague et al. (2012) seek to determine whether an automated score on a detect-and-correct version of the test predicts a good score on a traditional hand-graded full translation test. To determine whether or not this occurs they compare the computer-

graded detect-and-correct scores to the human-graded scores of the full translation test. If a strong correlation is present between these scores, then this suggests that future research concerning tools like the CTT testing tool might be able to help organizations like the ATA quickly and cost-effectively "weed out" unqualified individuals from applying for its certification exam. Interestingly, the idea of a screening test has been considered by the ATA organization but was dismissed on account of technical and logistical difficulties (Stejskal, 2003). Further research like that of Hague et al. (2012) may be able to sort through those technical and logistical difficulties. To date, the results of this study of Hague et al. (2012) are still being analyzed and have not yet been published.

The goal of my research is not to develop a computerized screening test that weeds out individuals without the ability to pass the ATA certification test. My research seeks to find out whether computerized translation testing is a valid way to test translation quality. Thus, my research is a replication of the research of Hague et al. (2012) with the exception that my research furthers theirs by assessing the quality of a translation from given an Arabic-to-English language pair. Therefore, the main questions my study seeks to answer are as follows: Can a computer-graded test accurately, or even semi-accurately, predict translation quality? Specifically, do the scores of an individual taking the detect-and-correct translation assessment test correlate significantly with or indicate a good score on the hand-graded scores of a full translation test, using an Arabic source text and an English target text? If they do correlate significantly, what are the implications of that correlation, and if they do not correlate significantly, what are the implications of this lack of correlation, given an Arabic to English language pair?

REVIEW OF LITERATURE

A number of current discussions in the literature provide insight into the questions presented in the previous section. These discussions form the basis of the assumptions this research makes concerning translation competence and quality. One assumption holds that translation competence and language proficiency are two different, but related abilities. If this were not true, one could argue that translation quality testing is not an important or necessary topic of discussion because it belongs under the category of language proficiency testing. The second assumption is to define the quality of a translation product as only part of the quality of a translator. A third assumption is to use a functionalist approach to defining and assessing quality. This particular functionalist approach is based on fulfilling predetermined specifications, which makes it a manufacturing-styled approach to quality assessment.

Translation Competence and Language Proficiency

First, this paper assumes that translation competence and language proficiency are two different but related abilities. It is important to note a matter of terminology. Generally, in the literature, authors like Hague et al. (2011) use the word "competence" to discuss the concept that one possesses the knowledge, skills and abilities involved in translation while this in the field of second language acquisition (SLA), usually uses the word "proficiency" to discuss the knowledge, skills and abilities involved in communication through a given language as seen in the ACTFL Proficiency Guidelines. These practices will be used in this paper. Second, one may question why simply speaking a foreign language fluently does not qualify one to be a competent translator. This argument rests on the idea that a competent translator must have skill sets beyond being a proficient second language (L2) speaker. The word "beyond" is key because it denotes that a "good" translator must already have the ability of a proficient L2 speaker. The

opposite is not required: a proficient speaker need not have the skill sets of a competent translator in order to be a proficient speaker. One can prove this concept by distinguishing the skills of a competent translator from the skills of a proficient speaker. Both translation theory and acquisition theory provide various examples of criteria that denote proficiency and competence, respectively. Many of the skills on which each field founds its definition of a competent translator or proficient speaker are found in the measurement criterion of their proficiency and certification tests. By comparing the criterion by which L2 speakers and translators are measured, one can denote how a competent translator differs from a proficient speaker.

A Basis for Language Proficiency Testing

First, like an approach to translation theory defines how one assesses translation quality, a language acquisition theory influences proficiency testing. Many language acquisition theories may be placed on a spectrum whose poles can be defined as being "for" or "against" a specialized language acquisition device (LAD). Those that are "for" this device believe that language is a special, innate part of the human faculty and that it is so special that there is a device separate from other cognitive devices that allows a human to develop language. Theories like Universal Grammar (UG) apply to this side of the spectrum (Chomsky, 1965). On the other hand those who may be considered "against" a unique language acquisition device argue that language is not special but rather is learned through cognitive processes just as other aspects of life. For example, connectionist models hold that learners use every experience they have had and new experiences are compared to the old (McClelland et al., 1986). These models fall towards this side of the spectrum.

However, a connectionist approach is only one viewpoint on the side of the spectrum against a specialized LAD. In fact, Elis describes viewpoints that fall on this side of the spectrum as "constructivist views." In general, these viewpoints believe that

> …simple learning mechanisms operating in and across human systems for perceptions, motor action, and cognition while exposed to language data in a communicatively rich human environment navigated by an organism eager to exploit the functionality of language are sufficient to drive the emergence of complex language representations (2003, p. 63).

Elis further describes those who adhere to various versions of this approach as: connectionists like Christiansen and Chater 2001; functional linguists like Bates and MacWhinney 1981; emergentists like Elman, Bates, Johnson, Karmiloff-Smith, Parisi, and Plunket 1996; cognitive linguists like Croft and Cruse 1999; constructivist child language researchers like Slobin 1997; or computational linguists like Bod 1998. While not everyone would agree that these ideas can be grouped together or placed on this spectrum, the main point is to understand that for these approaches language is not unique but like any other cognitive process. Naturally, approaches exist that span the spectrum as a whole.

In their review of SLA assessment, Norris and Ortega (2003) enumerate various SLA theories like those above, emphasizing that the definition of L2 acquisition depends on theoretical assumptions. Generativist theories "view language as a symbolic system, autonomous from cognition, and too complex to be acquired from training or inductive or deductive learning from the input" (p. 725.) Thus, generativist theories fall on the side of the spectrum that is "for" a specialized language acquisition device. However, interactionalist theories "focus on the relationship between learner-internal and external processes in L2

acquisition" and fall in the middle of the spectrum (p. 724). Moreover, Norris and Ortega define generativist and interactionalist theories, at the time of their publication, as the mainstream of SLA research (2003).

However, they note the appearance of other theories such as emergentist and sociocultural theories. Emergentist theories view acquisition as "the byproduct of the [brain's] establishment of networked connections" which are based on patterns in linguistic input (p. 724). Thus, these theories fall on the side of the spectrum that is "against" a specialized language acquisition device. One the other hand, sociocultural theories may not fall within the spectrum at all because they view language acquisition as a social process, something that happens entirely outside the individual (p.724). It is important to note that not all theories can or should be placed on this spectrum. However, understanding the assumptions of the generativist, interactionalist, emergentists, and sociocultural approaches helps one understand how and why definitions of acquisition vary.

Norris and Ortega (2003) then go on to discuss how the assumptions of these theories change the definition of L2 acquisition. Acquisition in a generative theory includes correctly assessing grammaticality (p. 725) while acquisition in an emergentist theory is based on a subject's speed and accuracy in a confined testing environment (p. 728). Interactionalist theories define acquisition with a more functionalist view that includes acquisition as something "noticed" or where "acquired" means "understood with awareness" (p. 727). It is important to note that Norris and Ortega do not necessarily condone these differences; rather, they claim that constructs should be connected firmly in theory. They state, "Constructs should be defined in specific terms, such that observable behaviors may be linked with them, and they should provide a clear

indication of the theoretical assumptions that they represent" (p. 720). Thus, Norris and Ortega show that the definition of acquisition depends on the theoretical paradigm in which one operates.

The ACTFL Proficiency Guidelines

Acquisition is the process of acquiring language skills while proficiency is the product of that process. The ACTFL Proficiency Guidelines rely on criterion descriptions of each proficiency level. Some may take issue with the idea of using a definition of proficiency based on measurement criterion of proficiency and certification tests. This is especially true in the case of the ACTFL Proficiency Guidelines, which describe the various levels of criterion against which L2 learners are measured in tests such as the Oral Proficiency Interview (OPI). One could criticize tests like the OPI for their inability to adequately assess an individual's performance in reality, outside of a testing situation. Some authors (Douglas 1998, Vladman, 1988,and Clark and Clifford 1988) note that this may be due to the nature of the OPI test as a criterion-referenced test. A criterion-referenced test reflects something about the subject's ability if they reach a particular level of achievement on a continuum of pre-defined standards. This is in contrast to norm-referenced testing, which tests a subject's ability in a certain skill and then reflects the "relative" ability of that subject in comparison with a certain population of subjects (Glaser, 1963). One author, Douglas (1988), states that criterion-referenced testing "will necessarily be reductionist" (p. 251). While this accusation may be valid, it draws attention to the issue of differing paradigms.

Other scholars also criticize the OPI for its failure to test "real world" situations while at the same time noting its value as the best that is available. In the June 1988 issue of the periodical *Studies in Second Language Acquisition* (SSLA), which was a special issue devoted to "the assessment of foreign language oral proficiency" (Valdman, 1988), every article addresses

the inability of the OPI to test "real world" situations. Valdman, the editor of this special issue of SSLA claims:

> "This issue… is an attempt to bridge the gap between SLA and language proficiency testing …to acquaint SLA researchers with the testing instrument that is shaping FL teaching policy in the United States…[and] provide thoughtful critiques of the OPI and ACTFL guidelines and suggest alternatives for the assessment of oral proficiency" (1988, p. 122).

Thus, he recognizes the differing paradigms in which SLA research and OPI testing tend to operate, showing that neither side need throw out its assumptions completely. Rather, each should recognize the paradigm in which the other operates. In this manner, Clark and Clifford (1988) also note the discrepancy between oral proficiency testing and the "real world." However, they also state that the results of the OPI test are "by and large, substantially greater than that of more traditional measurement approaches" in terms of actual assessment (p. 142). Thus, one may establish the definition of proficiency by the criterion defined in the ACTFL Guidelines.

The criteria of the ACTFL Guidelines do not define proficiency in terms of a native ability. The ACTFL Proficiency Guidelines contain four separate parts: speaking, writing, listening and reading. Each part denote a scale of proficiency ranging through five different levels of distinguished, superior, advanced, intermediate, and novice, with all levels except distinguished and superior being subdivided into three additional categories of high, medium, and low ("ACTFL," 2012). It is important to note that even the distinguished level allows for "non-native accents, a lack of a native-like economy of expression, a limited control of deeply embedded cultural references, and an occasional isolated error" ("ACTFL," 2012). Thus, not even the distinguished level indicates a complete native-like proficiency.

According to the most recent publication of the ACTFL guidelines, the said guidelines are in part an adaptation of the Interagency Language Roundtable (ILR) guidelines, which scale proficiency from 0 (novice) to 5 (native-like), with the ACTFL superior level being equivalent to the ILR level 3 ("ACTFL," 2012). One may question whether the reason a superior level of proficiency only reaches the third level is because of the idea in SLA theory of maturational constraints, which revolve around the idea that biological factors, such as age, constrain one's ability to acquire language. The idea of maturational constraints is highly debated in SLA theory. Hyltenstam and Abrahamsson (2003) claim that current evidence supports at least some form of "a maturational constraints hypothesis" though they do mention that there are occasional exceptions to this idea" (p. 542). However, whether or not one can reach a native-like level is not as critical when measuring proficiency as the actual proficiency level achieved. It is important to note though, that proficiency for an L2 speaker must not necessarily reflect a native-like ability.

Understanding that one can be proficient without reflecting a native-like ability allows one to understand how the ACTFL guidelines can define a person who makes errors as "proficient." It is important to note that the superior level is used in this paper to represent a "proficient" speaker. The reason for this is that until the recent update of the guidelines, the superior level was the highest achievable ACTFL level and involves exceptional fluency. The distinguished level appears from its description to be for those persons who are highly "specialized" in language especially in terms of persuasion ("ACTFL," 2012). Indeed one may question how many native speakers would be able to attain the level of a distinguished speaker.

The following is a summary of a superior level speaker as characterized by the ACTFL Guidelines for speaking.

> A superior speaker's speech is characterized by ease and fluency, no patterned errors, the ability to communicate with concrete and abstract thoughts and defend and organize opinions, an ability to speak at length and in detail on a variety of topics even specialized topics, the use of interactive and discourse strategies, the use of syntactic and lexical devices, and the use of intonational features ("ACTFL," 2012).

The superior level in writing is characterized by similar descriptions but with additional requirements, such as "the use of writing protocols, especially those that differ from oral protocols, and an ability to write a variety of types of correspondence" ("ACTFL," 2001). Again, it is important to note that both the speaking and writing versions of the ACTFL guidelines allow for error. Superior speakers are expected to make "un-patterned errors, especially with low-frequency constructions, while superior writers may fail to conform to all of the cultural, organizational, syntactic, or stylistic patterns of a target language" ("ACTFL," 2009). However, these errors are not frequent and are without pattern.

A Criterion-Based Model for Translation Competence

As with the notion of "proficiency" in the field of language acquisition, the notion of translator competence also varies according to theory. Hague et al. (2011) synthesize the varying concepts of translator competence according to more functionalist theories in current literature. Their paper compares three criterion-based models of translation competence, specifically those of Albrecht Neubert (2000), the PACTE group (2000-2005), and Dorothy Kelly (2005). Each model describes translation competence in terms of "sub-competences," or specific skill sets that denote translation competence. The Neubert model delineates five levels of competence, while the PACTE and Kelly models delineate six and seven distinctions, respectively. Hague et al. (2011) provide a rough comparison of these models within a chart (p. 249). Each of the Neubert

and PACTE distinctions roughly correspond with five of the seven distinctions of the Kelly model.

The terms given to the distinctions of the Kelly model will be used in this paper; however, the description of each distinction will combine the corresponding descriptions from each model. It is important to note that in their inquiry, Hague et al. (2011) discuss translation competence as "performance… [in] producing a target text" (p. 244). With this in mind, Hague et al. (2011) report the following seven distinctions in translator competence: 1) communicative and textual skills including the ability to use grammar, lexical, and textual systems in both languages; 2) subject area competence including knowledge about the text topic; 3) cultural and intercultural competence including knowledge of the community of the text; 4) professional and instrumental competence in the translation field (meaning "knowledge about using necessary resources"); 5) strategic competence in solving problems (meaning "an ability to solve translation problems and make revisions"); 6) psycho-physiological or attitudinal competence (meaning "psychological factors like self-confidence and memory"); and 7) interpersonal competence including the ability to work effectively with others (pp. 248-250). As Hague et al. (2011) claim, these criteria show a trend in translation theory to look outside the text for factors that affect translation competence (p. 247).

One additional concept commonly mentioned in translation theory is congruity judgment. ASTM International uses the term congruity judgment to mean "the ability to choose an equivalent expression in the target language that both fully conveys and best matches the meaning intended in the source language for the audience and purpose of the translation" ("ASTM F2575," 2006). One may wonder where the idea of congruity judgment fits into this particular model. In their discussion of the Neubert, PACTE and Kelly models, Hague et al.

(2011) note that the fifth competence in the Kelly model, "strategic competence," is equivalent to the PACTE "strategic sub-competence" and Neubert's "transfer competence." However, they also note that these particular competences are each described as types of super-competences that "create links between different sub-competences as they control the translation process" (p. 247). Furthermore they note that Kelly's "strategic competence" includes both a linguistic approach to translation and competence beyond a linguistic approach that includes "factors that affect a translator's decision-making process and metacognitive ability to explain strategies and decisions" (p. 248). Congruity judgment falls in this category of strategic competence. Thus, congruity judgment is not forgotten in the Kelly model.

The Skill Set Involved in Translation Competence

A comparison of the criterion of the ACTFL Proficiency Guidelines for speaking and writing with the criterion of the models evaluated by Hague et al. (2011) shows some overlap, but more importantly it shows a distinction between a competent translator and a proficient speaker. Obviously, the first criterion for translation, the "communicative and textual skills," corresponds with most of the criterion laid out in both the speaking and written parts of the ACTFL Guidelines. Likewise, the second criterion for translation competence, "subject area competence," may also be compared with the requirement for a proficient speaker to be able to converse or write on a variety of topics, including specialized topics. However, the third criterion for translation competence, a "cultural and intercultural competence," is lacking in the ACTFL Guidelines at the superior level. In fact, the description in the ACTFL written criterion specifically states that a superior writer may not reflect the cultural patterns of a native writer ("ACTFL," 2001). However, the recent addition of the distinguished level does describe what may be termed "cultural and intercultural competence" as well. However, because the

distinguished level is highly specialized and little mention of cultural competence is made at the superior level one may question whether translation competence requires more cultural awareness than the average proficient L2 speaker who receives a superior level score. The last translation criterion, "intrapersonal competence," may appear to naturally be a part of a proficient L2 speaker's ability. However, as with the general population, the ability to speak proficiently does not mean that an L2 speaker possesses intrapersonal skills, which are described in the Kelly model as "the ability to work with different [and varied] people one encounters in the translation process" (Hague, 2011, p. 248). There are no corresponding criteria in the ACTFL Guidelines for the remainder of the translation criterion. Therefore, one may conclude that the translation criterion that do not have a corresponding criterion in the ACTFL Guidelines denote the skills that a competent translator must have beyond those of a proficient L2 speaker.

A distinction does exist between the skills required of a proficient L2 speaker and those required of a competent translator. Based on the criterion given in the current fields of language acquisition, testing, and translation, these skills include the following: a cultural and intercultural awareness that is better than the average "proficient" L2 speaker (meaning a superior level speaker), knowledge of the translation profession and resources, strategic competence in translation, which includes congruity judgment, psychological-physiological competence, and extended intrapersonal competence (Hague et al., 2011, 249). It is important to remember that the basis of these skills stems from the assumption that proficiency is defined by criterion-referenced evaluation. Other definitions of proficiency are not explored here. However, the skills noted above clearly show that translator competence includes a number of skills not necessary to be a proficient speaker.

It is interesting to note how these four translation competences compare to the ILR skill level descriptions for translation performance. The ILR denotes 5 levels of translation performance. The first two levels include minimal and limited performance. The third through fifth levels are all labeled "professional." However, the biggest difference between the levels of professional performance appears to be concerned with two of the four translation competences mentioned above: a cultural and intercultural awareness and strategic competence with includes congruity judgment. The ILR skill descriptions of translation performance at the professional level from level three to level five each include a statement that indicates cultural awareness is greater than the previous level. In fact the definition of being at a professional level is defined as "familiarity with the cultural context of both languages" ("ILR," n.d.). In addition, the professional levels also include many statements showing congruity judgment increases with advancement through each professional level such as "expression reflects native usage and consistent control of target language conventions" ("ILR," n.d.). Some evidence for the translation competence criterion "extended intrapersonal competence" is also found in the fifth level, which states that a translator has the "ability to finalize the product within time constraints and according to specifications" ("IRL Translation," n.d.). The fact that many of the translation competences discussed above are also found in the IRL skill descriptions for translation performance also helps support the idea that translation competence goes beyond language proficiency.

