JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Mariana, V., Cox, T. & Melby, A. (2015). The Multidimensional Quality Metric (MQM) Framework: a new framework for translation quality assessment. *The Journal of Specialised Translation, 23*, 137-161. https://doi.org/10.26034/cm.jostrans.2015.343

The Multidimensional Quality Metrics (MQM) Framework: a new framework for translation quality assessment

Valerie Mariana, Troy Cox, Alan Melby, Brigham Young University, Provo

ABSTRACT

Determining translation quality requires a precise measure of the traits being examined. This article evaluates a new framework for translation quality evaluation, Multidimensional Quality Metrics (MQM), to the task of grading student translations. It demonstrates the viability (i.e. the practicality, validity and reliability) of using the MQM framework by novice raters to judge translations based on the American Translators Association's translator certification exam grading system. The data gathered for this study were drawn from 29 student translations of a single news story that were rated by nine novice and two expert raters. The study used average time on task, correlations between novices and experts and Many-Facet Rasch Measurement to identify the extent to which this use of the MQM framework was viable. The findings indicate that this implementation of MQM can be viable with novice raters under the right conditions.

KEYWORDS

Translation quality, translation evaluation, translation assessment, Multidimensional Quality Metrics, MQM, practicality, validity, reliability.

1. Introduction

Determining translation quality requires a precise measure of the traits being examined. Multidimensional Quality Metrics (MQM) is unique to the translation community because:

1. It is a comprehensive framework for developing translation quality assessment metrics;
2. All metrics developed within the framework draw on the same hierarchy of error categories, making it easier to discuss and compare different studies;
3. It is customisable to the user's needs;
4. Each metric is tied to a set of specifications, which are vital to determining the standards for quality; and
5. The MQM framework is available online at no cost.

Due to its availability as a free resource, it is likely to attract novice raters in translation quality assessment. Their fledgling use of the framework viably would be strong evidence that experienced raters would only improve in the reliable and valid measure of translation quality. This article introduces MQM, highlights the importance of establishing the reliability and validity of translation quality assessment metrics, and presents a study that evaluates the viability (practicality, reliability and validity) of using the framework, customised to reflect the ATA's translator certification exam, with novice raters judging the quality of student translations.

There are several methods for evaluating the quality of a translation, including the holistic method and the analytic method. The holistic method focuses on the text as a whole, while the analytic method focuses on segments of a text, from the paragraph-level to the word-level (see glossary for further detail). Both aim to determine whether a text is a quality translation relative to appropriate specifications. The MQM framework allows the user to create a translation quality assessment tool that can fall under either method. That is, the framework presents a variety of error categories that can be drawn on to create customised metrics *based* on the end user's needs, and those error categories can be used to evaluate the text as a whole (holistic method) or on a sentence-by-sentence basis, as this study did (analytic method).

This study examined MQM operating as an analytic method and aimed to determine whether the MQM framework gives viable results (which, for this study we will define as results which are valid, reliable, and practical) in an educational or translation testing environment that uses the error categories of the American Translators Association (ATA) translator certification exam.

Certification exams are typically criterion-referenced and in this case the analysis tool created using the MQM framework was criterion-referenced. However, the MQM framework can be used in developing either criterion-referenced or norm-referenced exams (see Appendix 1 for an explanation of these two types of exams).

In this study, 29 student translations of a short French newspaper article about a schoolteacher (*Nouvel Observateur* 2009) have been acquired and rated with the tool we created using the MQM framework.

1.1 Research Question

This study aims to answer the following question:

- To what extent is the MQM framework for rating translations viable (practical, reliable and valid) when it is operationalised based on the test architecture laid out by the ATA certification exam?

Originally, this study aimed to also examine an application of MQM based on Pre-Selected Items Evaluation, or PIE, a system designed by Hendrik Kockaert and Winibert Segers which is suitable for formative assessment in translator training (Kockaert and Segers forthcoming). Some of the raters used in this study were trained and began rating translations with the PIE tool before they rated using the MQM tool, which may have influenced their MQM ratings.