Quality of a Translator and Quality of a Translation Project

Second, just as translation competence and L2 proficiency are separate but related abilities, so too are the quality of a translator and the quality of a translation product. Namely, the quality of a translation product is only one aspect of the quality of a translator. This idea can

be seen by looking at the comparison of Hague et al. (2011) mentioned above concerning the three criterion-based models for translator competence. Each of the translation competence models, Neubert (2000), PACTE (2000-2008) and Kelly (2005), include some type of description of source and target language proficiency. Again, Hague et al. (2011) define the first competence as "communicative and textual skills including the ability to use grammar, lexical, and textual systems in both languages" (pp. 249). The fact that each model includes numerous aspects for judging the quality of a translator shows that L2 proficiency is only one part of being a proficient translator. In addition, the comparison of the ACTFL criterion and the summary criterion model for a proficient translator further show a total of five additional skills that a translator needs beyond proficiency in both languages.

A General Functionalist Approach

Finally, it is fundamental to understand that this research assumes a general functionalist approach to translation theory. This approach took hold in Germany in the middle of the latter half of the twentieth century and focuses on the purpose of a text, or *Skopos*, of a text (Schäffner, 1998). Nord, a well-known proponent of the functionalist approach, clarifies the relationship between translation theory and the concept of *Skopos* by stating that the methods and strategies a translator selects all depend on the purpose of a text for a specific audience (2006). This incorporation of purpose and audience found in the functionalist approach embraces two of the most recent and important theoretical shifts in translation theory:

> "The two most important shifts in theoretical developments in translation theory over the past two decades have been (1) the shift from source-text oriented theories to target-text oriented theories and (2) the shift to include cultural factors as well as linguistic elements

> in the translation models. Those advocating functionalist approaches have been pioneers in both areas" (Gentzler, 2001, pp. 70).

While the purpose of this paper is not to argue that a functionalist approach is better than other approaches, the fact that this approach has been able to incorporate these major theoretical shifts suggests at the very least it has a strong basis.

In the functionalist approach quality is judged by the fulfillment of project specifications, which are created based on the *Skopos* for a particular audience. After overviewing various arguments for the origin of the *Skopos* of a text, Gentzler (2001) surmises that "for all practical purposes, then, *Skopos* is not found but negotiated between the client and the translator, with reference to both the source text and the receiving audience" (73). Melby et al. (2005) refer to this negotiation process as defining project specifications and as a result even take the liberty of labeling the functionalist approach as a "specifications approach." They claim that in this theory, "… there is no one-best type of translation." Instead, a quality translation is one that conforms to the particular specifications established for the project at hand based on the audience and purpose" (pp. 405). Thus, specifications play an important role in quality assessment as defined by this functionalist approach because they are formed based on audience and purpose.

This idea of specifications goes hand in hand with the notion of the "translation brief." The idea of a "translation brief," has a solid foundation in the functionalist approach to translation theory. Nearly all translation models developed from a functionalist standpoint incorporate a "translation brief," or guidelines for a given translation project (Gentzler, 2001). While the *general* idea of a "translation brief" is widely accepted by those adhering to a functionalist approach, Hague et al. (2011) argue for the use of specifications created from a set of 21 standard parameters, which are derived from the Standard Guide for Quality Assurance in

Translation ("ASTM F2575," 2006). The values of these parameters are the specifications for a specific translation project. This argument for specifications goes beyond the simple notion of translation brief and seeks to establish a constant standard from which all specifications for any translation project can be derived. A recent publication of the International Organization of Standards (IS0) also includes this same set of 21 parameters, such as complexity and obstacles, register, and production tasks. This shows this in-depth version of the "translation brief" is becoming more recognized ("ISO," 2012). For these reasons, this project will incorporate the use of a "translation brief" in the form of specifications derived from this standard set of parameters.

In a theory describing definitions of quality found in the business world, a functionalist approach that includes specifications can be described as a manufacturing based approach to quality. Among others, Russell (1998) uses Garvin's theoretical framework to classify three approaches to defining quality: transcendent, where quality is a theoretical and absolute notion; user-based, where it possess the ability to fulfill needs; and manufacturing-based, where quality is based on the fulfillment of specifications. While Russell (1998) does discuss other approaches, in an interview, Melby (2012) clarified that three of these approaches, transcendent, user-based, and manufacturing-based, are well-known in discussions of translation quality. It is the manufacturing-based definition of quality that best complies with the functionalist-based approach. The manufacturing-based definition to quality is really a business-world definition. It states, "quality pertains to a product's degree or conformance to engineering and design specifications" (Russell, 1998, p. 14). However, the definition of translation quality based on specifications by Melby et al. (2005) mentioned above complies with it well.

It is important to note that Russell's description of Garvin's framework differs somewhat from other authors concerning the definition and assessment or translation quality. For example, House (2001) outlines three major categories: mentalist approaches, text-based approaches and response-based approaches. Her description of the mentalist approaches as assigning quality based on global judgments because this theory holds that "translation is an act that depends solely on the artistry and skill of the translator" (224) is similar with Garvin's transcendental category as a theoretical notion. The other two categories do not seem to correlate as well to Garvin's more business framework for quality. Specifically, House (2001) places functionalism approaches in the category of response-based approaches along with behavioristic approaches. She claims that "the notion of 'function' is never made explicit or operationalized in any satisfactory way" (245). However, the explanation of Melby et al. (2005) of a functionalist approach as a specifications approach does allow one to operationalize quality by measuring whether or not agreed upon specifications are met. Perhaps a specifications approach to functionalism strengthens this approach.

Colina (2009) also sets up a framework of approaches to quality, only she puts functionalism in its own category. The other categories include experiential, theoretical, reader-response and text and pragmatic. Interestingly, she defends functionalism as an approach to quality assessment claiming that it can "achieve middle ground between theory and applications" by addressing approaches simultaneously (p. 239) using a client-defined definition of quality. Specifically, her assessment tool allows for clients to choose which aspects of quality they want to emphasize. In this way Colina (2009) seems to fall both under Garvin's user-based and a specifications approach. However, the fact remains that having a client choose which aspects of quality to emphasize at least is similar to the notion of a client and translator negotiating *Skopos*

as Gentzler (2001) claims or negotiating specifications as Melby et al. (2005) describes. In addition, Dunne (2011) similarly argues for client involvement in a translation project as a means of assessing quality. In fact, Dunn even uses the term "client specifications" and refers to meeting them as a way of assessing quality. Suggesting that in the real world, more and more people are recognizing the specifications approach as a model that can be operationalized.

In summary, this research will assume a functionalist-based approach to defining and assessing quality. Translation quality assessment is viable subject of study and does not fall under L2 proficiency assessment because skills involved in translation competence incorporate and go beyond proficiency skills. Instead of relying on a framework of translation quality that is more oriented to translation theory, this paper relies on Russell's description of Garvin's theoretical framework, which is a more business-oriented framework of quality assessment as restructured by Melby (2012) who claims only three categories: transcendental, user-based and manufacturing. Specifically, a specifications approach falls under the manufacturing classification. However, Melby (2012) has developed a more encompassing definition for translation quality based on a functionalist specifications approach that gives credence to both transcendental and user-based ideas. The definition states: "a quality translation achieves sufficient accuracy and fluency for the audience and purpose, while, in addition, meeting all other negotiated specifications that are appropriate to end-user needs." It is this definition of quality that will be adhered to in this research as it looks to determine the quality of a translation product, which in turn reflects a part, but not all of a translator's translation competence.

DESIGN

The methods used for data collection involved in this research included four phases: test design, test administration, scoring and analysis. The online testing tool and scoring algorithm used by Hague et al. (2012) comprises the general test design. However, many steps were involved in the creation of an Arabic-language specific version of the general testing tool including choosing appropriate level texts, creating model and faulty translations, and writing a translation brief of specifications for each of the two passages chosen. The test administration phase included recruiting subjects and the logistics of administering the test. The third phase, scoring, occurred in two parts: hand scoring both versions of the test and retrieving the automated scores provided by the algorithm designed for the ongoing project of Hague et al. (2012). Specifically, scores from each style of test were gathered for each subject that participated. Finally, various methods of analysis were applied to the data, including descriptive statistics, in hopes of determining whether a high score on one test predicted a high score on the other test.

General Test Design

While the testing tool and scoring algorithm used by Hague et al. (2012) does comprise the majority of the Arabic-language test, it is important to note the differences between my test design and that of the aforementioned research. The overall test design of Hague et al. (2012) consists of a five stage process: (1) a language background survey, (2) an L2 reading comprehension test, (3) an English writing proficiency test, (4) a traditional full translation test, and (5) a detect-and-correct style translation test. The second and third items were included in the Hague et al. process as a way to look for connections between language proficiency and translation competence. My research is limited in that it only includes the first, fourth and fifth

stages of the Hague et al. process. This project did not seek to replicate all five stages due to limited funding. When Hague et al. (2012) administered their French and Spanish tests, each subject was paid sixty dollars for participating in the approximately four-hour long test. It was feared that without monetary compensation, including the second and third stages would have made the length of my test too long and an already limited number of Arabic speakers would not be willing to participate.

The first, fourth and fifth stages of the Hague et al. (2012) process were included in the present research because they encompass the crucial stages of the overall testing process. The first stage was completely replicated in the current research in that the same language background questionnaire via the same password protected website used by Hague et al. (2012) was administered to the subjects before taking the translation tests. The fourth and fifth stages of their research were also replicated, but with differences. The most notable difference was that this research produced Arabic versions of the full translation and detect-and-correct translation sections of the test. Thus, based firmly on the test design of Hague et al. (2012), this research includes a three stage testing process: (1) a language background questionnaire, (2) a traditional full translation test, and (3) a detect-and-correct style translation test. The design of the second and third stages of this test was performed using the online CTT testing suite created for the aforementioned research of Hague et al. (2012). This CTT testing suite includes means to create and administer both a traditional full translation and a detect-and-correct style translation test in a given language.

The final products of the traditional full translation and detect-and-correct style tests include various parts. First, both tests include a copy of the source text, a translation brief, and a glossary of uncommon terms (see Appendix). The interface design of the traditional full

translation test includes a free response space to type the target text as seen in Image A. This test is graded by hand. However, the detect-and-correct style test already includes a target text translation into which intentional errors have been introduced. The subject's goal is to fix all the errors correctly and leave the parts of the text without errors alone. When a subject identifies an error in a particular part of the text, he or she clicks on the text and a box appears in which he or she types a correct translation as seen in Image B. This test is scored by an automated algorithm.

Image A- Full Translation Test

Image B- Detect-and-correct Test

Test Design for Arabic Version

Numerous steps had to be taken to create Arabic-language specific versions of the full translation and detect-and-correct style tests. The first step included choosing the source texts to be used. The CTT testing suite created for the research of Hague et al. (2012) requires the input of two L2 texts to serve as source texts. Two Arabic texts were chosen and an approximately 175 word passage from each text was used as the source text (see Appendix). One source text deals with poverty and the West's role in it while the other deals with the 2011 Egyptian revolution in the context of the Middle East. Consequently, each will hereafter be referred to as "Povtest" and "Revtest" respectively.

The reasoning behind selecting each source text was founded in research. First, the researcher determined to follow the pattern of Hague et al. (2012) in using a source text that would not require too much specialized knowledge of the source language culture or another specialized topic. Second the research chose to follow the pattern of Hague et al. (2012) in

selecting texts at an ILR Level 3. ILR Level 3 is a classification of the Interagency Language Roundtable that is approximately equivalent to the ACTFL superior level ("ACTFL," 2009). It includes texts that go beyond reporting information, like news articles, and includes texts that among other things use language to describe an opinion ("LangNet," 2006). They chose this standard because it is the minimum text level used by the ATA in translation tests for any given language pair (Koby and Champe, in press). To ensure that two appropriate ILR Level 3 texts were chosen, a native English speaker and a native Arabic speaker participated in extensive training on how to rate texts using the five point scale of the Interagency Language Roundtable. Afterwards they rated numerous Arabic texts before selecting the Arabic source texts used in this research.

The second step that had to be taken was to create Arabic-language specific versions of the tests, which included writing a translation brief (specifications) for the translation project. In order to keep the two versions of the test as similar as possible, and because the goal of both tests were similar, a single translation brief (see Appendix) was written for both Arabic source texts and used for both the full translation and the detect-and-correct tests. The brief was written by choosing from the set of 21 parameters described by Hague et al. (2011) and found in the recent ISO document on translation guidelines ("ISO," 2012). This brief included 8 of the 21 parameters including: audience, volume to be translated, complexity and obstacles, languages and regions, content correspondence, usage/register, directions concerning reference materials and technology. Special emphasis was given to the parameters "complexity and obstacles, content correspondence and usage/register" in the hopes that the subjects would be consistent in their translations and graders could be consistent in their grading. In addition, two other

specifications: terms of delivery deadline and compensation were not included in the translation brief, but explained in the recruitment process.

The third step in creating Arab-language specific versions of the tests was creating a model translation. For each source text passage, a model translation (see Appendix) was created by a team consisting of three native Arabic speakers and three native English speakers. Two of the native speakers were post-secondary students and the third was a graduate student working towards a master's project. Two of the natives were male the other female. The native English speakers consisted of the researcher, a graduate student studying Arabic, and two seasoned Arabic professors. The process of creating the translation was as follows. The researcher and one of the native Arabic speakers created a translation draft according to the guidelines provided in the translation brief discussed above. This draft of the model translation was subsequently looked over by the other native Arabic and English speakers. Any differences in opinion were settled by looking at the specifications.

The fourth step involved creating a faulty text for the detect-and-correct style test. This faulty translation text was created by taking the model translation and introducing between ten to fifteen errors into the text. This is similar to the practices used by Hague et al. (2012). All errors were local errors and not global. This is because the technology of the testing suite is not currently able to handle global errors. Note that global errors cross sentence and paragraph boundaries, local errors do not. Errors were created by a team of two native English speakers: the researcher (a graduate student studying Arabic) and a seasoned Arabic teacher. Because the specifications of the translation required that the text "sound like an article originally written in English, free of cultural, lexical and grammatical errors," all errors introduced not only had to be errors that a translator would actually make, but also had to not be too obviously in violation of

the requirement to read "like an article originally written in English." Error types included lexical and grammatical errors. Lexical examples include mistranslations of words such as translating the word سياسات as *politics* instead of *policies*. Grammatical examples include errors such as flipping the subject and object in a short clause. (See Appendix for more sample errors and responses.) The final versions of the faulty translations used in the detect-and-correct portions of the tests included thirteen errors for the "Povtest" and thirteen errors for the "Revtest."

Test Administration

The administration phase of the project included recruiting subjects and managing the logistics of test administration. The recruiting of subjects was done by word of mouth. Finding a subject population was difficult due to the low number of student retention in advanced Arabic classes. BYU was seen as a good location to find Arabic-speaking students because of the extent of the Arabic Language Program. All but one of the students took an advanced-level online reading proficiency test belonging to the National Middle East Language Resource Center that is being developed in connection with the American Council on the Teaching of Foreign Languages (ACTFL). The average score of the students was 73.04% and the median was 79%. A score of 80% or higher represents an advanced-mid score according to the ACTFL guidelines and a 67% or higher represents an advanced low score. Thus, on average the subjects could read at an advanced level.

There are limitations to these reading scores in that they may not accurately measure the ability of the students. Specifically, students may not have tried their best because the scores were not factored into their grade. In addition, many of the students who had studied Arabic aggressively quit studying after the study abroad and the tests were conducted approximately seven months or seventeen months after the 2009 and 2010 Arabic Study Abroad programs. The

students who participated in this study attended either the BYU Egypt Study Abroad during the summer of 2010 or the BYU Jordan Study Abroad in the summer of 2009. The majority of the students attended the former. Therefore, most of the subject population had taken a reading proficiency exam and more than half received an advanced-mid rating despite limitations. This population was chosen because it was the largest one known to the researcher with these proficiency requirements.

Once a student had been recruited, a time was arranged for the student to take the test. Test administration included many policies. First, a student was allowed to take the test on campus or off campus, at the convenience of the student so that more students would be willing to participate. This procedure helped boost the number of students that participated. Second, students were allowed to use a paper copy of an Arabic-English dictionary, but were not allowed access to online resources. All students chose to use the Hans Wehr Dictionary of Modern Arabic. This is in congruence with the protocol of both the translation tests administered by Hague et al. (2012) and the translation tests administered by the ATA, who allow a test-taker to use general paper dictionaries ("ATA Certification," n.d.). In addition, translators in the real world use dictionaries when working on projects. This policy allowed for a more authentic atmosphere with the test. Third, all subjects were proctored while taking the test, including those who took the test at off-campus locations, ensuring that they followed directions indicated in the translation brief.

Another important policy to note is that subjects were allowed to take as long as needed to finish the test. The subjects were told during recruiting that the test would take approximately two hours. This time frame was based on the fact that subjects that participated in the Hague et al. (2012) tests were allotted only twenty minutes for each section of the test, which included

passages of about 100 words, totaling 60 minutes. Hague et al. (2012) suggested that the algorithm they use would work better if a longer passage were used allowing for a greater number of "chunks" in a text. Subsequently, this research used passages of about double the number of words. Therefore, it was expected that it would take approximately double the amount of time to complete. However, unlike Hague et al. (2012) students were allowed to take longer than the projected 120 minutes. In fact most students took approximately 2.5 hours. Students were not given a time limit to ensure that a good number if not all students would actually finish both sections of the test. This is a weakness in this research because while most students took about two and a half hours, there were a few who took more, which may have given them an advantage.

The last important policy to note is how it was decided which version of the test each subject took. As described above, the testing suite requires two source texts which in turn allows for the creation of two different versions of the text used in this research. In the first version, a student takes a detect-and-correct test created from the faulty translation of the "Povtest" and then takes a full translation test involving the "Revtest." Version II of the test switches the roles of the "Povtest" and the "Revtest." In order to have approximately the same number of students taking each test, every other student was given version I of the test, while the others were given version II. In the end, twenty-six students took a test, with thirteen students taking the "Povtest" and thirteen taking the "Revtest."

Scoring

The scoring phase involved hand scoring both the full translation and detect-and-correct versions of the test in addition to retrieving the automated scores provided by the algorithm designed for the ongoing project of Hague et al. (2012). A total of twenty-six subjects took the

test. Two subjects' responses had to be discarded. One subject's response was discarded because that person left early and only finished about three lines of the full translation section of the test. All other subjects at least attempted to get through the entire test. The second subject's response was thrown out because that subject was recruited to be a grader for this research due to a lack of human resources with the ability to score the full translation responses. The full translation tests were hand scored by two graders using a modified version of ATA grading methodology for grading certification exams. This modified ATA process involved a brief grader training (reading ATA grader training texts), the creation of a model translation, creation of a document of projected errors (see Appendix) and finally grading by two individuals. The model translation previously created for the testing suite was used in the grading process by the two graders. In addition to the model translation, both graders read the ATA grader training texts and participated in the creation of a document of projected errors before grading.

The document containing passage-specific guidelines concerning projected errors was created in two steps. First, the two graders briefly discussed projected errors, or errors that a translator is likely to make given the source text. Then, they individually graded the full translation responses of two subjects that took the "Povtest" and two subjects that took the "Revtest." After that, the graders came together again to discuss projected errors and how certain types of errors should be interpreted according to the ATA grader training texts. Two different documents of passage-specific guidelines were created, one for each source passage. About half of each document consists of general guidelines dealing with how to score varying degrees of errors such as omission or subject-object confusion, etc. The rest of each document consisted of guidelines for scoring specific errors in certain chunks of the text that are projected to be difficult. After creating passage-specific guidelines, the graders scored the tests. (See

Appendix for scoring samples from each grader.) It is important to note two logistical weaknesses. Originally one of the graders rescored those four responses used to create the passage specific guidelines after the said guidelines were created. The other grader did not re-grade the four test responses according to the passage specific guidelines at first. However, this logistical error was discovered during the analysis phase and was corrected by having the second grader rescore those four responses according to the passage specific guidelines.

Second, the original passage-specific guidelines did not specify whether a sentence boundary was to be counted according to the Arabic source text or the English target text, which could vary. In addition whether one should cap the sentences at sixteen error points as the ATA grading methodology suggests or continue to score beyond sixteen error points per sentence was confused. This caused issues with inter-rater reliability between the score of the full translations. An attempt was made to correct this error by re-tallying the scores of the grades in which sentence boundaries were clearly marked in the source text and where any sentence over sixteen errors points was capped at sixteen. Of course, this does not change the fact that the mentality of the graders while grading was different, meaning that one grader went into the task thinking, "I need to cap sentences at sixteen points," which may have caused the person to score errors in larger chunks. In addition, following the ATA Methodology, if the scores between the two raters were drastically different, in this case more than a twenty-five point difference on the weighted scale, then the graders came together to discuss any major discrepancies in grading particular parts of that passage. If an agreement was reached, all other responses where then checked for a similar mistake so that each grader would remain consistent in how he or she graded. The scoring samples included in the appendix reflect these practices.

The detect-and-correct section of the test was scored using an algorithm developed for the research of Hague et al. (2012). This algorithm is based on a system of chunking. When a subject begins the detect-and-correct test the system divides the intentionally error-filled target text into chunks. The subject's goal is to fix all the faulty error chunks correctly and leave the non-error chunks alone. Each chunk is labeled in the computer system as an "error chunk" or a "non-error chunk." An "error chunk" contains an error introduced into the system by a test developer who creates a "faulty passage" full of parsed errors. Parsed errors are errors that are tagged in the system. Each one is connected to a list of possible correct answers that is stored in the system. The system records which chunks a subject does and does not fix, whether the act of fixing the chunk was an error, and whether chunks with errors are fixed correctly. The automated grading system uses this information to score each passage once submitted.

This algorithm produces five numeric scores. The first score entitled "sensitivity" gives a score for how well a subject can identify errors. This score is calculated by adding the number of error chunks in a text that a subject fixed correctly to the number of error chunks that a subject fixed incorrectly and dividing that number by the total number of chunks in the text with errors. This gives the total percentage of how many error chunks were changed at all. The second score called "specificity" provides a score for how well a subject is at detecting chunks without error. It is calculated by taking the number of non-error chunks the subject did not fix and dividing that by the total number of non-error chunks. This gives a percentage for how many non-error chunks the user did not change. The third score entitled "diagnostic skill" gives a percentage for how well a subject is at identifying both errors and non-errors. It essentially combines the sensitivity and specificity scores and shows the total percentage of non-error chunks left alone plus error chunks that were changed, even if they were not correct. The fourth skill is the

"prescriptive skill." It gives the percentage for how well a student fixes the errors (that they are able to identify) correctly. It is calculated by taking the number of error chunks a subject fixed correctly and dividing that by the total number of error chunks that a subject attempted to correct. The final encompassing score provided by this algorithm is the "translation rating" score. It is calculated by adding the number of error chunks fixed correctly to the number of non-error chunks left alone and then dividing that number by the total number of chunks in the passage. This shows which subjects can fix the errors that they are supposed to, while leaving the remainder of the text alone because it does not have errors.

This algorithm has certain weaknesses. One is that when a subject detects an error chunk and changes it, if he or she does not provide an answer that is included on a list of acceptable answers entered into the testing suite, he or she will automatically be scored down for that correction. Hague et al. (2012) argue that while this a weakness, they believe this will not be a problem once a test has had enough subjects take it. Each time a subject takes a test, a human intervenes and judges whether an error chunk that was counted wrong could be correct according to the specifications. If a subject were correct, his or her response would then be added to the list, and at some point the list of acceptable answers would steady and the test would be able to run without human intervention, hypothetically. While this is only one scenario for how to solve this issue, it is how this issue was dealt with in this project. Each error chunk fixed incorrectly was analyzed by a human to judge whether the response needs to be added to the list of acceptable answers and thereafter the test was rescored to reflect these changes. A second weakness is that the scoring algorithm does not have the means to account for unexpected changes to non-error chunks. This means that human intervention would need to occur in order to judge whether

changes to non-error chunks are harmful to the translation product or not. However, fixing this weakness in the algorithm is beyond the scope of this research.

Analysis

The final phase of this research involved various types of analysis on all the data collected. This analysis included the use of correlations and Bland-Altman plots of agreement in order to analyze inter-grader reliability. It also included descriptive statistics. All were useful in analyzing the original research question of whether a good score on the detect-and-correct version of the test is a decent indicator of a good score on the full translation version.

RESULTS AND DISCUSSION

Three stages of analysis were performed on the data collected. First, the five numeric scores automatically calculated for the detect-and-correct type tests were analyzed using descriptive statistics. Second, the hand-graded scores of both graders for the full translation test responses were analyzed using a Pearson correlation and a Bland-Altman Plot of Agreement to determine inter-rater reliability. Finally, a comparison was made between all the scores of the detect-and-correct tests and the scores of the full translation test using a Person correlation.

In answer to the original research question, the results do not strongly suggest that a high score on the detect-and-correct portion of the test most likely indicates a high score on the hand-graded full translation test for this particular subject population. However, this research has still provided great insight into translation testing, especially concerning whether the detect-and-correct portion of the test actually is measuring translation competence, concerning SLA programs and their intentions, and concerning logistical issues in testing including the impact text difficulty and length had on the detect-and-correct testing as well as the negative impact the ATA grading practices of weighting errors and capping errors can have on an experiment such as the one described in this research. These insights were discovered in various stages of the analysis, a description of which follows.

Before describing the various stages of analysis, it is important to note that the results reveal no difference between the "Povtest" version of the test and the "Revtest" version of the test in either the detect-and-correct or the full translation portions. This can be seen by the even distribution of both the circle and diamond data points in Figures 1-12, which represent a "Povtest" or "Revtest" response, respectively. As a reminder, a "Revtest" version of the test uses a passage about the Egyptian revolution in a given task (i.e. detect-and-correct or full translation

task) while a “Povtest” version of the test uses a passage about poverty. Because there were no apparent distinctions found in using different passages, this research combines the results of both the “Povtest” and “Revtest” versions when analyzing.

The Five Detect-and-Correct Scores

In the first stage, a descriptive analysis was performed on the five scores of the detect-and-correct tests. This analysis generally indicates that the particular subject population used in this research is better at identifying errors than correcting them. This means that students could tell something was wrong with the faulty translation but could not fix it, as can be seen in Figure 1 and in a later figure, Chart A, which shows among other things the r-values and their significance for the correlations between the automated scores. The overall numbers for the diagnostic skill, or identifying chunks of text with both errors and non-errors, is much larger than the overall numbers of the prescriptive skill, which involves correcting errors correctly. In fact, Figure 1 shows two important revelations that further confirm this statement. First, the majority of the prescriptive scores were below the line of equality, which means that more than half of the time the subjects are performing worse at correcting errors. Note that the line of equality in all figures 1 to 6 shows the points at which the subjects would have scored the same in each of the skill sets placed on the x and y axes, i.e. a point located on the line in Figure 1 would show equal scores for the prescriptive and diagnostic skills. Second, when looking at this graph, one automatically notices the shear number of subjects who received a zero for the percentage of time he or she fixed errors correctly. Both of these revelations indicate the population more often than not knew something was wrong with the text but could not fix what was wrong.

When looking at Figure 1, one notices two apparent outliers who appear to have very good prescriptive ability. These two points are above the line of equality. It is important to note

again exactly what the prescriptive ability describes. Prescriptive ability calculated by the algorithm for Hague et al. (2012) shows the percentage of errors fixed correctly divided by the number of errors that one attempted to correct. This means that one could receive a perfect score for prescriptive ability without detecting all thirteen errors in the detect-and-correct portion of the test. In fact, as seen in Figure 1, one subject did receive a perfect score for prescriptive ability. However, it was because he or she fixed the three errors chunks he identified correctly. Interestingly, the second data point that is above the line of equality represents a subject who fixed five of the six error chunks he identified correctly. This shows that prescriptive ability, when looked at individually can be misleading. Figure 2 shows a new calculation of prescriptive ability recalculated as the percentage of error chunks fixed correctly over total error chunks. This new calculation is plotted against diagnostic ability showing, as Figure 1 does that diagnostic skills were much better than prescriptive skills.

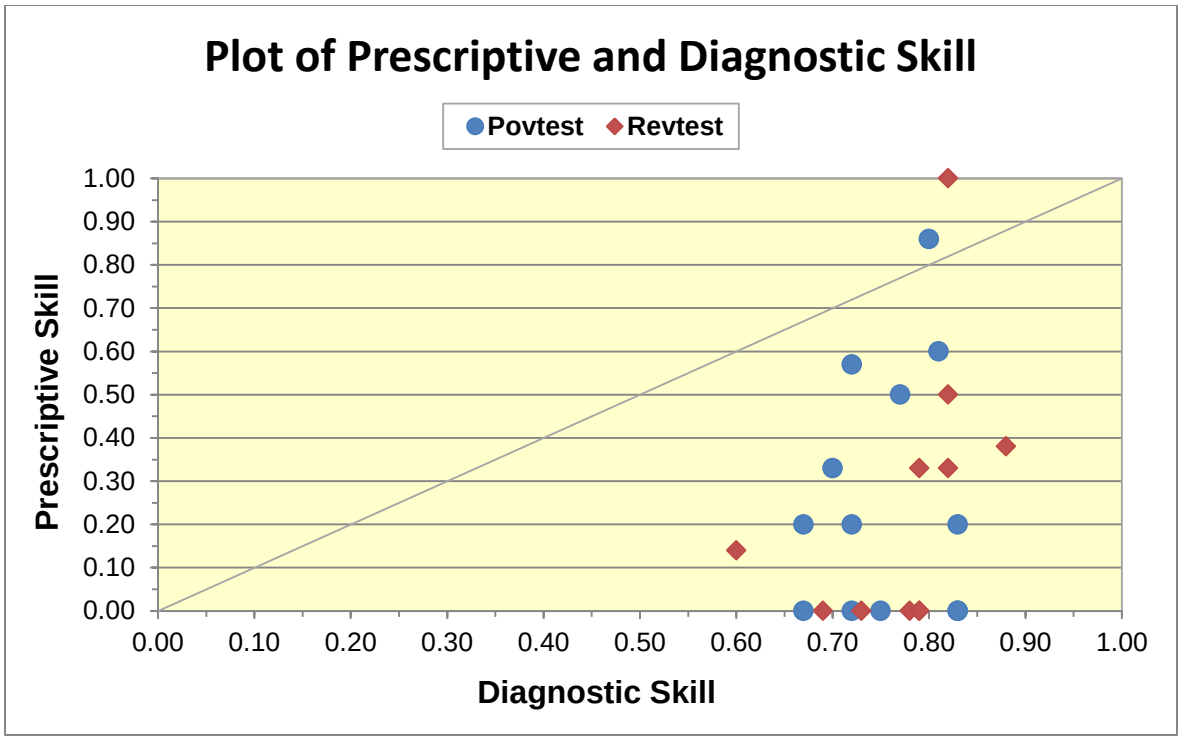


Figure 1

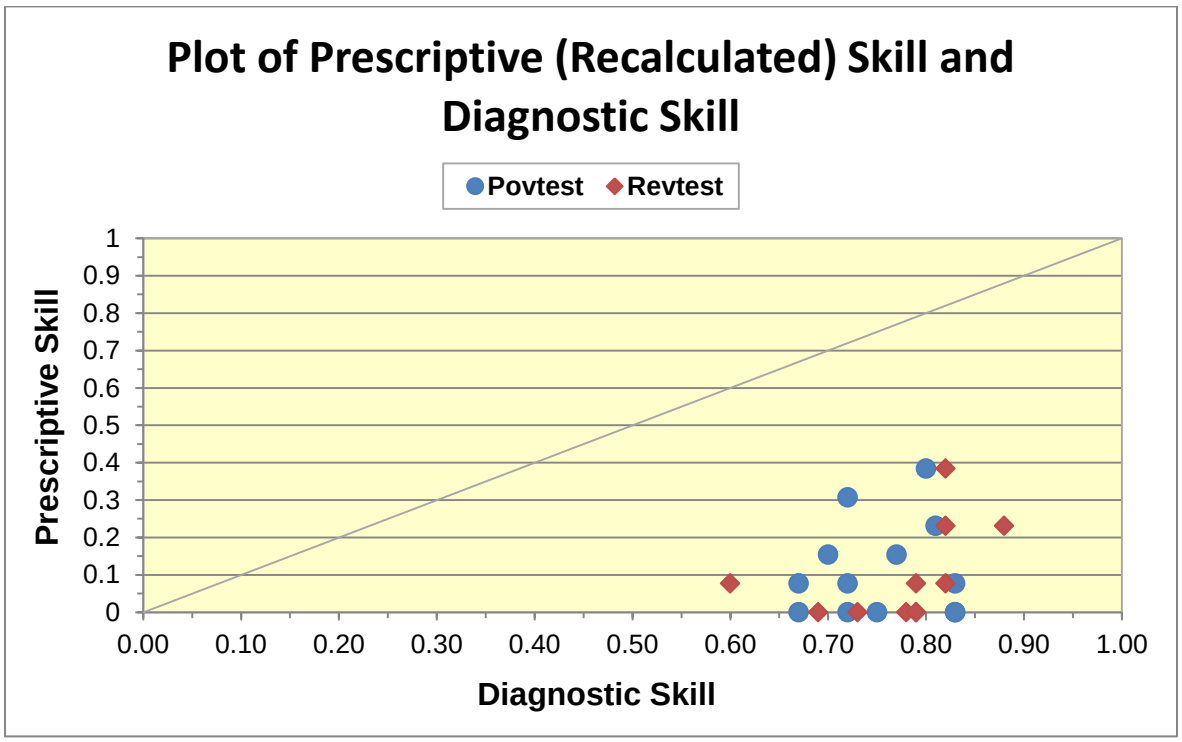


Figure 2

Naturally, one must question the reasons for why subjects had higher scores when identifying errors rather than correcting them. One reason may simply be that these results actually reflect the ability of this particular subject population. Perhaps the prescriptive skills of the majority of this population really are poor. Another possible reason may question whether the training of these students specifically prepared them to identify rather than correct errors. All of the subjects tested were students at BYU or students that had recently graduated and were still around campus. Perhaps something innate in the instructional process of language acquisition prepared students to recognize errors but not fix them. Furthermore, perhaps the ability to fix errors correctly is the mark of an advanced translator that comes with experience and current or recently graduated students simply do not have the experience necessary to perform well on a prescriptive task.

This notion that language proficiency skills and translation competence are different but related abilities is found in the literature and seems to apply here. As mentioned in a previous section, if one takes the synthesized set of skills that denote translation competence noted by Hague et al. (2012) and then only look at those skills in the set that do not have a correspondence with ACTFL proficiency skills, one is left with the skills that a translator requires that goes beyond proficiency skills. These skills include: (1) a cultural and intercultural awareness, (2) knowledge of the translation profession and resources, (3) strategic competence in translation, (4) psychological-physiological competence, and (5) extended intrapersonal competence (Hague et al., 2011, 249). One might conclude that the subjects tested in this research have not yet developed these abilities. Indeed the ability to correctly fix errors could go beyond proficiency and involve at least cultural and intercultural awareness as well as strategic competence in translation, if not knowledge of the translation profession and resources as well. The language

programs in which these students participated at least at BYU focused mostly on language proficiency, as most programs do. Any translation exercises were used for the purpose of increasing language proficiency not in order to improve translation competence. Therefore, subjects may lack enough experience with the skills that go beyond language proficiency, which may have caused them to have low prescriptive ability.

Indeed the idea that translation competence is developed through experience is not new. Castellano, as quoted in Baker (1992) states that, "Our profession is based on knowledge and experience. It has the longest apprenticeship of any profession. Not until thirty do you start to be a useful translator, not until fifty do you start to be in your prime" (p.3). Thus, one may hypothesize that language proficiency training inherently grants a subject the ability to identify errors, a diagnostic ability, but something more is required through specialized translation training or sheer experience that allows a subject the ability to correct errors, a prescriptive ability.

A third explanation for the diagnostic ability of subjects being higher than the prescriptive ability may be due to a design flaw in the test. One flaw may be that the length of the text and the number of chunks into which it is divided may not be enough to give the subject a fair chance to demonstrate his or her prescriptive abilities. Numbers are important in statistics and the diagnostic ability has been analyzed using descriptive statistics. Larger numbers may make the results more reliable. In addition, if there are too many non-error chunks or too many error chunks a student may be swayed to think that the majority of the test falls one way or the other rather than relying on his or her skills to identify and correct errors.

Another possible design flaw that may have influenced the results could be the difficulty of the text. Extensive measures were taken to ensure that both texts chosen would be at an ILR

Level 3 as is customary in ATA testing procedure as outlined in the project design. However, for the subject population in this research, this level may have been too difficult. The majority of the subjects had only studied Arabic at BYU and on a BYU Arabic study abroad. They had all taken 300-level classes and the majority had tested at a reading level of advanced-mid to advanced-high on the ACTFL proficiency scale. While some ILR Level 3 texts were a part of the curriculum, this level of texts was not assigned very often and was meant to push students out of their comfort zone. One must therefore hypothesize that perhaps the proficiency level of the subject population was lower than needed to translate ILR Level 3 texts and that if proficiency were higher the subjects could have performed better in both prescriptive and descriptive abilities.

This hypothesis contradicts the possibility explored above that language proficiency ability grants one the ability to identify errors while extra training or experience allows one to actually correct errors correctly in that a higher language proficiency may have allowed subject to perform better at correcting errors. However, because researchers like Hague et al. (2011), Neubert (2000), Kelly (2005) and the PACTE group (2000-2005) all define competencies of a translator that go beyond language proficiency ability, one assumes that even if this group had the highest language proficiency they still may not have performed well due to a lack of special training or experience to develop the skills that go beyond proficiency. Obviously further studies with varying populations could explore this possibility.