The ATA certification system is described in detail in Koby and Champe (2013). Defining and justifying thresholds of acceptability in grading translations is very important but is beyond the scope of this study.

2. Background

MQM was recently created by the Quality Translation Launch Pad group (QTLP 2013). It was based on many other quality evaluation tools, and most heavily draws from the LISA QA model (which was developed by the now defunct Localization Industry Standards Association), a model which is often used in a modified form. It was designed to be applicable to a professional production environment, (the translation industry, where translations are produced for pay) as well as a testing environment. It is important to note that in a translation testing environment it is appropriate and even necessary to have a reference translation, whereas in a production environment, it is rare to have a reference translation, as there is no need to translate a document for a client when there is already an acceptable translation.

The MQM website presents a list of possible error categories and their definitions (QTL, definition). The MQM error categories are arranged hierarchically, as pictured below in Figure 1.1.

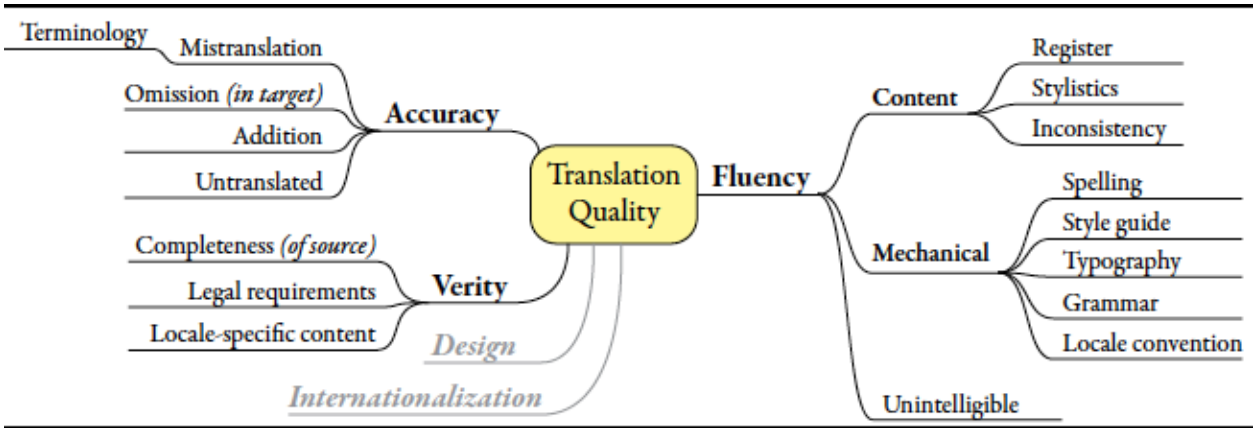

Figure 1.1: A tree graph of the hierarchy of the MQM error categories from the website http://www.qt21.eu/mqm-definition/ (Section 5.2 MQM Core).

Another description of the MQM error categories, along with an explanation of error severity, both of which were given to the novice raters who participated in this study, is found below in Figure 1.2. A description of the difference between minor, major and critical errors is found in Sharon O'Brien's article "Towards a Dynamic Quality Evaluation Model for Translation" (O'Brien 2012: 62). O'Brien states that minor errors "are noticeable but … do not have a negative impact on meaning… [,m]ajor errors … have a negative impact on meaning … [and] critical errors … have major effects not only on meaning, but on product usability…" (O'Brien 2012: 62).