Besides diagnostic and prescriptive skills, other components of the diagnostic and prescriptive skills such as sensitivity and specificity were also analyzed in order to find additional insight on the subject population and the detect-and-correct test itself. Specificity is the ability to identify non-error chunks while sensitivity is the ability to identify error chunks. It

is important to note again that the diagnostic skill mentioned above is derived from the combination of the ability to identify errors and the ability to identify non-errors. This particular population group was better at identifying non-errors (specificity) rather than identifying errors (sensitivity) as can be seen in Figure 3. In fact all subjects scores for specificity fall below the line of equality, which means that all subjects performed better at identifying non-errors than errors. In addition, all but one of the specificity scores is above seventy percent. This means that subjects, overall, were very good at leaving chunks without any errors alone but they were not as good at identifying chunks with errors. The sensitivity scores were low, and by extension, this also means that the prescriptive skill scores are also rather low because one needs to be able to identify errors before one can fix them. This can be seen in Figure 4. Therefore, one can see that overall this subject population did not perform well on this test, which fits in line with the hypotheses above that the proficiency skills and/or level of translation competence of this subject population may not have been advanced enough for the ILR Level 3 texts used in this test.

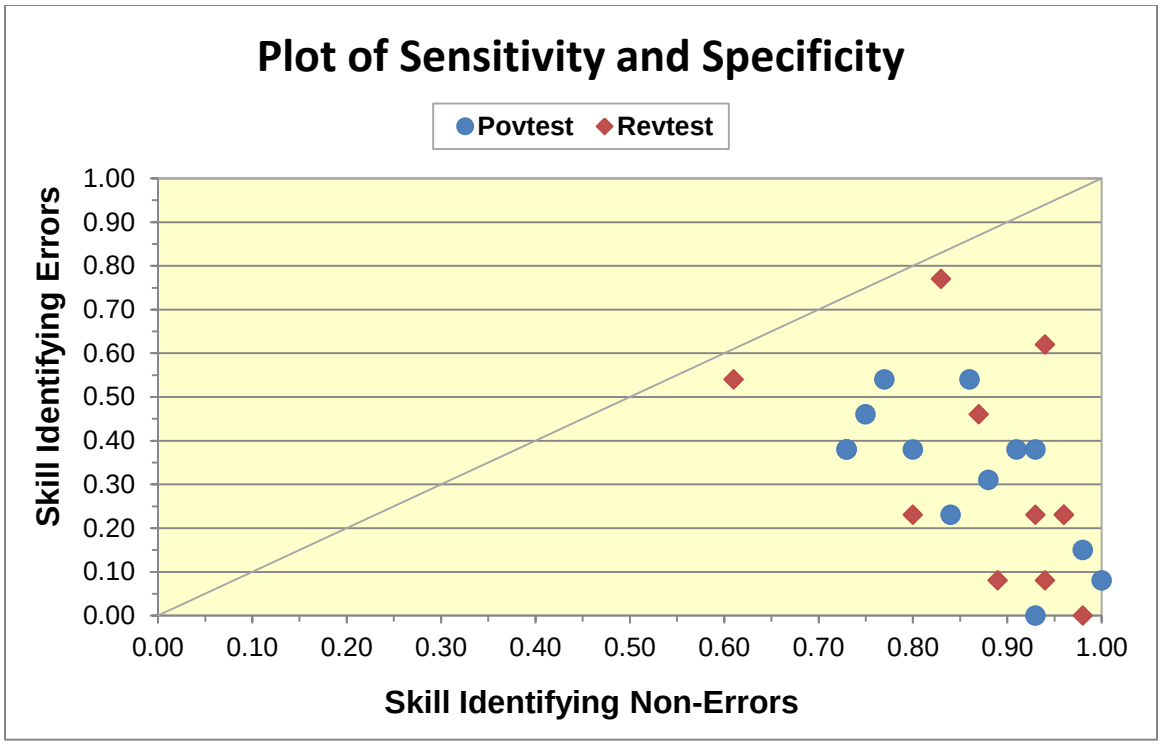


Figure 3

In addition, one must also wonder whether the higher specificity scores combined with a large number of zero percentages in the prescriptive ability is due to a large number of students submitting the test without making any changes. While none of the subjects did this, the test responses do show that three subjects only changed one thing in the entire text. In addition, five other subjects only changed between three and six chunks. Each version of the detect-and-correct test included thirteen error chunks. This means that a total of eight subjects changed less than half the number of chunks with errors, which may have influenced the number of zero percentages on the prescriptive scores. However, that could also be accounted for by subjects who "corrected" a lot of both error and error free chunks but did not correct the error chunks correctly or missed them entirely.

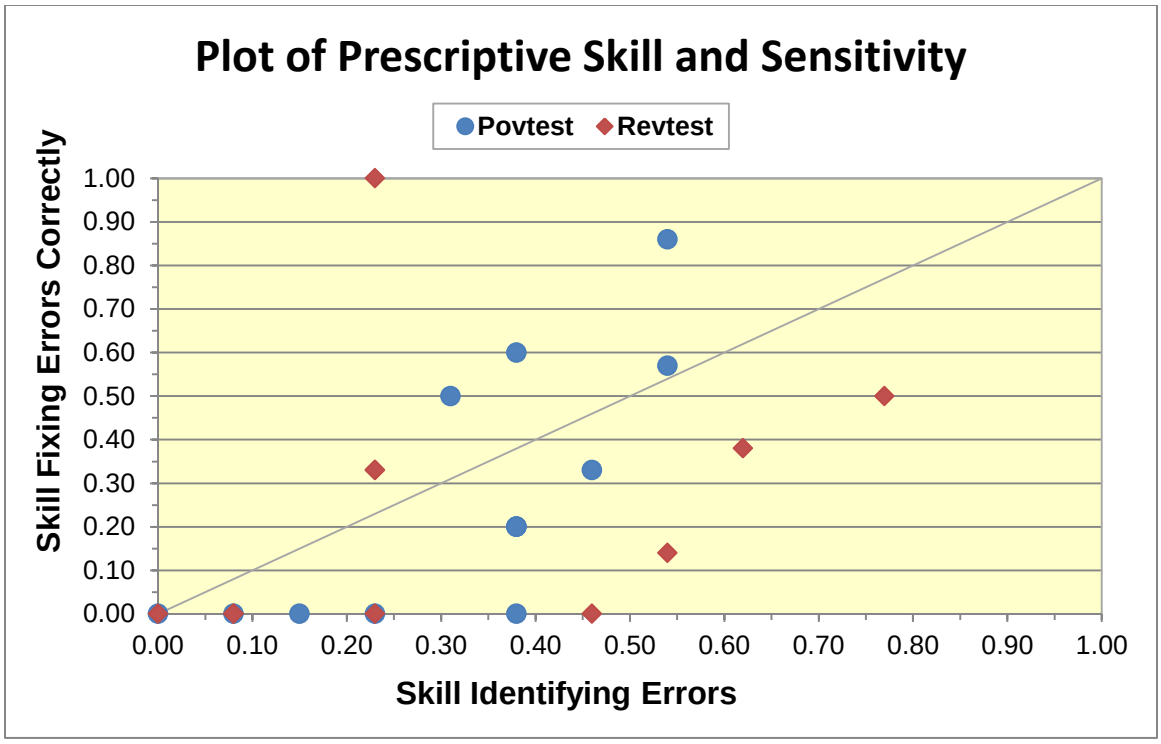


Figure 4

The last of the five scores calculated for the detect-and-correct test is translation rating, which also indicates that this population was better at identifying errors rather than fixing them. Translation rating is the percentage of both errors fixed correctly and non-errors left alone. Translation ratings were overall very high. Figures 5 shows high scores for both translation rating and diagnostic skill while Figure 6 clearly shows high translation rating with low prescriptive skill. The importance and many reasons for this have been discussed above.

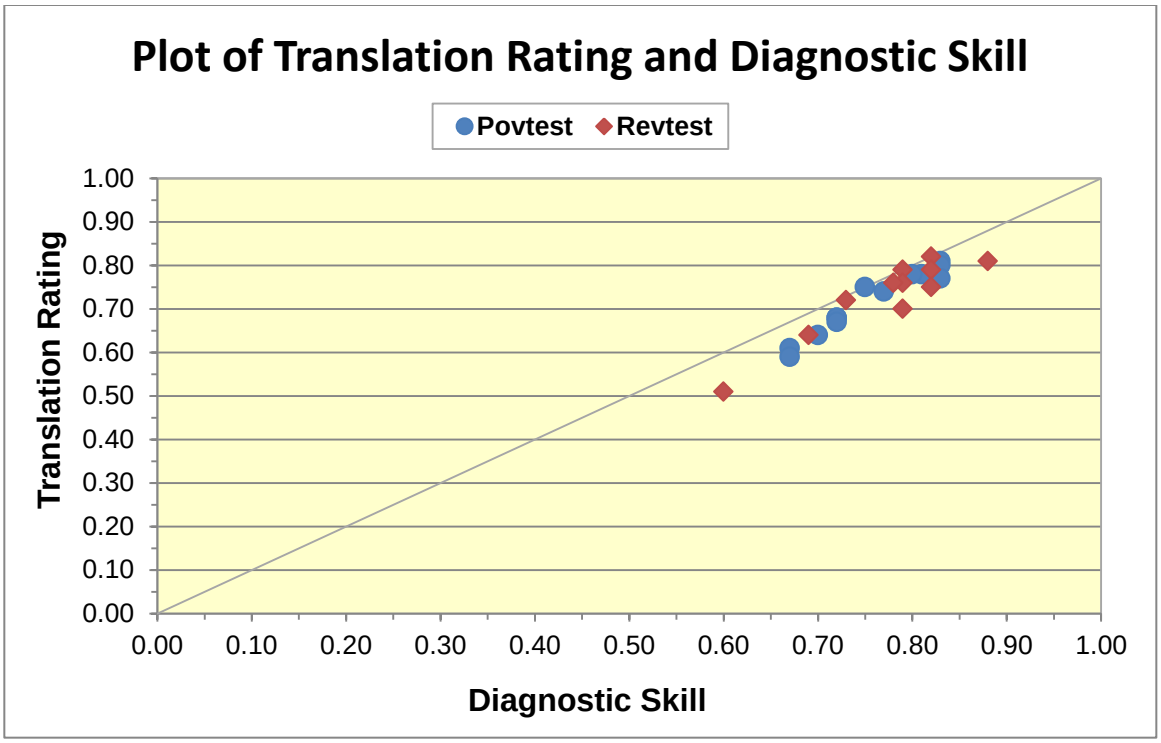


Figure 5

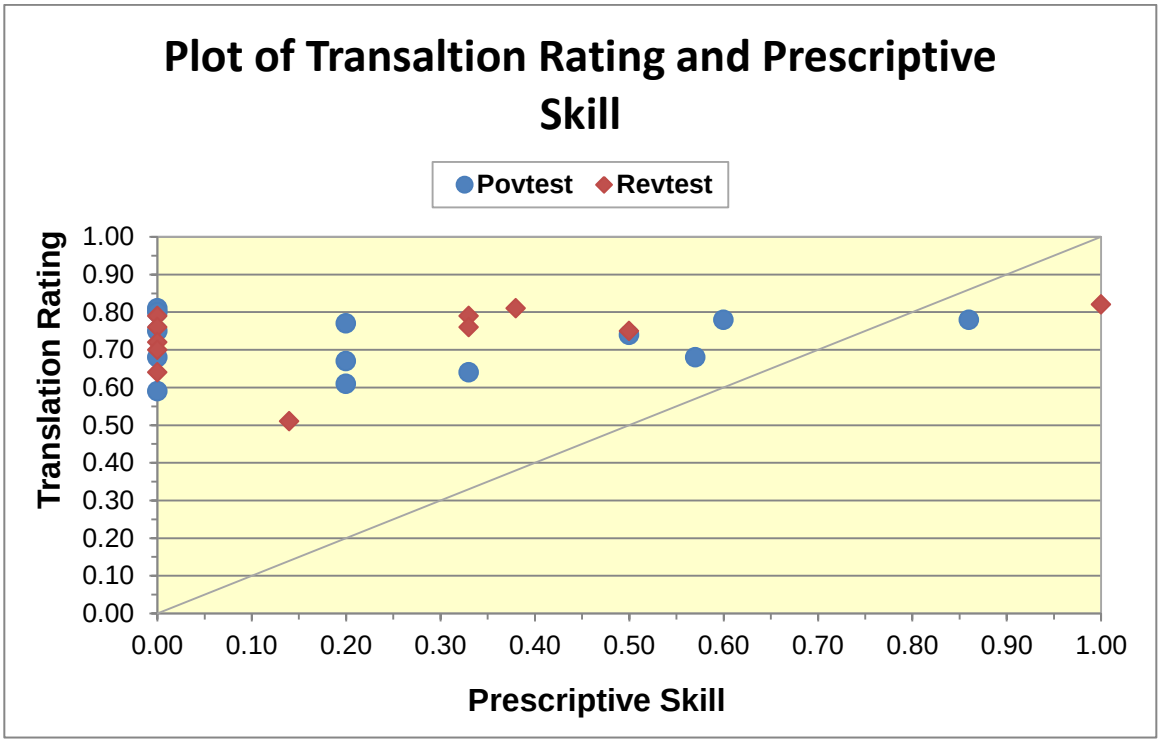


Figure 6

Full Translation Scores

The second stage of analysis focused on the scores given by the two graders on the full translation test. In this stage, a Pearson r correlation and a Bland-Altman Plot of Agreement were used to investigate the inter-rater reliability between the two graders of the full translation tests. It is important to reemphasize that the scores of the full translation test are presented in two numbers. The first number is simply a count of the number of errors each rater gave to a full translation response. The second number is the total of the weighted errors a rater gave each test using the ATA's core documents for grader training including the "Framework for Standardized Error Marking." Each grader recorded both scores for each response.

A Pearson correlation was computed to assess the relationship between the graders. This was done once for the total errors and once for the total of the weighted errors each grader gave a response. There was a positive correlation between both graders when simply counting up the number of errors a rater gave each passage [$r = 0.530$, $p = 0.008$] as seen in Figure 7. This correlation is significant at the 0.01 level. In addition, there was a positive correlation between both graders when totaling the weighted errors given by each rater [$r = 0.986$, $p = 0.000$] as seen in Figure 8, which is also significant at the 0.01 level. While both these correlations indicate inter-rater reliability between the graders, the latter correlation is strong at the 0.08 level. Additionally, the strength of this correlation is uncommon considering the guidelines set by some scholars in the social sciences such as Cohen (1988), who considers an r value of 0.5 or higher as large. This suggests that something may be incoherent with the data or process used in this research.

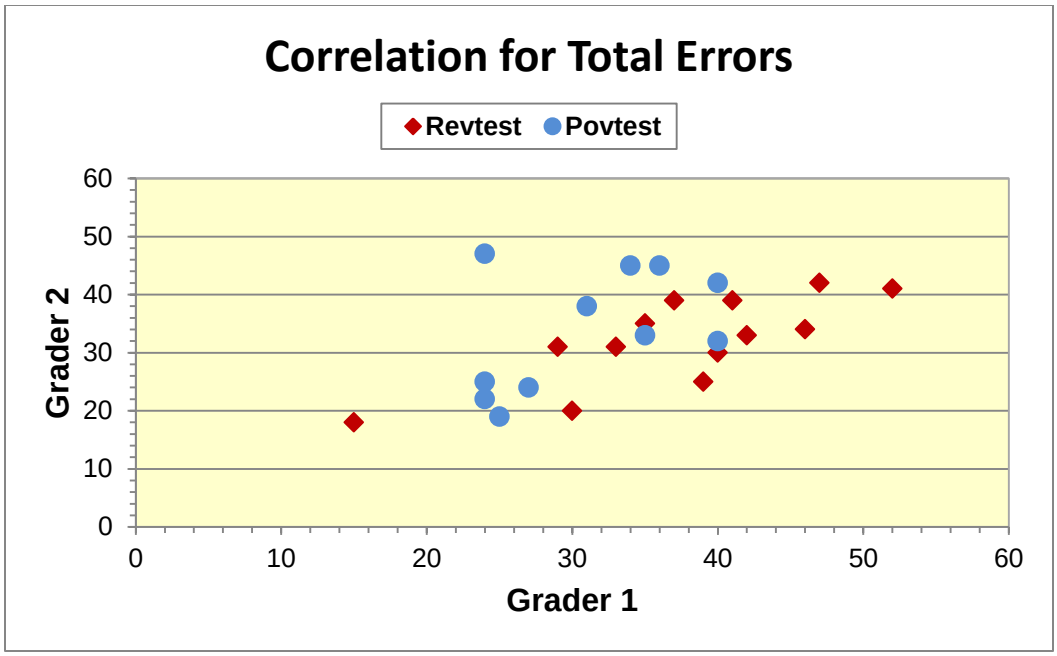


Figure 7

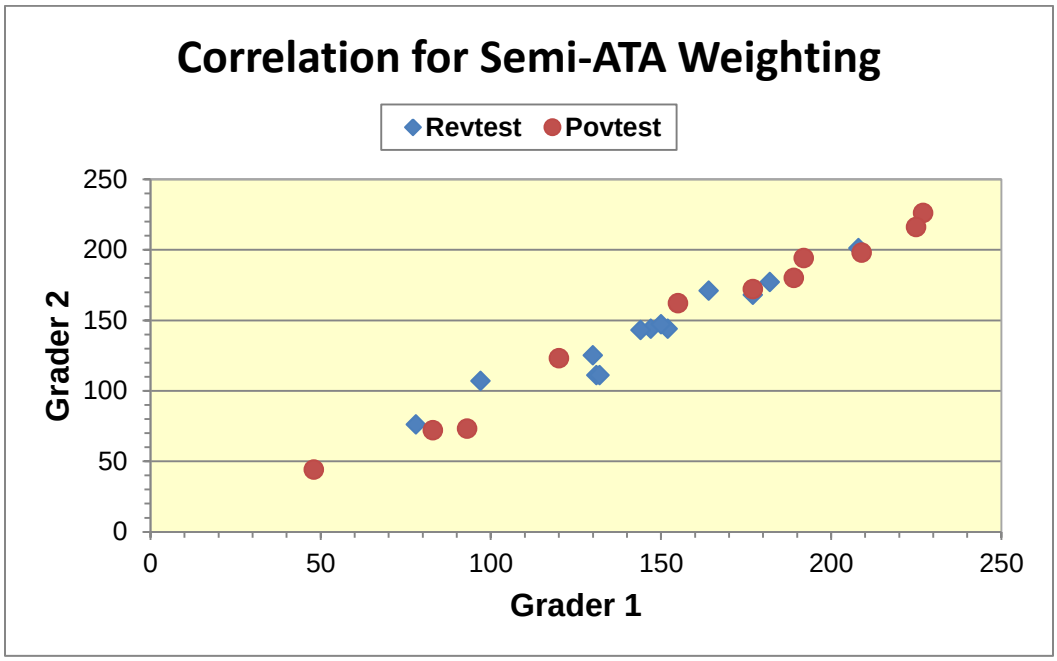


Figure 8

Multiple explanations may account for such a strong correlation between the ATA weighted scores given by the graders. First, one may suspect that something in the grading procedure used in this research caused a bias in the grades. The grading procedures are outlined in the design section. However, it is important to note two major procedural steps, which may be have had an impact here. First, after the graders scored each test response individually, they came together to discuss the scores that varied drastically in order to come to a compromise. This is in line with ATA grading procedures. It was decided that a 25 point difference would be the cut off for grades that would be discussed in order to be methodical. The process of coming to an agreement between graders included the following steps: (1) examining the scoring of both graders and looking for large differences in point value due to different weighting, (2) discussing the reasoning for a given point deduction, (3) agreeing on one grader's method or coming to a middle ground (note the right to disagree was reserved and used at times) and (4) changing the point deduction not only for the one passage but also for any other response with the same error. Step four in this process especially shows that great care was taken to make sure each grader maintained consistency within his or her scores while still allowing for a method to reconcile large differences in a score. This procedure obviously increased inter-rater reliability and that was its intention.

However, it is important to note that ATA weighting may hide differences in the scoring of the two graders. In fact, the process outlined above concerning coming to an agreement on the test responses that differ 25 points or more specifically involves looking for drastic differences in points due to weighting. Therefore, this practice naturally helped increase inter-rater reliability between the ATA weighted scores and did not affect the grading method that just involves counting up the total errors. However, the procedure is similar to the one used by the

ATA though perhaps using the number 25 as a cut off for deciding which tests to discuss may have been too low.

It is important to consider that the effect of weighting may simply be because the entire subject group performed extremely poor. According to the ATA weighting scale, a passing score is one that is less than seventeen error points. Most students' scores were substantially higher, with the overwhelming majority of the types of errors being lexical. (See Appendix for an error type frequency chart.) This again suggests that the level of the text was too high. However, even with a higher-scoring subject group, one could postulate that the ATA weighting procedure hides differences between graders in that similar scores do not necessarily mean that the graders found remotely similar errors in the text, which may indicate issues in inter-rater reliability.

Similarly, due to the extremely poor performance of the subjects, the ATA process of capping a sentence at 16 error points appears to have hidden differences in the scores of the two graders. When a sentence reached 16 errors or more, the sentence was counted as one error in the total errors score and it became one error worth sixteen points according to the weighted ATA scale. This practice is naturally not found in the method of simply totaling the number of errors. Because of poor performance, a large portion of the subject responses included sentences full of errors in addition to very large omissions of phrases and sentences. Therefore, if all the subjects did poorly enough hypothetically they would get 16-point error on all of the sentences and then all subjects would be assigned the same scores. While this extreme did not happen, the correlation above suggests that many of the subjects' scores may have converged because of this practice.

Additionally, the practice of capping sentences may also prevent assessing distinctions between the ability of the subjects. This is important because the capability of the test to reflect

the breadth of a subject's ability is necessary for comparison in the design of this test. However, the fact that the end goal of this research and that of ATA testing is different means that this method may work well for ATA but not in a study like this one. The ATA is interested in determining a pass/fail score for its members. This study was attempting to look at the breadth and range of a translator's competence, so capping actually limited the ability of this study to achieve its goals. As explained above, this phenomenon may be the result of the low performance of the subject group.

In addition, this study uses Bland-Altman Plots of Agreement to further show inter-rater reliability. Both the total error scores and the ATA weighted scores were examined using this tool of analysis, which was developed for the medical field in order to assess two modes of measurement (Bland and Altman, 1986). It was used in this research for two primary reasons. First, the plot of agreement is specifically designed to assess the agreement of two methods of measurement and only two graders were used in this study. Second, it is important to note that there is a difference between agreement and correlation. Bland and Altman 1986 note that when looking at a correlation graph, "We have perfect agreement only if the points lie along the line of equality, but we will have perfect correlation if the points lie along any straight line" (p. 2). Thus, Figures 9 and 10 note whether or not grader 1 and grader 2 agreed, not whether they correlated. Thus, this tool of analysis adds another dimension to the evaluation of inter-rater reliability

Figures 9 and 10 show Bland-Altman Plots of Agreement for the total error scoring method and the ATA weighted scoring method, respectively. The outer two lines in the graph indicate the boundaries of agreement in graphical form and represent two standard deviations of the differences (Bland-Altman, 1986). All but one of the points on the plot lie within these two

lines indicating that there is a reasonable degree of agreement between the two graders. Remember the grading procedure used in this research called for the graders to discuss scores that varied more than 25 points between the graders. The outlier in Figure 9 was never reviewed because it fell within the 25 point requirement. However, it is interesting to note that all scores that were reviewed fall within the boundaries of agreement.

The dotted line in the middle represents the mean, which indicates the bias of one grader over another grader (Atman and Bland, 1983). The mean for the total errors method is 1.50 while the mean for the ATA weighted method is 5.2. The mean here indicates the bias of one grader over another by a difference of 1.50 for the total errors method and 5.2 for the ATA weighted method. This means that one grader was consistently a little harsher than the other in each method. The bias for the ATA method is higher. However, as mentioned before the scores for both tests still fall within an acceptable range of agreement.

In addition, the plots in Figures 9 and 10, like the correlations mentioned above, also show that there is considerably less variance between the graders using the ATA weighted method. This is seen by noting and comparing the standard deviation between grading methods. Figure 9 indicates that the standard deviation between graders stands at 8.44 while Figure 10 shows a standard deviation of 8.11. These numbers are similar but they are not directly comparable because the grading scales between the two methods differ. To resolve this issue it is necessary to calculate the Coefficient of Variation, or the relative standard deviation to the mean of all the data for each scoring method. The Coefficient of Variation for the total errors scores is 25.1% while it is 5.5% for the ATA weighted scores. This shows that while both plots indicate inter-rater reliability, the scores calculated with the ATA weighted scale vary considerably less than the scores of the total errors method just as the Pearson correlations above

show. Again this is worthy of note because it may indicate that the nature of the ATA weighting scale, especially the method of capping sentences, may actually be hiding variations between graders and variations between the ability of this particular subject group. This may have produced less information about the quality of translation.

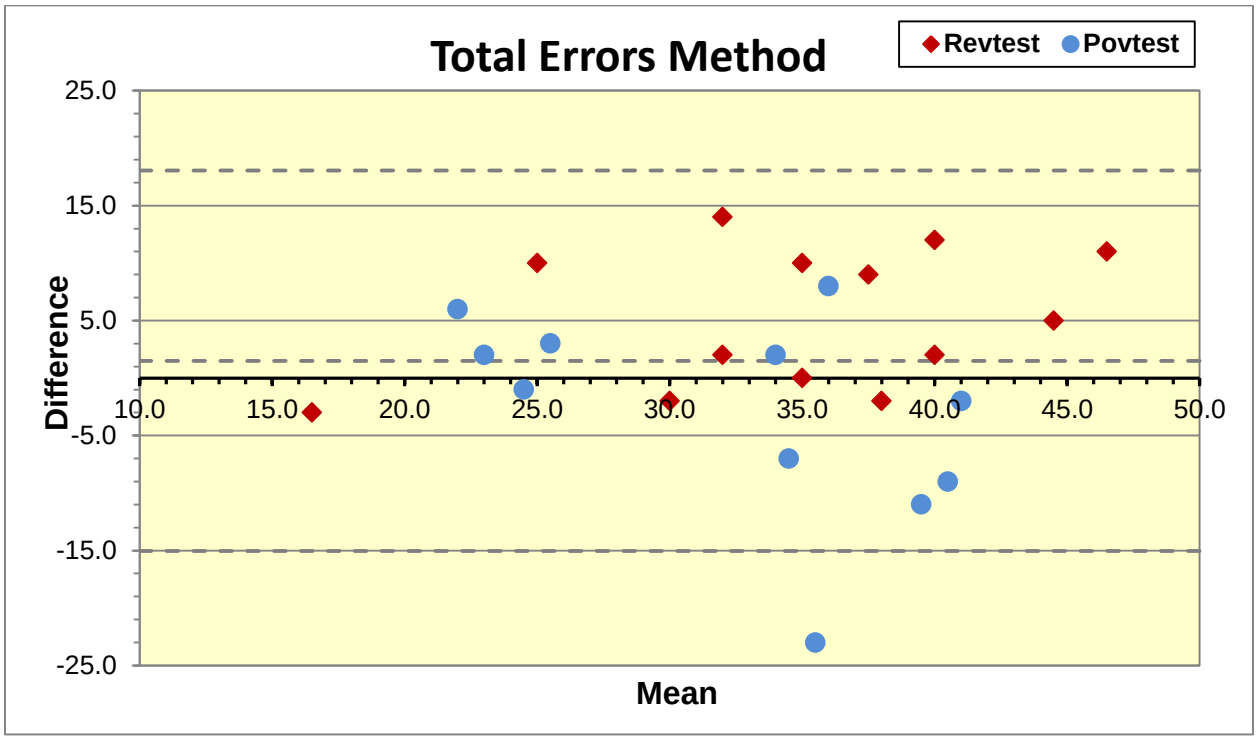


Figure 9

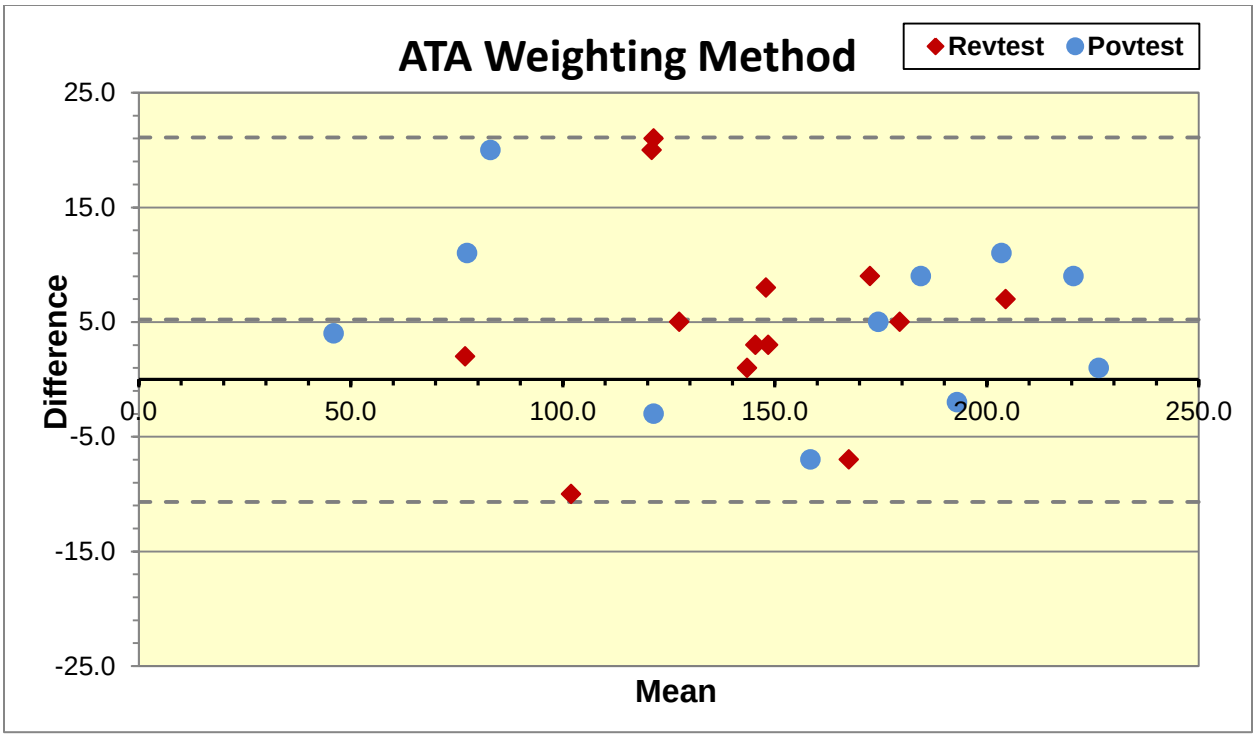


Figure 10

Comparison of Full Translation and Detect-and-correct Methods

Finally, a Pearson correlation was used to analyze the relationship between the overall translation rating score for each response of the detect-and-correct portion of the test and the scores of the full translation portion of the test. This was done once for the total errors grading method and once for ATA weighted grading method. There was a negative correlation between the two versions of the test when simply counting up the number of errors a rater gave each passage [r = -0.454, p = 0.026] as illustrated in Figure 11. This correlation is significant at the 0.05 level. In addition, there was a negative correlation between the scores of both graders when totaling the ATA weighted errors given by each rater [r = -0.318, p = 0.130] as seen in Figure 12. This correlation is not significant. The direction for both correlations is very critical. First, when totaling the number of errors or the ATA weighted errors the higher the score the worse the performance. However, the translation rating score is a percentage out of one hundred so a

higher score is a good score. This means that one expects the correlations between these scores to be negative.

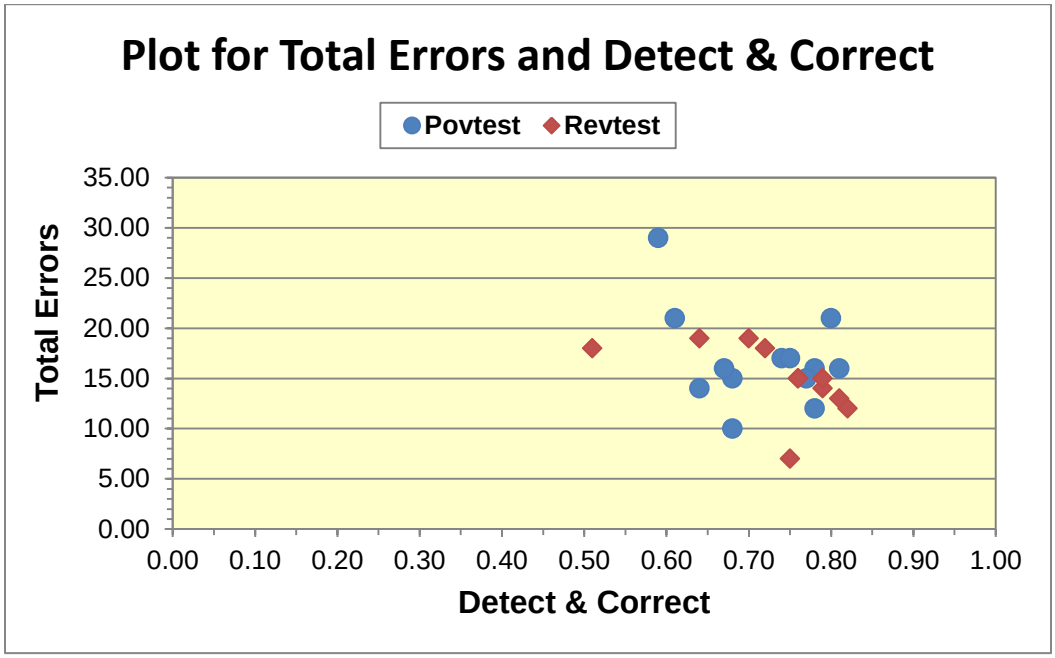


Figure 11

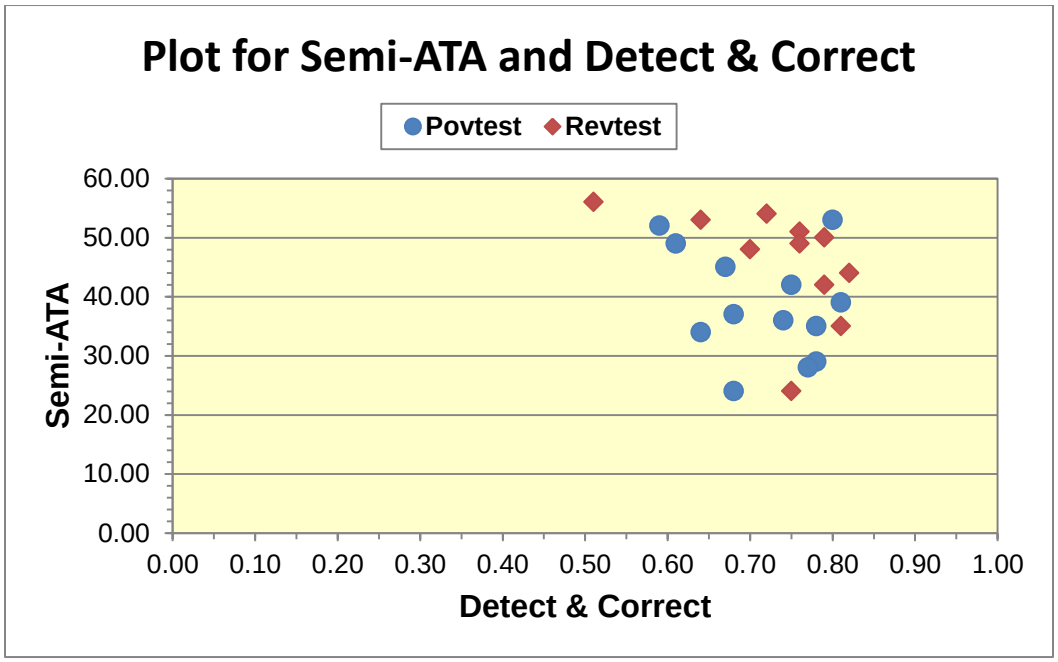


Figure 12

The fact that one scoring method achieved significance while another had no significance may be the result of a number of situations. One situation is that the subject population used in this study is very limited. It is possible that if more subjects were tested the correlation with no significance would become significant, but it is important to note that the opposite could happen as well. In addition the fact that the detect-and-correct translation rating score correlated more with the total errors method of grading the full translation brings up an important issue. Namely, is a method of grading that simply totals errors sufficient enough to assess the complexity of translation? Arango-Keith and Koby (2003) call for "more reliable translation quality assessment instruments that would go beyond error marking protocols towards a more comprehensive approach" (p. 129). The fact that the detect-and-correct test correlates more with the total error marking assessment method may suggest that the detect-and-correct portion of the test, as it is currently, relies too heavily on totaling scores. Future studies could help confirm this.

A Pearson correlation was also performed between all five scores of the detect-and-correct test and the two scores of the full translation test. The results are displayed in Chart A. The r-value is given for each pair with varying shades indicating the strength of each value. One thing this chart shows is that using the ATA weighted scale did not add any more information about a subject's prescriptive ability. This is seen by the fact that the r-value between the total errors method of grading and prescriptive ability [r = -0.5667, p = 0.004] and the r-value between the ATA weighted method of grading ("Semi-ATA Scale") and the prescriptive ability [r = -0.5714, p = 0.004] are similar. This, in addition with the fact that the ATA scale seemed to give graders more congruent scores when testing low-performing populations, brings into question the validity of using the ATA weighted scale for the testing translation ability of this population type.

Most importantly, Chart A also gives evidence that the full translation portion of the test following the ATA method of grading may not be testing the same thing as the detect-and-correct portion. This is first seen in the negative correlations between the weighted ATA scores and the sensitivity [r = -0.5319, p = 0.007] as well as prescriptive skills [r =.-0.5714, p = 0.004] both of which are significant at the 0.01 level. Note that this negative correlation is expected because a good score according to the ATA grading method is a low score while a good score on any of the detect-and-correct scores is a high score. It is also important to note that both the sensitivity score and the prescriptive score deal with identifying and fixing error chunks correctly. Thus, given this particular subject population, the ATA weighted scale appears to be designed to test prescriptive skills.