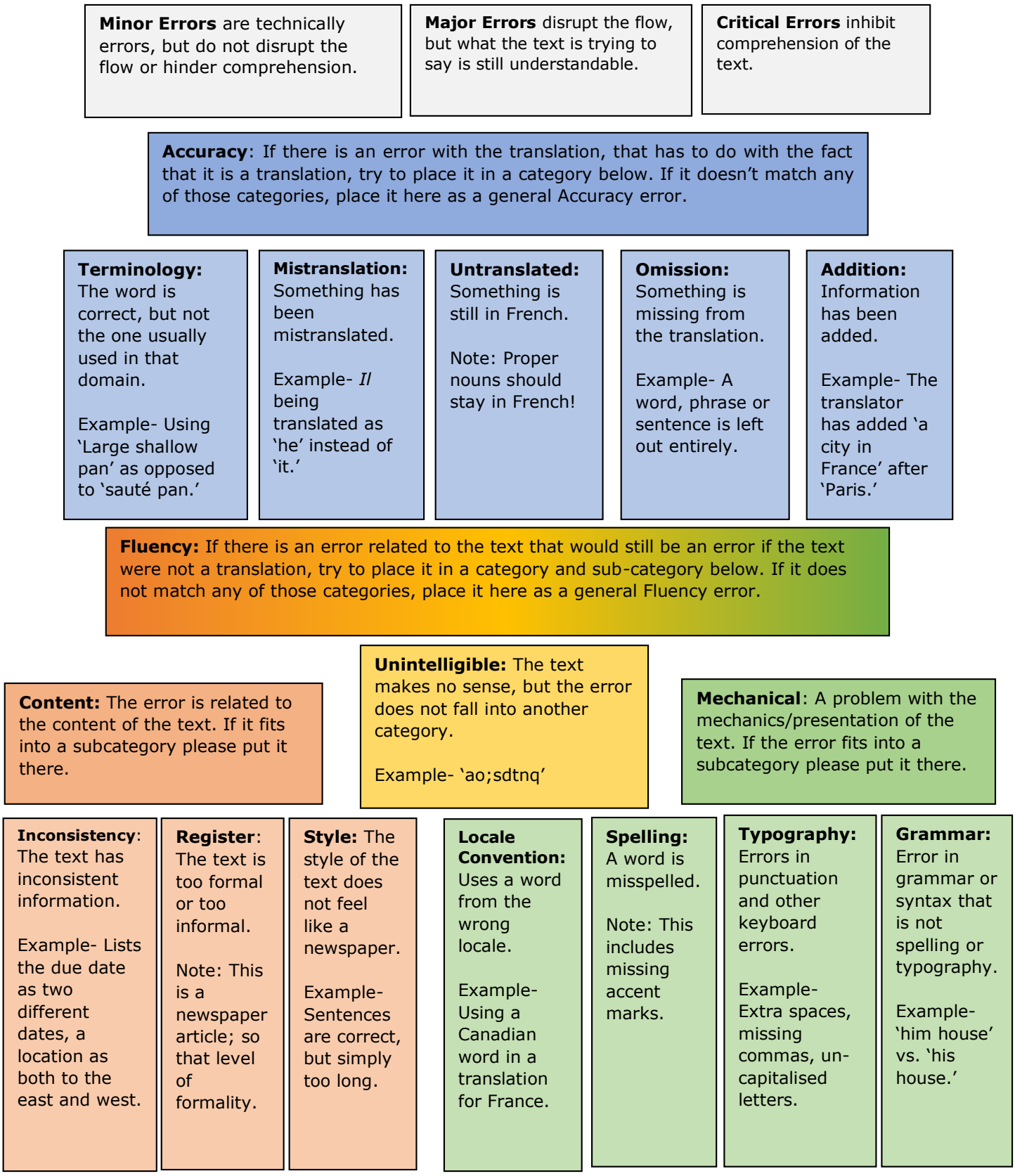

Figure 1.2 Chart explaining error severity and error categories used to aid raters in applying the MQM method.

In order to operationalise the MQM framework based on the ATA's translator certification exam, we needed to discuss how ATA raters measure quality. ATA raters refer to a mapping of error categories (see Table 1) when evaluating the quality of the translations. When an error is found, the type is identified; and the severity is determined (Koby and Champe 2013). For example, if there were five errors in a 100 word text: two minor accuracy/grammatical errors (one point each) and three moderate fluency/cohesion errors (two points each), the final score would be eight. Any score above 18 would be a failing grade (Koby and Champe 2013: 166).

To equate ATA to MQM, the ATA categories were matched to the error categories available in the MQM framework. This mapping, which is seen in Table 1, was created by Geoffrey Koby (Chair of the ATA Certification Committee), Arle Lommel and Alan Melby. It is similar to other such mappings found under section 10, "Mappings of existing metrics to MQM (non-normative)" (QTLP, section 10). The online MQM scorecard creation software available on the site (QTLP, metric builder) was used to create a customised scorecard which used the error categories from the ATA mapping, and was used to identify all the errors present in a translation and classify them by category as well as by severity. The scorecard was designed to reflect ATA grading standards as closely as possible. Not all of the error categories suggested by the MQM site were used due to the fact that some did not map to any ATA category. In addition, only 22 of the 24 ATA categories could be adequately represented by the MQM framework. These unmappable categories are the 'Other' category, which is essentially a catchall category, and 'Usage' which corresponds with multiple categories in MQM but no single exact category. It was deemed appropriate to leave these categories out since errors that would fall under the Usage category in the ATA error system would still be counted, just as part of various other categories. Although there is an Other category in MQM there is no need to include this category, since errors that do not fit into a specific subcategory can be placed into a more general category such as 'general fluency,' if the error has to do with the transfer from the source to the target, or 'general accuracy,' if the error has to do with the target language mechanics. Once an error is matched to a category the rater determines if the error is minor, major or critical.

ATA category	**MQM category**
Unfinished	Accuracy: Untranslated
Illegibility	Fluency: Unintelligible
Indecision (i.e., the translation included more than one translation)	Accuracy
Mistranslation	Accuracy: Mistranslation
Misunderstanding of source text	Accuracy: Mistranslation
Addition	Accuracy: Addition
Omission	Accuracy: Omission
Terminology, word choice	Accuracy: Terminology
Register	Fluency: Register
Faithfulness	Accuracy
Literalness	Accuracy: Mistranslation
Faux ami (false friend)	Accuracy: Mistranslation
Cohesion	Fluency: Inconsistency
Ambiguity	Fluency: Ambiguity
Style (inappropriate for specified type of text)	Fluency: Style
Grammar	Fluency: Grammar
Syntax (phrase/clause/sentence structure)	Fluency: Grammar
Punctuation	Fluency: Typography
Spelling/Character	Fluency: Spelling
Diacritical marks/Accents	Fluency: Spelling
Capitalisation	Fluency: Typography
Word form/part of speech	Fluency: Grammar
Usage	Multiple categories, no exact match
Other (describe)	No need for this category

Table 1: ATA and MQM Mapped Error Categories

Once all errors are entered into the scorecard by the rater, the scorecard program calculates the overall score of the translation based on the number of words in the source versus the target, using the following equation developed by the MQM project. Terms that were not used in this study were removed to simplify the equation.

$$TQ=100-AP-(FPT-FPS)$$

TQ = translation quality score
AP = accuracy penalties
FPT = fluency penalties for the target language
FPS = fluency penalties for the source language

In this equation, penalties are relative to the number of words in the source and target. For example, a text with only one sentence suffers much more for the omission of one word than does a text of several thousand sentences, thus the penalty for omitting a word is greater in the shorter text. More information on the calculations used in the MQM framework can be found in the scoring section of their website (QTLP, section 8).

As an example, if a text had about 100 words, if the rater marked two minor accuracy errors, the term AP would be two, since minor errors carry a

weight of one. If the rater marked three major fluency errors, the term FPT would be six, since major errors carry a weight of two (note the weights can be adjusted). If the source text were either unavailable or had no errors, FPS would be zero. Thus the translation would receive a score of 92% since:

TQ=100-2-(6-0)
TQ=92

Despite mapping the ATA's error categories onto the MQM framework, the two rating systems are inversely related. With ATA, a zero indicates the absence of any errors, whereas 100 indicates that quality in MQM. The biggest difference, however, is that ATA sets the cut score for failure at 18 points, whereas MQM has no preset judgment of pass/fail. A translation that received 18 error points from the ATA grading system (a failing score) would receive a score of 82% from the MQM scorecard. If we wanted to use a translator's MQM scorecard score to predict whether a candidate would pass when their translation was graded by the ATA method, we would simply have to set 82% or below as 'failing' with higher scores being 'passing.'