On the other hand, Chart A gives evidence that the culminating translation rating score of the detect-and-correct portion of the test focuses more on testing diagnostic skills. This is seen in a very strong positive correlation at the 0.08 level between the translation rating scores and the specificity [r = 0.9465, p = 0.000] and diagnostic skill [r = 0.9480, p = 0.000] scores. In addition, weak correlations at the 0.08 level are shown between translation rating and prescriptive [r = 0.3039, p = 0.149] and sensitivity skills [r = -0.2965, p = 0.159]. In fact the correlation between the translation rating scores and the sensitivity scores is negative, which may indicate that they were testing different things. Note that a positive correlation is expected when comparing the five detect-and-correct scores with each other because a good score for each is a high score. Thus, the results of this Pearson correlation analysis appears to suggest that there is some but not a significant degree of correlation at the 0.08 level between the full translation portion of the test and some of the detect-and-correct scores. However, there is also strong

evidence that the detect-and-correct portion of the test and the full translation portion are testing different things.

r-Values for Pearson Correlation between All Scores

	Specificity	Sensitivity	Diagnostic Skill	Prescriptive Skill	Translation Rating	Total Errors
Sensitivity	-0.5423**					
Diagnostic Skill	0.8698**	-0.0604				
Prescriptive Skill	0.0285	0.4765*	0.2995			
Translation Rating	0.9465**	-0.2965	0.9480**	0.3039		
Total Errors	-0.2473	-0.3009	-0.4436*	-0.5667**	-0.4544*	
Semi- ATA Scale	-0.0780	-0.5319**	-0.3946	-0.5714**	-0.3177	0.6801*

**Correlation is significant at the 0.01 level.
* Correlation is significant at the 0.05 level.

Specificity	Skill Identifying Non-Errors
Sensitivity	Skill Identifying Errors
Diagnostic Skill	Skill Identifying Errors and Non-Errors
Prescriptive Skill	Skill Fixing Errors Correctly
Translation Rating	Skill Fixing Errors Correctly & Not Fixing Non-Errors

Chart A

The subsequent question, then, is what are each of these tests testing? For the purposes of this research, because the full translation section of the CTT Arabic test resembles the full translation test given to those seeking ATA Certification, it is assumed that the full translation test is testing translation. However, it appears from the data above, which is limited to one

sample, that the detect-and-correct styled test may be testing another skill. It is important to note that it is possible that the detect-and-correct form of the test is testing post-editing skills rather than translation competence. Serenda (1982) defines post-editing as a "process …[that] involves two main tasks: to identify errors in the translation and to find solutions for the errors" (p. 121). Even though post-editing always refers to machine translation, and the detect-and-correct test does not deal with machine translation, but rather human translation, one can put this aside to notice that the process of the test at first glance appears similar to the post-editing process described by Serenda above.

A number of similarities appear between the detect-and-correct test and the post-editing process. First, Obrien (2002) notes the practical difference between translation and post-editing, the pivotal point of her argument being:

> Post-editing and translation differ on the practical level. Translation usually involves one source text and the creation of one target text to a level of publishable quality. Post-editing, on the other hand, involves two source texts, i.e. the text authored in the source language and the raw MT output [Note this use of the term "source text" is highly unconventional], which a translator uses to help produce a final version. (p. 101)

Similar to the post-editing task, the process of the detect-and-correct test involves two source texts as well: the text in the source language and the faulty translation in the target language. The target text in this procedure is the final response after the subject changes all the chunks they feel necessary.

Another similarity between the detect-and-correct form of the test and post-editing concerns the type of errors involved in the test. While Serenda (1982) and Lavorel (1982) list various types of errors often found in post-editing such as verb form errors, these errors appear to

be similar to those found when dealing with revising human translation as well. However, Obrien (2002) notes a more substantial difference between the errors of post-editing and the errors of human translations, specifically that "with post-editing misconstructions are more likely to be local rather than global" (p.101). This fact is also true of errors in the detect-and-correct version of the test. Because the CTT testing suite does not have the ability to automatically score global errors, all errors in the test are local not global. This point is a strong basis for future research, which may be able to determine whether the detect-and-correct portion of the test is actually testing post-editing skills. Based on the current evidence in this research, one can only offer this as a hypothesis for further testing.

CONCLUSION AND FUTURE WORK

The purpose of this research was to determine whether a good score on the detect-and-correct portion of the CTT Arabic test is a good indicator of a good score on a full translation test. It has been determined that for this particular subject population this is not the case. In fact, the data indicates that it is likely that these two types of tests do not necessarily test the same thing. It is assumed in this research that the full translation test, because of its foundation in the methodology of the ATA certification exam, does actually test translation ability. However, one may question whether the detect-and-correct version of the test actually tests translation ability. One can hypothesize that the detect-and-correct version of the test may actually test skills that resemble post-editing skills because of the nature of the procedure of the test, including its reliance on two source texts as well as the fact that the type of errors used in the test are local, not global. Local errors are more common in post-editing work, which deals with machine translation, while global errors are more common in human translation (Obrien, 2002). This finding is noteworthy in that it helps provide a foundation for future work in not only computerized testing research but also post-editing research.

A second finding provides insight concerning Second Language Acquisition (SLA) programs and their intentions. While this research was limited in its subject population, it clearly shows that subjects were better at identifying errors rather than correcting them. The importance of this idea is found in possible causes for this phenomenon. One cause may be because the subjects were students or recently graduated students and their training up to that point had only prepared them to identify errors rather than correct them. This implies that the skills that are involved in translation competence come with a special kind of training and experience, which fits well into the literature especially that which calls for a special translation pedagogy.

However, this finding has no meaning for SLA if the purpose of a Second Language Acquisition program does not claim to provide translation skill training that advances one's competence. Although if the opposite is true, it may indicate that SLA programs should review their curriculum objectives and course descriptions to provide a more accurate description of their purposes. This would be especially important if further research indicates that prescriptive ability is primarily affected by acquiring experience and less affected by language proficiency training or training in the skill sets involved in translation competence.

In addition, this research has also provided three additional valuable insights into the logistics of testing design. First, this research has shown that the difficulty of the source text may skew results when the skill level of the subjects is very low. The source texts used in this passage were at the ILR 3 level. The subjects used in this research were students or recently graduated students who had not been greatly exposed to this level of text. It is important to note that a pilot test could have detected this issue of text difficulty if one had been preformed. This was a great limitation to this research. It is important that future studies avoid this mistake by conducting pilot tests to determine whether a test is too hard or to easy for a particular sample population. In addition, testing subjects with a broader range of ability and using a large spectrum of passage difficulty could clarify the extent of this effect on test results.

Second, this research has shown reason to question whether the length of the source and target text and the number of chunks into which it was divided gave subjects a fair chance for success on the detect-and-correct version of the tests. This research cites the relatively small number of chunks into which a test passage was divided as a possible reason for the overall low performance on the detect-and-correct portion of the test. One of the reasons for the small number of chunks was the limited length of the test passages, at approximately one hundred-

seventy words. Studies that vary the length of a source text and the chunks into which it is divided may be able to determine how much the number of chunks affects a subject's score.

Third, evidence suggests that when a sample group performs extremely poorly, the ATA practices of sentence capping and weighting errors appears to hide discrepancies both between the ability of subjects and between grader scoring. Essentially, since the majority of subjects performed poorly, many of their sentences had to be capped at sixteen points, i.e., be weighted heavily. If the majority of sentences are weighted heavily and capped at 16 points, the scores of the test become very similar. Thus, it becomes difficult to determine inter-rater reliability or to analyze the differences in performance. It is important to note that the issue of capping sentences may not interfere with the purposes of the ATA in attaining a pass or fail result. However, this issue is important in studies like this one because the ability of the test to reflect the breadth of a subject's ability is necessary for comparison in the design. Therefore, future tests should not score tests using the ATA practice of capping and weighting scoring on known low-scoring sample groups.

However, it is important to note that even in higher scoring sample groups, the ATA method of weighting may hide differences between the types and number of errors each grader marks. Indeed just because two graders give a passage the same score does not mean that they weighted each error the same way. In fact, they may not even mark the same errors. One may question the effect of this on inter-rater reliability. Future research may help determine the relationship between this practice and inter-rater reliability.

My research, while it did not show strong support for its original hypothesis, it still produced many unforeseen findings that have added insight to current knowledge concerning the challenges of assessing translation competence. These findings have been outlined above and

can provide direction in designing future testing methods for translation ability. However, it is important to note that future work that looks at which types of errors are harder or easier for various subject samples to correct may lead to a detect-and-correct style test that can in part predict a good score on a full translation test, as originally hypothesized. In addition, the limitations of this research should be noted again and steps should be taken to avoid them. This research is limited in scope and number. Other weaknesses include flaws in procedure such as the absence of a pilot test, not limiting the time allotted to take the test and the three logistical flaws in the grading procedure outlined in the design section. However, this does not nullify the need to further research the significance of the findings mentioned above.

Specifically, in connection to the research suggestions mentioned above, multiple studies need to be conducted using a broad range of subject groups from untrained to experienced translators. Only then will patterns become evident. These groups may include bilingual speakers, translation students, translators with varying years of experience, certified translators including multiple forms of certification, etc. The wide range of the ability of practicing translators makes this point crucial. Of course, these studies should first attempt to address the various issues in logistical testing found in this study before testing multiple sample groups. Then, with enough subjects in each population, research on this scale could detect patterns that would expand the key findings of this research outlined above, which would give insight into multiple fields including translation theory, translation quality testing, SLA theory and SLA Testing.

In addition, the use of the think-aloud method may also be of interest in analyzing how various sample groups perform on translation (or post-editing) assessments created by the CTT testing suite. Jääskeläinen 1998 states this method "involves asking a translator to translate a

text and, at the same time, to verbalize as much of his or her thoughts as possible" (p.266). Note that a written transcription of a recording of a session like this is called a think-aloud protocol (TAP). Wakabayashi 2003 suggests the use of this method to "address reoccurring problems in how students approach the task of translation" (p. 61). Similarly, a future study may be able to test a wide variety of translators to analyze how their approach, their experience and their score on a test created using the CTT testing suite are related. A similar study could be set up for post-editing assessment if future studies reveal the detect-and-correct test is actually testing post-editing skills. In fact, Krings 2001, as cited in Obrien 2005, defends the use of the think aloud method to analyze post-editing at a technical level and while Obrien does note many valid critiques of this method including the impact it has on the post-editing process, the method may still be useful in gaining insight on topics such as the differences in approach between post-editing and translation tasks.

Finally, further research should also analyze on a large scale whether a computerized detect-and-correct test could successfully be used to assess post-editing skills. As outlined above, the results of this research when compared with findings in the literature concerning post-editing suggest that the detect-and-correct styled test used in this research may test post-editing skills. Thus, the detect-and-correct styled test used in this study, which is based on the ongoing research of Hague et al. (2012) appears to be a good candidate for future testing in this area given the research here. However, the development of other types of detect-and-correct styled tests would also be a valuable course of action.

REFERENCES

ALTMAN, D. AND BLAND, J. 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. *Statistician* 32(1983) 307-310.

ARANGO-KEITH, F. AND AND KOBY, GEOFFREY S. 2003. Assessing Assessment: Translator training evaluation and the needs of industry quality assessment. In BAER AND KOBY, 117-134.

ATA (AMERICAN TRANSLATORS ASSOCIATION). n.d. Certification Exam. *In Certification.* Online: http://www.atanet.org/certification/upcoming.php.

ACTFL, 1999. *ACTFL Proficiency Guidelines Speaking.* Online: http://www.actfl.org/files/public/guidelinesspeak.pdf.

ACTFL, 2001. *ACTFL Proficiency Guidelines Writing.* Online: http://www.actfl.org/files/public/writingguidelines.pdf

ACTFL. 2012. *ACTFL Proficiency Guidelines.* Online: http://actflproficiencyguidelines2012.org/.

ASTM. 2006. *ASTM F2575-06: Standard Guide for Quality Assurance in Translation*. West Conshohocken, PA: ASTM International.

BAER, BRIAN J. AND KOBY, GEOFFREY S, (eds.) 2003. Beyond the Ivory Tower: Rethinking translation pedagogy. Amsterdam: John Benjamins Publishing Company.

BAKER, M. 1992. *In Other Words: A coursebook on translation*. London: Routledge.

BAKER, M. (ed.) 1998. *Routledge Encyclopedia of Translation Studies.* London: Routledge.

BLAND, J. AND ALTMAN, D. 1986. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* 327(8476) 307-310.

CHOMSKY, N. 1965. *Aspects of Theory of Syntax.* Cambridge, MA: MIT Press.

CLARK, J. AND CLIFFORD, R. 1988. The *FSI/IRL/ACTFL* Proficiency Scales and Testing Techniques: Development, Current Status, and Needed Research. *Studies in Second Language Acquisition* 10(2) 129-148.

COHEN, J.1988. *Statistical power analysis for the behavioral sciences (2nd ed.).* Mahwah, NJ: Lawrence Erlbaum.

COLINA,S. 2009. Further evidence for a functionalist approach to translation quality evaluation. *Target: International Journal on Translation Studies* 21. 235–264.

DEKEYSER, R. 2003. Implicit and Explicit Learning. In DOUGHTY AND LONG, 717-761.

DIKLI, S. 2006. An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5. 1-35.

DODDINGTON, G. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. *In proceedings of the Human Language Technology Conference (HLT), San Diego, CA.* 1*28*–132.

DOUGHTY, C.J. AND LONG, M. H. (eds.) 2003. *The Handbook of Second Language Acquisition. Malden, MA*: Blackwell Publishing.

DOUGLAS, D. 1988. *Studies in Second Language Acquisition* 10(2) 121-128.

DURBAN AND MELBY. 2008. Translation: Buying a Non-commodity. ATA publication. Online: http://www.atanet.org/docs/translation_buying_guide.pdf.

DUNNE, K. 2011. *From vicious to virtuous cycle: Customer-focused translation quality management using ISO 9001 principles and Agile methodologies*. Amsterdam: John Benjamins Publishing Company.

ELIS, NICK, C. 2003. Constructs, Chunking and Connectionism: The Emergence of Second Language Structure. In DOUGHTY AND LONG, 63-103.

GENTZLER, E. 2001. *Contemporary Translation Theories, 2nd Ed.* Clevendon: Multilingual Matters, LTD.

GLASER, R. 1963. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. *American Psychologist* 18. 519–521.

HAGUE, D., MELBY, A. AND ZHENG, W. 2011. Surveying Translation Quality Assessment. *The Interpreter and Translator Trainer.* 5(2). 243-267.

HAGUE, D., MELBY, A AND ZHENG, W. 2012. Personal Communication. Provo, UT.

HOUSE, J. 2001. Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta : journal des traducteurs / Meta: Translators' Journal.* 46. 243–257.

HYLTENSTAM, K AND ABRAHAMSSON, N. 2003. Maturational Constraints in SLA. In DOUGHTY AND LONG, 19-42.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. 2012. *ISO/TS 11669: Translation projects – General guidance*. Geneva, Switzerland: ISO.

ILR (INTERAGENCY LANGUAGE ROUNDTABLE). n.d. ILR Skill Level Descriptions for Translation Performance. Online: http://www.govtilr.org/Skills/AdoptedILRTranslationGuidelines.htm.

JÄÄSKELÄINEN, R. 1998. Think-aloud protocols. In BAKER 1998, 265-269.

KOBY, G AND CHAMPE, G. in press. Welcome to the Real World: Professional-Level Translator Certification. *The International Journal of Translation and Interpreting Research 4 (1).*

LAWSON, V. (ed.) 1986. Practical Experience of Machine Translation: Proceedings of a Conference, London, ..., Amsterdam: North-Holland.

LAVOREL, B. 1982. Experience in English–French Post-Editing. *In* LAWSON, 105–109.

MCCLELLAND, J., RUMELHART, D. AND THE PDP RESEARCH GROUP (eds.) 1986. *Parallel Distributed Processing: Explorations in the Microstructure oaf Cognition. Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

MELBY, A, MANNING, A AND KLEMETZ, L. 2005. Quality in Translation: a lesson for the study of meaning. *Linguistics and the Human Sciences 1 (3)* 403-446.

MELBY, A. 2012. Personal Communication. Provo, UT.

NATIONAL FOREIGN LANGUAGE CENTER. 2006. LangNet Virtual Institute: Passage Rating Course, Maryland: The National Foreign Language Center at the University of Maryland.

NORD, C. 2006. Translating as a Purposeful Activity: A Prospective Approach. *TEFLIN Journal 17 (2)* 131-143.

NORRIS, J. AND ORTEGA, L. WHITE, L. 2003. Defining and Measuring SLA. In DOUGHTY AND LONG, 717-761.

OBRIEN, S. 2002. Teaching Post-editing: A Proposal for Course Content. *In 6th EAMT Workshop Teaching machine Translation, Manchester.* 99-106.

OBRIEN, S. 2005. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation 19 (1)* 37-58.

PAPENINI, K., ROUKOS, S., WARD, T. AND ZHU, W. 2002. BLEU: A method for automatic evaluation of machine translation. *In proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* 311-318.

RUSSEL, G. 1998. The Definition and Perception of Quality in ISO-9000 Firms. *Review of Business 19 (3)* 13-16.

SEREDA, S. 1982. Practical Experience of Machine Translation. In LAWSON, 119–123.

SCHAEFFNER, C. 1998. Skopos theory. In BAKER 1998, 235-238.

STEJSKAL, J. 2009. International Certification Study. *The ATA Chronical. 1*5-18.

VLADMAN, A. 1988. Introduction. *Studies in Second Language Acquisition 10(2).* 121-128.

WAKABAYASHI, J. 2003. Think-alouds as a pedagogical tool. In BAER AND KOBY, 61-82.

WHITE, L. 2003. On the Nature of Interlanguage Representation: Universal Grammar in Second Language. In DOUGHTY AND LONG, 19-42.

APPENDIX

Excerpt from Source Text: “Povtest”

الفقر ليس رجلاً

قبل زهاء خمسة عشر قرناً تحسّر الخليفة الراشد عمر انه "لو كان الفقر رجلاً لقتلته". لكن الفقر ليس رجلاً، وهو ما زال حياً.

الفقر نتاج سياسات وممارسات.

واليوم تريد منظمات غير حكومية ان تجعل من الفقر ماضياً يذهب الى غير عودة. لكنها ستفشل. فالفقر باق ما بقيت اسبابه.

وأسبابه طبيعة انسانية في بعضها، وفشل سياسي واخلاقي في بعضه الاخر.

يتحمل الغرب مسؤولية مساعدة شعوب يقتل الفقر اطفالها، ويحرمها معالجة مرضاها. فالغرب استغل الدول الفقيرة مستعمراً.