3 Methods

To determine the viability of the MQM rating framework, a group of novice raters were given the framework, were trained to use it and rated student translations. To examine practicality, the raters recorded the amount of time it took to complete each rating, and the average time was compared to the average reported time from the translation industry. To judge validity, we compared the ratings done by the novice raters to ratings done by ATA certified translators, experts in the field. Finally, to judge reliability we ran a Rasch measurement statistical analysis using the program Facets (Linacre 2013)

3.1 Source Text

Although the source text used in this study is considerably shorter than those required by ATA (144 words rather than the required 225-275 words) it mirrors the ATA's passage A, as it is a "general text written for the educated lay reader in expository or journalistic style" (Koby and Champe 2013: 161). It is appropriately nontechnical and "material specific to the culture … is common knowledge" (Koby and Champe 2013: 161). However, it may not necessarily meet the difficulty requirements for the ATA's exam. The French source text, altered from its original version, can be found below. (*Nouvel Observateur* 2009).

> Le prof "désobéisseur," sanctionné, perd 7.000 euros
>
> Alain Refalo se voit retirer 7.000 euros de son salaire pour avoir refusé d'organiser les heures d'aide personnalisée pour les élèves en difficulté. Il dit ne rien regretter et appelle « à l'insurrection des consciences. »

> L'Inspecteur d'académie de la Haute-Garonne, Michel Baglan, a décidé d'abaisser d'un échelon pendant quatre ans le salaire de ce professeur des écoles de Colomiers, a-t-on appris vendredi 24 juillet de son comité de soutien. Initiateur du mouvement des « professeurs désobéisseurs, » Alain Refalo était le premier professeur des écoles à avoir refusé d'organiser les heures d'aide stipulées par les réformes gouvernementales. Il explique que ces actions « ont permis à des milliers d'enseignants du primaire d'en montrer toute l'inefficacité tout en ayant une attitude responsable vis-à-vis des élèves en difficulté. » Il avait alors proposé d'organiser d'autres activités pour les élèves.

3.2 Specifications

Specifications are necessary for stakeholders to know what will be considered a quality translation (see Hague and Melby 2011). The idea of specifications has its roots in Skopos theory which states that a translation must be completed in relation to "a target setting… target purpose… target addressees [and] target circumstances" (Vermeer 1987). Skopos theory is further expanded by functionalism (Nord 1997), as well as Koby and Melby (2013), who build further on this foundation, affirming that a quality translation "demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs" (Koby and Melby 2013: 178). The use of specifications is compatible with the ATA grading metric, which provides the translator with a prompt along with the source text. This project analysed a subset of 29 translations from a larger project in which 59 intermediate to advanced French students each translated one of two articles based on the following specifications, which are compatible with the ATA metric:

This newspaper article is to be translated for American expatriates living in Paris. Assume the audience has a basic knowledge of French geography and customs, but does not speak French.

Language: English (United States)
Purpose: To inform on current events
Register: Semi-formal newspaper style

3.3 Sample Student Translation

Below is an example of a student translation of the original French source text. Any misspellings (such as 'refusedto' being one word) or other errors are the work of the student translator.

> The disobeying professor, sanctioned, is losing 7,000 euros
>
> Alain Refalo will have 7,000 euros taken out of his salary for having refused to organize time for individualized help for his struggling student. He says he has no regrets and is calling to "the raising of consciences."
>
> His support committee informed us on Friday, June 24, that Michel Baglan, Inspector of the Academy of Haute-Garonne (a French province), has decided to lower the salary of this professor at the schools of Colomiers by one rung for four years. Ringleader of the movement of the "disobeying professors," Alan Refalo was the first professor from the schools to have refusedto organize the one-on-one time mandated by the governmental reforms. He explains that these actions have allowed thousands of teachers from the elementary school to show the entire inefficiency of this practice, while maintaining a responsible attitude towards the students in difficulty. Il had then proposed to organize other activities for the students.

4. Ratings

For this study, ten native English speakers were recruited as novice raters from upper-level French speakers (nine university students and one high school French teacher). We classified upper-level speakers as people who were at the university 300-level or above; most were majoring or minoring in French or French teaching. However, none of them had prior formal training in translation.

Three raters were involved in a pilot of training materials. The pilot consisted of a training session, followed by the raters completing a rating of a sample translation, then a moderated discussion of their reasoning behind their ratings took place, wherein any differences were reconciled and the trainer clarified any confusion. Appropriate changes were made to the training materials to reflect the questions that were brought up in the pilot. After the completion of the pilot, raters did not confer with one another. The translation that was designated as a practice for training purposes was not included in end calculations.

To answer the question on practicality, raters were asked to record the amount of time they took to rate the translations. To answer the question of validity, these raters were all asked to rate the same translation as two ATA certified raters. To judge reliability, a Rasch measurement statistical analysis was run using the program Facets (Linacre 2013). A ten-rater rating design was created, based on a design by Eckes (Eckes 2011: 111), such that every translation would be rated by two raters and no rater would be paired with the same person twice. Each rater was assigned to rate seven-eight of the translations so that all 29 translations were rated, most by two raters. The rating design can be seen below in Figure 2. Overall, due to a rater dropping out before the ratings began, two raters completing one additional rating each (highlighted in blue), and due to one rater not finishing their last rating (highlighted in red), 61 ratings were completed using the scorecard based on the MQM framework. Note that Translation 1

is placed at the bottom of the table because two of the raters had seen it prior to training, so it could not be the Training nor Anchor translation.

Rater ->	1	2	3	4	5	6	7	8	9	10
Training (Tr#2)	X	X	X	X	X	X	X	X	X	X
Anchor (Tr#3)	X	X	X	X	X	X	X	X	X	X
Translation#4	X	X								
Translation#5	X		X							
Translation#6	X			X						
Translation#7	X				X					
Translation#8		X	X							
Translation#9		X		X						
Translation#10		X			X					
Translation#11			X	X						
Translation#12			X		X					
Translation#13				X	X					
Translation#14						X	X			
Translation#15						X		X		
Translation#16						X			X	
Translation#17						X				X
Translation#18							X	X		
Translation#19							X		X	
Translation#20							X			X
Translation#21								X	X	
Translation#22								X		X
Translation#23									X	X
Translation#24	X					X				
Translation#25		X					X			
Translation#26			X					X		
Translation#27				X					X	
Translation#28					X					X
Translation#29					X	X				
Translation#1	X	X	X					X		

Figure 2: The rating design used in this project.

As was stated earlier, initially this study was also going to gather information on the PIE method. For that reason, four raters were trained on and completed the MQM ratings first, while the other five raters completed PIE ratings before using the MQM scorecard. It has been decided to omit the PIE ratings from this paper and devote an article to them at another time. We have not detected a difference between the scores of the raters who started with MQM and those who started with PIE, but it is worth noting this possible source of error.

Some raters were trained in person, some over videoconference or telephone and others via email, depending on their physical location and their availability. After rating the sample translation, each rater gave their feedback to the trainer, who clarified any questions they had. In addition to rating the sample translation, all raters rated an anchor translation on their own. An anchor is an item that is rated by all of the raters to give a common point for comparing the raters to one another.

5. Results

This section presents the results of the analysis of the MQM ratings, which show the extent to which the results of applying the MQM framework in the manner done by this study are practical, valid, and reliable.

5.1 Practicality

Practicality is defined by Bachman and Palmer as "the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (Bachman and Palmer 1996: 36). In other words, practicality concerns whether the test to determine translation quality can be created and implemented within the constraints of the test designer's given resources.