وهو الذي يثري متاجرة في اسواقها التي تستهلك ولا تنتج ولا تستطيع ان تنافس.

ويساعد الغرب في كبح جماح الفقر من منطلقات مصلحية. فالفقر يعني هجرة الى اسواقه. وهذا حراك سكاني ترفضه دول غربية عديدة وترى فيه شعوبها خطر على اقتصادها وهويتها.

بيد أن تخفيف حدة الفقر يتطلب خطوات لا يبدو ان الدول الغربية مستعدة لاتخاذها.

وهذه الخطوات ليست مقتصرة على توفير ما يكفي من دعم يتطلبه اطلاق مشاريع اقتصادية ناجحة او الغاء ديون تستنفد خدمة فوائدها موارد الدول الفقيرة.

محاربة الفقر تتطلب سياسات تجارية تأخذ بعين الاعتبار ضعف اقتصاديات الدول الفقيرة، وتخلف قدراتها التنافسية.

Full Text:

http://alghad.com/index.php?article=1741&searchFor=%C7%E1%DD%DE%D1

Model Translation: "Povtest"

Fifteen centuries ago the Caliph Omar lamented, "If poverty were a man I would kill him." However, poverty is not a man and it is still alive. Poverty is the product of policies and practices.

Today non-governmental organizations want to make poverty a thing of the past that leaves and never returns, but they will fail. Poverty is something that will remain as long as its causes remain. Some of its causes are human nature and other causes are political and moral failure.

The west bears responsibility for helping peoples where poverty is killing their children and prohibiting them from the treatment of their sick. The west exploited the poor countries through colonialism. It is the one that is getting rich by trading in their markets, which consume, but do not produce and cannot compete.

The West is helping overcome poverty from a perspective of self interest. Poverty means a migration to its markets. This is a population movement that numerous Western nations refuse and whose people see it as a danger to their economy and identity.

However, the alleviation of the severity of poverty demands steps that the Western nations do not seem ready to take.

These steps are not limited to providing sufficient support, which is required by the launching of successful economic projects or the cancelation of debts whose interest is exhausting the resources of the poor countries

The fight against poverty requires commercial policies that take into account the weakness of the economies of the poor nations and the lag in their competitive ability.

Excerpt from Source Text: "Revtest"

ثورة مصر في كل بيت عربي

على الصعيد الداخلي عانى المصريون من النظام أكثر من أي أحد آخر. عانوا من القمع وغياب الحرية والتعددية. عانوا من الفساد الذي استهلك مقدراتهم. عانوا من الفقر والبطالة. وصار من حقهم أن ينتفضوا في وجه ذلك كله، ويبحثوا عن الحرية والعيش الكريم في ظل نظام يختارونه بحرية.

هكذا كان النظام المصري يذلّ المصريين، ويذل الأمة من ورائهم، لكن الذي لا يقل أهمية في ثورة الشعب المصري هو أنها تفتح باب الحرية لجميع العرب، إذ ليست مصر وحدها التي عانت من القمع وغياب الحرية والفساد والفقر، وليس نظامها وحده من رهن قراره للخارج، فمن ورائه تصطف أنظمة كثيرة، ما يعني أن نجاح الثورة المصرية سيكون مقدمة لتحرير الأمة بأسرها، وسيبزغ أمامها فجر جديد بإذن الله.

بعد ثورة المصريين الشرفاء لن تقبل الأمة أن يستعبدها حاكم، أو تتحكم فيها نخب تهمين على السلطة والثروة وترهن قرار البلاد للخارج، فقد أدركت جماهير الأمة سر قوتها، وهي لن تستكين أبدا حتى يفتح الله بينها وبين من يضطهدونها وهو خير الفاتحين.

سلام على مصر وعلى شعبها الأبي، وسلام على شهدائها وأبطالها ومن يحملون راية الحق إلى يوم الدين.

Full Text:

http://aljazeera.net/NR/exeres/7C9EB97C-7A78-4344-9187-5AB022E4D027.htm

Model Translation: "Revtest"

At the domestic level, the Egyptians suffered from the regime more than anyone else. They suffered from repression, an absence of freedom and pluralism. They suffered from corruption, which consumed their capabilities. They suffered from poverty and unemployment. And it became their right to rise up in the face of all of this and pursue freedom and a respectable living under a regime that they freely choose.

Consequently, the Egyptian regime used to humiliate the Egyptians and it use to humiliate the Arab nation behind its back but what is no less important in the revolution of the Egyptian people is that it opened the door of freedom for all Arabs since Egypt is not the only one that has suffered from repression, the absence of freedom, corruption and poverty. Its regime is not the only one who subjected decision-making to foreign nations. Behind it stands many regimes, which means that the success of the Egyptian revolution will be a prelude to the liberation of the Arab nation as a whole and a new dawn will break forth, god willing.

After the revolution of the honorable Egyptians the Arab nation will not accept being enslaved by a ruler or being controlled by groups that dominate power and wealth and subject the country's decision-making to foreign nations. The masses of the Arab nation realized the secret of its strength and it will never be silent until god grants its victory over its persecutors. He is the best of all conquerors.

Peace on Egypt and on its proud people and peace on its martyrs and its heroes and those who bear the banner of justice until the Day of Judgment.

Translation Brief: “Povtest” and “Revtest”

The translation should only correspond to the section of the source text indicated. It should sound like an article originally written in English, free of cultural, lexical, and grammatical errors. However, this is not a grammar test. See the glossary for preferred translations of certain words

Your translation should not assume that the reader has lived in the Middle East or knows much about the language or the area. The content of the source text should correspond closely with the content of the target text, although sentences and phrases may be divided, merged or otherwise altered to maintain English flow. English equivalents of Arabic metaphors and colloquialisms may be substituted. Names should be transliterated unless otherwise stated.

The errors in this text are LOCAL not GLOBAL. You will not have to make changes across paragraphs and sentences. There are multiple ways to translate a text "the right" way. Do not make changes based solely on your personal preferences.

An appropriate register for English journalism should be maintained throughout the text.

You may consult a general purpose bilingual Arabic-English dictionary but no electronic or internet dictionaries or other outside sources such as the internet, a friend, etc.

Passage Specific Guidelines: "Povtest"

- It should sound like an article originally written in English, free of cultural, lexical, and grammatical errors.
- Your translation should not assume that the reader has lived in the Middle East or knows much about the language or the area.
- The content of the source text should correspond closely with the content of the target text, although sentences and phrases may be divided, merged or otherwise altered to maintain English flow.
- English equivalents of Arabic metaphors and colloquialisms may be substituted. Names should be transliterated unless otherwise stated.
- An appropriate register for English journalism should be maintained throughout the text.
- Preferred translations of certain words:
    1. الخليفة الراشد عمر "the Caliph Omar"
    2. الراشد is an epithet meaning "rightly guided" that is used in connection to the first four caliphs. It is usually omitted when translating.
    3. خدمة فوائد "interest" (pl.)

(FOR BOTH PASSAGES) Model	Projected Errors	Suggested Method for Scoring
Large Omission	Whole sentence or nearly whole sentence missing from translation	O (16)
Other Omission	Omits idea (such as verb) but the English still makes grammatical/logical sense without it	O (2)
Other Omission	Omits idea (such as a verb) but the English does not make grammatical/logical sense	O(4) if the thing omitted is limited in scope
Other Omission	Nearly half of a sentence or longer idea is omitted.	O(8) unless chunking phrases is less points
Subject/Object Confusion	Any flipping of subject/object in a chunk of text around a verb (does not include changing text to passive voice if appropriate)	MU (4)

Definiteness	Making something definite in English that is not definite in the source text –unless it NEEDS to be definite for the sake of English syntax.	
Translation of "wa" at the beginning of a sentence	Should be left out unless the text sounds awkward without connecting words. Any connecting words or phrases should be considered an addition	A (1) –for a one to three word "connecting phrase" added to a text
PS	Any word translated into the wrong part of speech	PS(2) i.e. "the wealthy" vs. "wealth'; unless it is very slight, then PS(1) –i.e. "is destroying" "destroyed"
relative clauses	misses relative clause causing information to be confused	MU (4)
"The West" (or other capitalization errors)	"The west"	C (1) -only counted once for all encounters. (2 points may be taken off if serious capitalization problem)
FOR THIS PASSAGE		
"Fifteen centuries ago"	"Before 15 centuries"	L (1)
"Fifteen centuries ago"	"Before the time period of fifteen centuries"	L(2)
"Fifteen centuries ago"	"Before _(any mistranslation of زهاء)	MT (2)
"lamented"	"sighed"	T (2)
"lamented"	"saddened"	SYN (4)
"If poverty were a man I would kill him"	confuses subject/object	MU (8)
"I would kill him"	"he"; "we"	MT (1)
"still alive"	"did not cease to live"	L (2); other variants possible
"is the product of"	"produces"; etc.	MT (4)
"policies"	"politics"	MT (1)
"poverty"	"the poor"	MT (4) if once secluded (16) –if continues throughout
"non-governmental organizations"	-all else	MT (2)
"a thing of the past"	"a past"	L (2)
"a think of the past"	"of the past"	MT (4)
"that leaves and never returns"	"that goes without returning"	L (2)
"helping peoples"	"helping its people"	MU4
"where poverty is killing their children"	"kill the poverty of their children"	MT (4) (8)

“their sick”	“their illnesses”	MT (2)
“The West exploited”	“profits off”	R (1)
“The West exploited”	a verb that does not fit English semantic rules but IT DOES CONNOTE a negative action	MU/MT (4)
“The West exploited”	The West helps” –or any other positive verb	MU (4)
“countries”	“role”	MT (4)
“It is the one that is getting rich by trading”	Confuses subject and object in various ways	MU (4) –two part confusion may call for MU(8)
in their markets	In its markets	MT (2)
“consume”	“strive”, etc.	MT (4)
“is helping curb poverty” كبح جماح	any mistranslation of جماح as an idea separate from “curbing.” The words are a phrase together in Arabic.	MU/MT (4)
“perspective of self interest”	“from awesome freedom” or	MU(8) ; if less severe MU/MT (4)
“population movement”	“people movement”	MT (2) ; more severe MT (4)
“that numerous Western nations refuse”	confuses subject and object	MU(4)
“that numerous Western nations refuse”	“that Western nations usually refuse”	MT (2)
“which their people see as a danger”	confuses subject and object	MU (4)
“The alleviation of the severity of poverty”	“the alleviation of poverty”	O(2)
“The alleviation of the severity of poverty”	confuses subject and object	MU (4)
“These steps are not limited to”	confuses subject and object	MU (4)
“providing”	“producing”	MT (2)
“which is required to launch successful economic projects”	confuses subject and object	MU (4)
“and the lag in their competitive abilities”	“and the difference in their competitive abilities”	T (1)

Passage Specific Guidelines: "Revtest"

- It should sound like an article originally written in English, free of cultural, lexical, and grammatical errors.
- Your translation should not assume that the reader has lived in the Middle East or knows much about the language or the area.
- The content of the source text should correspond closely with the content of the target text, although sentences and phrases may be divided, merged or otherwise altered to maintain English flow.
- English equivalents of Arabic metaphors and colloquialisms may be substituted. Names should be transliterated unless otherwise stated.
- An appropriate register for English journalism should be maintained throughout the text.
- Preferred translations of certain words:
    1. يوم الدين "Judgment Day" or "Day of Judgment"

(FOR THIS PASSAGE) Model	Projected Errors	Suggested Method for Scoring
"On the domestic level"	"On the portion of Upper Egypt"	MT (4)
"suffered"	any mistranslation not close to suffer	MT(4) if only once MT (16) if continues throughout
"more than anyone else"	"bigger than anyone else"	L (4)
"pluralism"	"diversity"	T (2)
"pluralism"	anything else that does not connote multiplicity	MT (4)
"capabilities"	"abilities"	WF (1)
"And it became their right"	"And it came from their right"	L (4)
"in the face of all this"	"to face this"	MU (4) –sub/obj. confusion
"and pursue"	"and discuss"	MT (4)
"and a respectable living"	"and a noble living"	T (2)
"under a regime"	"in shadow of a"	T (1)
"that they freely choose"	misses relative clause causing information to be confused/confuses subject and object	MU (4)
"used to humiliate"	misses "used to " idea one or both times	MT (4)
"Arab nation"	"nation"	O (4) whether done once or throughout
"but what is no less important"	"but that does not state the" importance"	MT 4 (of 'less')
"for all Arabs"	"for the Arab community"	MT (1)
"Egypt is not the"	"Egypt was not the"	MT (1) only count once if continues throughout second paragraph

“repression, corruption, poverty and an absence of freedom”	“repression, an absence of freedom, corruption and poverty”	SYN (1)
“subjected decision making”	“pledged”; “pawned”; etc.	MT/R 2
“foreign nations”	“leave”; “outside”; etc.	MT (4)
“Behind it stands many regimes”	“from behind it follow many regimes”	L (2)
“a prelude”	“advancement” etc.	MT (4)
“as a whole”	“with its families”	MT (4)
“will break forth”	“make prominent”	T (2)
“a new dawn will break forth”	confusing subject and object	MU (4)
“God willing”	“with the blessing of God”; “thank god”	L (1)
“After the revolution of the honorable Egyptians”	“After the honorable revolution of the Egyptians”	MU/F (2)
“being enslaved by a ruler or being controlled by groups that”	confuses subject and object	MU (4) ; make need to be MU (8) if continues
“decision-making”	“decisions”	MT (2)
“strength”	“support”; “nourishment”	T (1) ; unless completely off then, MT (4)
“be silent”	“live”	MT (4)
“God grants its victory over”	“God opens between”	MT/L (4)
“its persecutors”	“those who persecute/oppress it”	L (2); completely off, then MT/MU (4)
“proud people”	“people of its fathers”; “fatherly people”	MT (4)
“martyrs”	“witnesses”	MT (2)
“who bear the banner...”	misses relative clause causing information to be confused	MU (4)
“the banner”	“the soul”	MT (4)

Error Type Frequency Chart: For Full Translation Test Responses

Error Type	Grader 1	Error Type	Grader 2	Error Type	Both
Sentence Cap	101	Sentence Cap	95	Sentence Cap	196
High					
Mistranslation (MT)	268	Mistranslation (MT)	233	Mistranslation (MT)	501
Terminology/Word Choice T/WC	114	Misunderstanding (MU)	82	Terminology/Word Choice T/WC	171
Omission (O)	70	Terminology/Word Choice T/WC	57	Misunderstanding (MU)	129
Word Form (WF)	52	Faithfulness (F)	55	Omission (O)	116
Misunderstanding (MU)	47	Omission (O)	46	Faithfulness (F)	77
Mid					
Syntax (SYN)	27	Literalness (L)	45	Syntax (SYN)	70
Addition (A)	22	Syntax (SYN)	43	Literalness (L)	59
Faithfulness (F)	22	Capitalization (C)	23	Word Form (WF)	58
Capitalization (C)	21	Addition (A)	22	Addition (A)	44
Grammar (G)	16	Grammar (G)	15	Capitalization (C)	44
Punctuation (P)	16	Style (ST)	15	Grammar (G)	31
Literalness (L)	14	Register (R)	13	Usage (U)	22
Usage (U)	13	Usage (U)	9	Register (R)	20
Part of Speech/ Word Form (PS/WF)	9	Word Form (WF)	6	Style (ST)	18
Low					
Register (R)	7	Spelling (Spelling)	5	Punctuation (P)	16
Spelling (Spelling)	7	Indecision (IND)	4	Spelling (Spelling)	12
Ambiguity (Amb)	5	Ambiguity (Amb)	0	Part of Speech/ Word Form (PS/WF)	9
Style (ST)	3	Part of Speech/ Word Form (PS/WF)	0	Indecision (IND)	6
Indecision (IND)	2	Punctuation (P)	0	Ambiguity (Amb)	5

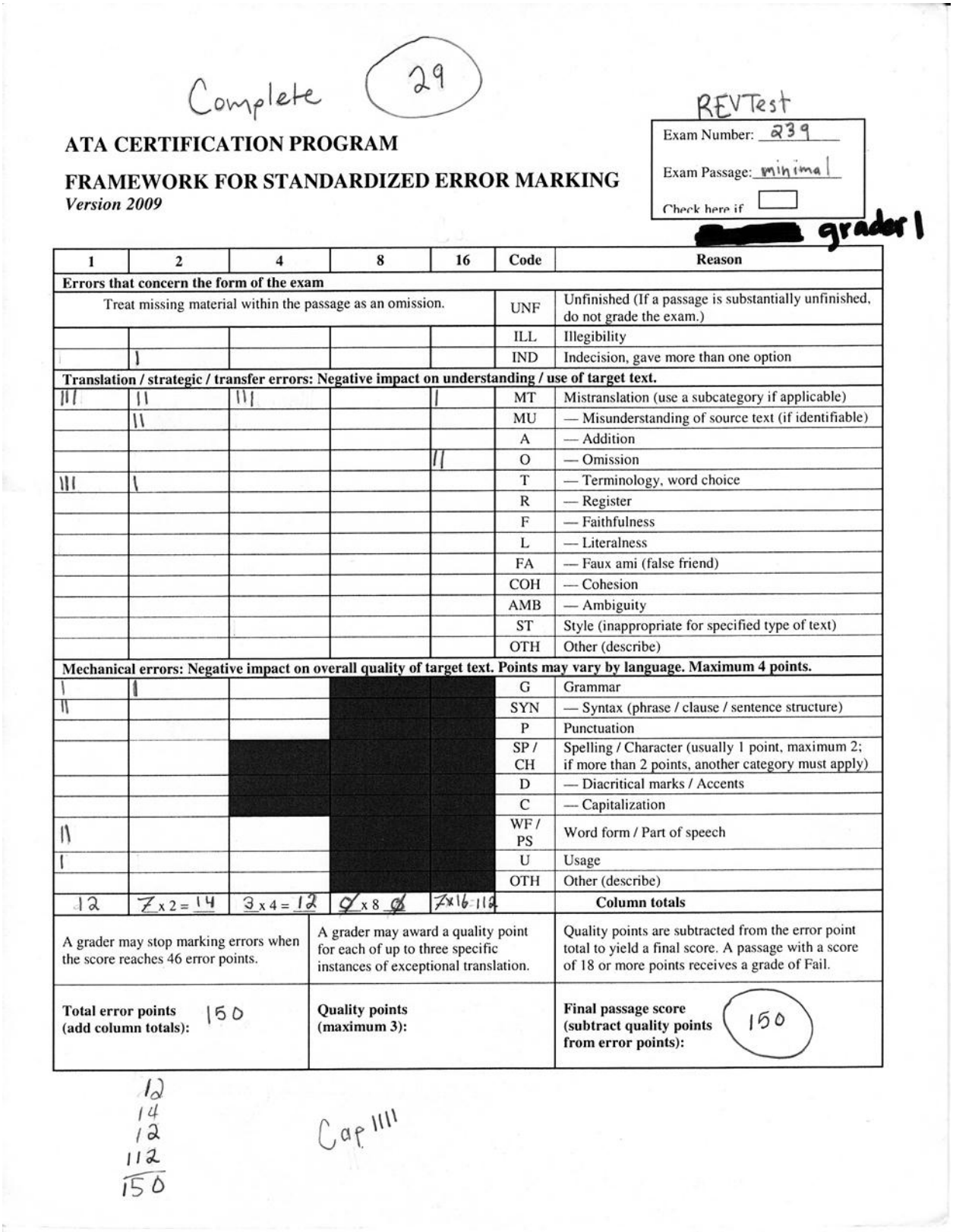
Note: Error type categories are taken from the ATA Framework for Standardized Error Marking, version 2009

Sample Error and Responses for "Povtest" and "Revtest"

SOURCE: "Povtest"	قبل زهاء خمسة عشر قرناً تحسّر الخليفة الراشد عمر انه…		
Faulty Translation:	In the fifteenth century, the Caliph Omar lamented…		
Model Translation:	Fifteen centuries ago the Caliph Omar lamented…		
Response 1 (Detect-and-correct):	"Before the fifteenth century, the Caliph…"	Response 4: (Full Translation)	"Before the brilliance of the 15th century the Caliph…"
Response 2 (Detect-and-correct):	"Before the splendor of the fifteenth century, the Caliph…"	Response 5: (Full Translation)	"About fifteen centuries ago, the Caliph…"
Response 3 (Detect-and-correct):	"Before the dawning of the fifteenth century, the Caliph…"	Response 6: (Full Translation)	"In front of roughly fifteen companions, the Caliph…"

SOURCE: "Revtest"	فمن ورائه تصطف أنظمة كثيرة، ما يعني أن نجاح الثورة المصرية سيكون مقدمة لتحرير الأمة بأسره.		
Faulty Translation:	…which means that the success of the Egyptian revolution will be a prelude to the liberation of the Arab nation with its families.		
Model Translation:	…which means that the success of the Egyptian revolution will be a prelude to the liberation of the Arab nation as a whole.		
Response 1 (Detect-and-correct):	…the liberation of the Arab nation and its people.	Response 4: (Full Translation)	"…the freedom of the Islamic Nation and her families…"
Response 2 (Detect-and-correct):	…the liberation of the Arab nation from its captivity	Response 5: (Full Translation)	"…the freeing of the nation from its bonds…"
Response 3 (Detect-and-correct):	the liberation of the entire Arab nation...	Response 6: (Full Translation)	"…the liberation of the Islamic community in full…"

Sample Scoring for Grader 1: ATA Framework



Complete 29

REVTest

ATA CERTIFICATION PROGRAM

FRAMEWORK FOR STANDARDIZED ERROR MARKING
Version 2009

Exam Number: 239
Exam Passage: minima 1
Check here if [redacted] grader 1

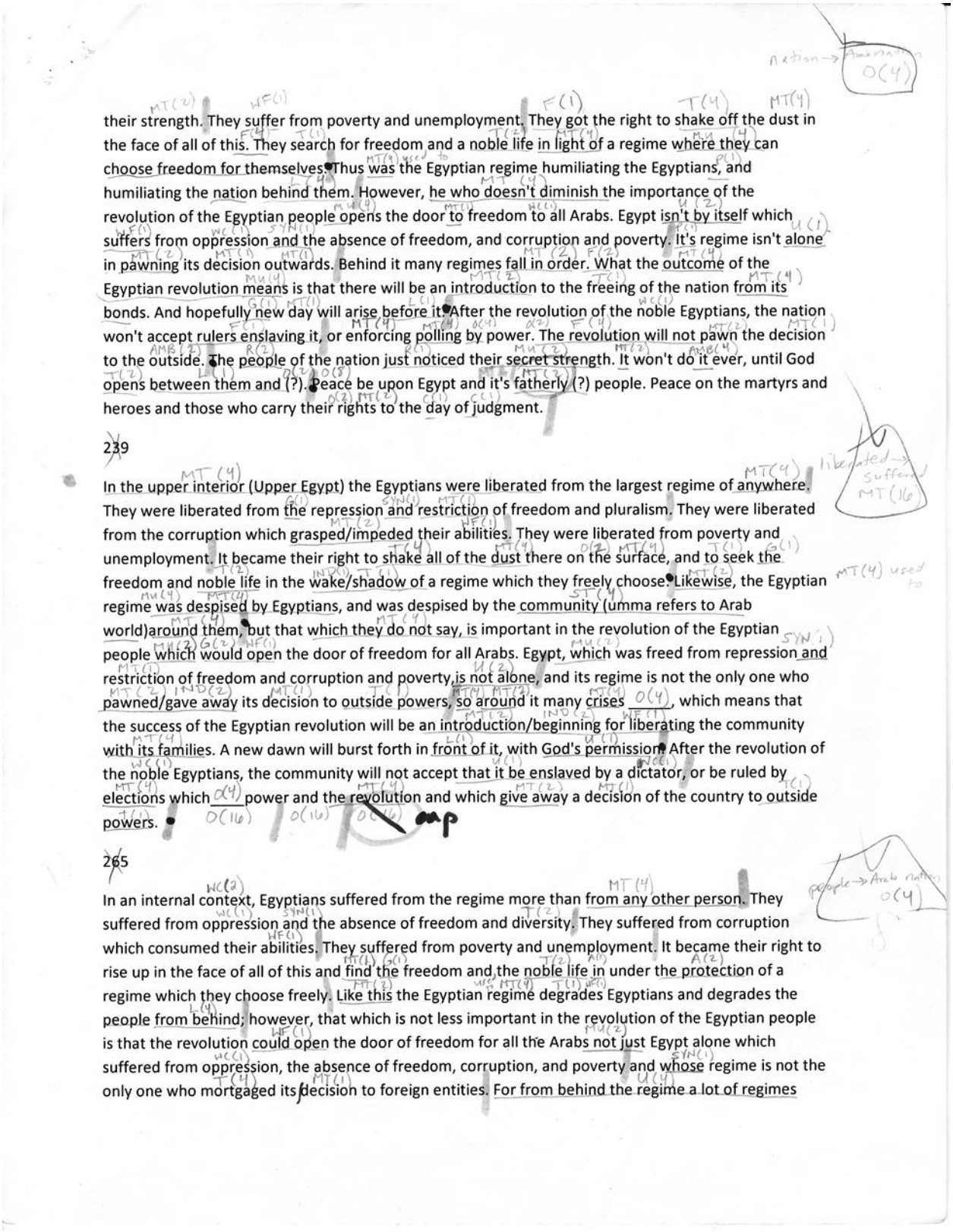
1	2	4	8	16	Code	Reason
Errors that concern the form of the exam						
Treat missing material within the passage as an omission.					UNF	Unfinished (If a passage is substantially unfinished, do not grade the exam.)
					ILL	Illegibility
	1				IND	Indecision, gave more than one option
Translation / strategic / transfer errors: Negative impact on understanding / use of target text.						
III	II	III		1	MT	Mistranslation (use a subcategory if applicable)
	II				MU	— Misunderstanding of source text (if identifiable)
					A	— Addition
				11	O	— Omission
III	1				T	— Terminology, word choice
					R	— Register
					F	— Faithfulness
					L	— Literalness
					FA	— Faux ami (false friend)
					COH	— Cohesion
					AMB	— Ambiguity
					ST	Style (inappropriate for specified type of text)
					OTH	Other (describe)
Mechanical errors: Negative impact on overall quality of target text. Points may vary by language. Maximum 4 points.						
1	1				G	Grammar
II					SYN	— Syntax (phrase / clause / sentence structure)
					P	Punctuation
					SP / CH	Spelling / Character (usually 1 point, maximum 2; if more than 2 points, another category must apply)
					D	— Diacritical marks / Accents
					C	— Capitalization
II					WF / PS	Word form / Part of speech
1					U	Usage
					OTH	Other (describe)
12	7 x 2 = 14	3 x 4 = 12	0 x 8 0	7x16 112		**Column totals**

A grader may stop marking errors when the score reaches 46 error points.	A grader may award a quality point for each of up to three specific instances of exceptional translation.	Quality points are subtracted from the error point total to yield a final score. A passage with a score of 18 or more points receives a grade of Fail.
Total error points (add column totals): 150	**Quality points (maximum 3):**	**Final passage score (subtract quality points from error points):** 150

12
14
12
112
150

Cap IIII

Sample Scoring for Grader 1



their strength. They suffer from poverty and unemployment. They got the right to shake off the dust in the face of all of this. They search for freedom and a noble life in light of a regime where they can choose freedom for themselves. Thus was the Egyptian regime humiliating the Egyptians, and humiliating the nation behind them. However, he who doesn't diminish the importance of the revolution of the Egyptian people opens the door to freedom to all Arabs. Egypt isn't by itself which suffers from oppression and the absence of freedom, and corruption and poverty. It's regime isn't alone in pawning its decision outwards. Behind it many regimes fall in order. What the outcome of the Egyptian revolution means is that there will be an introduction to the freeing of the nation from its bonds. And hopefully new day will arise before it. After the revolution of the noble Egyptians, the nation won't accept rulers enslaving it, or enforcing polling by power. The revolution will not pawn the decision to the outside. The people of the nation just noticed their secret strength. It won't do it ever, until God opens between them and (?). Peace be upon Egypt and it's fatherly (?) people. Peace on the martyrs and heroes and those who carry their rights to the day of judgment.

239

In the upper interior (Upper Egypt) the Egyptians were liberated from the largest regime of anywhere. They were liberated from the repression and restriction of freedom and pluralism. They were liberated from the corruption which grasped/impeded their abilities. They were liberated from poverty and unemployment. It became their right to shake all of the dust there on the surface, and to seek the freedom and noble life in the wake/shadow of a regime which they freely choose. Likewise, the Egyptian regime was despised by Egyptians, and was despised by the community (umma refers to Arab world)around them, but that which they do not say, is important in the revolution of the Egyptian people which would open the door of freedom for all Arabs. Egypt, which was freed from repression and restriction of freedom and corruption and poverty,is not alone, and its regime is not the only one who pawned/gave away its decision to outside powers, so around it many crises O(4), which means that the success of the Egyptian revolution will be an introduction/beginning for liberating the community with its families. A new dawn will burst forth in front of it, with God's permission. After the revolution of the noble Egyptians, the community will not accept that it be enslaved by a dictator, or be ruled by elections which O(4) power and the revolution and which give away a decision of the country to outside powers.

265

In an internal context, Egyptians suffered from the regime more than from any other person. They suffered from oppression and the absence of freedom and diversity. They suffered from corruption which consumed their abilities. They suffered from poverty and unemployment. It became their right to rise up in the face of all of this and find the freedom and the noble life in under the protection of a regime which they choose freely. Like this the Egyptian regime degrades Egyptians and degrades the people from behind; however, that which is not less important in the revolution of the Egyptian people is that the revolution could open the door of freedom for all the Arabs not just Egypt alone which suffered from oppression, the absence of freedom, corruption, and poverty and whose regime is not the only one who mortgaged its decision to foreign entities. For from behind the regime a lot of regimes

Sample Scoring for Grader 2: ATA Framework

Total Errors 31

Retest

ATA CERTIFICATION PROGRAM

FRAMEWORK FOR STANDARDIZED ERROR MARKING
Version 2009

Exam.Number: 239
Exam Passage: —
Check here if ☐

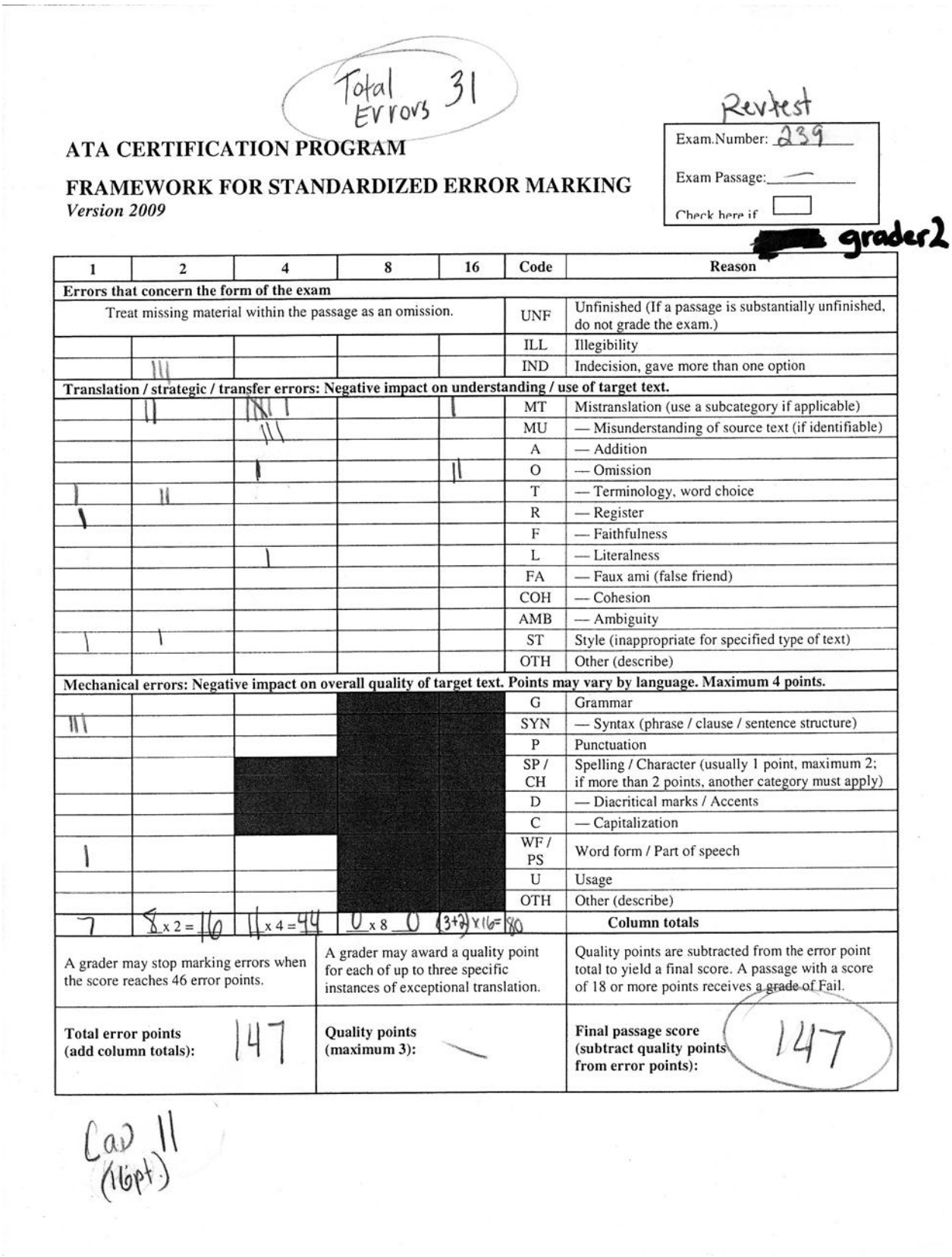
[redacted] grader2

1	2	4	8	16	Code	Reason
Errors that concern the form of the exam						
Treat missing material within the passage as an omission.					UNF	Unfinished (If a passage is substantially unfinished, do not grade the exam.)
					ILL	Illegibility
	III				IND	Indecision, gave more than one option
Translation / strategic / transfer errors: Negative impact on understanding / use of target text.						
	II	IIII I		I	MT	Mistranslation (use a subcategory if applicable)
		III			MU	— Misunderstanding of source text (if identifiable)
					A	— Addition
		I		II	O	— Omission
I	II				T	— Terminology, word choice
I					R	— Register
					F	— Faithfulness
		I			L	— Literalness
					FA	— Faux ami (false friend)
					COH	— Cohesion
					AMB	— Ambiguity
I	I				ST	Style (inappropriate for specified type of text)
					OTH	Other (describe)
Mechanical errors: Negative impact on overall quality of target text. Points may vary by language. Maximum 4 points.						
III			[redacted]		G	Grammar
					SYN	— Syntax (phrase / clause / sentence structure)
					P	Punctuation
		[redacted]			SP / CH	Spelling / Character (usually 1 point, maximum 2; if more than 2 points, another category must apply)
					D	— Diacritical marks / Accents
					C	— Capitalization
I					WF / PS	Word form / Part of speech
					U	Usage
					OTH	Other (describe)
7	8 x 2 = 16	11 x 4 = 44	0 x 8 = 0	(3+7) x 16 = 80		**Column totals**

A grader may stop marking errors when the score reaches 46 error points.	A grader may award a quality point for each of up to three specific instances of exceptional translation.	Quality points are subtracted from the error point total to yield a final score. A passage with a score of 18 or more points receives a grade of Fail.
Total error points (add column totals): 147	**Quality points (maximum 3):** —	**Final passage score (subtract quality points from error points):** 147

Cap II (16pt.)

Sample Scoring for Grader 2



Total Errors 31

Retest

ATA CERTIFICATION PROGRAM

FRAMEWORK FOR STANDARDIZED ERROR MARKING
Version 2009

Exam.Number: 239

Exam Passage: —

Check here if ☐

[redacted] grader2

1	2	4	8	16	Code	Reason
Errors that concern the form of the exam						
Treat missing material within the passage as an omission.					UNF	Unfinished (If a passage is substantially unfinished, do not grade the exam.)
					ILL	Illegibility
	111				IND	Indecision, gave more than one option
Translation / strategic / transfer errors: Negative impact on understanding / use of target text.						
	11	1111		1	MT	Mistranslation (use a subcategory if applicable)
		111			MU	— Misunderstanding of source text (if identifiable)
					A	— Addition
		1		11	O	— Omission
1	11				T	— Terminology, word choice
1					R	— Register
					F	— Faithfulness
		1			L	— Literalness
					FA	— Faux ami (false friend)
					COH	— Cohesion
					AMB	— Ambiguity
1	1				ST	Style (inappropriate for specified type of text)
					OTH	Other (describe)
Mechanical errors: Negative impact on overall quality of target text. Points may vary by language. Maximum 4 points.						
			[redacted]		G	Grammar
111					SYN	— Syntax (phrase / clause / sentence structure)
					P	Punctuation
		[redacted]			SP / CH	Spelling / Character (usually 1 point, maximum 2; if more than 2 points, another category must apply)
					D	— Diacritical marks / Accents
					C	— Capitalization
1					WF / PS	Word form / Part of speech
					U	Usage
					OTH	Other (describe)
7	8 x 2 = 16	11 x 4 = 44	0 x 8 = 0	(3+7) x 16 = 80		**Column totals**

A grader may stop marking errors when the score reaches 46 error points.	A grader may award a quality point for each of up to three specific instances of exceptional translation.	Quality points are subtracted from the error point total to yield a final score. A passage with a score of 18 or more points receives a grade of Fail.
Total error points (add column totals): 147	**Quality points (maximum 3):** —	**Final passage score (subtract quality points from error points):** 147

Cap 11 (16pt.)