In this study, practicality was based on time cost. All other costs were minimal, since rating materials were distributed via the Internet at no cost (QTLP, metric builder). The amount of time required by the quality assurance manager to prepare the translations for rating, to train the raters and to interpret the data were all taken into account. Finally, the time of the raters to judge each translation was calculated.

Table 2.1 illustrates the time required by a quality assurance manager to set up materials to be rated. For most of the tasks, this constitutes the set-up phase, which is considered to be a one-time cost, since it needs only be done once per source text. Formatting and uploading translations as well as data interpretation are not a one-time cost, as they must happen every time.

Task	***Hours**
Scorecard creation	0.25
Formatting and uploading translations to rater accounts	1
Training material creation	2
Rater training	1.25
Data interpretation	1
Total Time	5.5 hours

Table 2.1: Time spent preparing materials and training

*Note that this table does not include the time spent on performing the ratings themselves. The time is cumulative and represents all raters and ratings.

The number of minutes needed for each rater to rate each translation was recorded in Table 2.2. If a range of time was given (e.g. a self-reported range of seven-ten minutes) the mean average of the times was reported in this table (8.5 minutes).

Rater	**Self-reported time for rating translations via MQM (in minutes)**
Rater 1	25.0
Rater 2	10.0
Rater 3	8.5
Rater 4	Dropped out
Rater 5	10.0
Rater 6	7.0
Rater 7	13.8
Rater 8	13.0
Rater 9	15.0
Rater 10	6.5
Average time per translation	**12.1 minutes**

Table 2.2: Average time spent per rating.

According to Snow, the average time required for translation evaluation with experienced raters in the translation industry is roughly 30% of the total time required to complete the process from beginning a translation to finishing its evaluation (Snow forthcoming). In other words, if it takes an hour total for both a translator to create a translation and for a rater to rate it, the rating ought to take 30% of that hour, or about 20 minutes of the hour. Snow's industry report on translation evaluation time included the time it took to identify errors to give feedback to a translator, but not the time it took to correct the errors and make revisions.

The students whose translations were used in this study of novice raters took an average of 16.2 minutes to complete the translation. When the average rating time is added to the time it took to translate, for a total of 28.3 minutes, this operationalisation of the MQM method takes more time than the industry average, as 12.1 minutes is 42.8% of the total 28.3 minutes rather than the usual 30%. In comments, some of the novice raters reported taking less time with each subsequent rating. Note in Table 2.2 that two of the raters (six and ten) had times that would be roughly equivalent to industry standard.

Thus, while this implementation of the MQM framework may not be as practical for first time rating, it should approach the same time commitment as that reported for experienced industry raters. While beyond the purview of this study, it would be interesting to evaluate the amount of time experienced raters would use with MQM compared to the metrics they are currently using.

Some of the novice raters in this study were able to achieve a level of practicality equivalent to that seen in the industry, and over time the other novice raters would be expected to reach this threshold as well, since their rating times tended to go down over time. Thus we believe that this application of the MQM framework has potential to be just as practical as current methods used in industry.

5.2 Reliability

In their book on designing language tests, Bachman and Palmer define reliability as "a function of the consistency of scores from one set of tasks to another" (Bachman and Palmer 1996: 19). Koby and Melby add to this, saying an evaluation is reliable if "the candidate gets the same score, within a reasonable range of variation, regardless of who grades the examination" (Koby and Melby 2013: 176). This is particularly important for novice raters with very little training. If those raters can apply MQM consistently, then by extension, experts would be even more reliable.

Reliability of this application of MQM was examined via the statistics program Facets (Linacre 2013). This program was chosen because it can perform Many-Facet Rasch Measurement, and the Rasch system "provides estimates for every facet that is measured" whereas other programmes can only handle one facet and cannot compare multiple facets to one another (Evans *et al.* 2014: 38). The facets that were analysed were 'translation' and 'rater,' and the rater separation reliability was determined. Lunz and Stahl give another reason to use Rasch measurement, stating that even after "all reasonable efforts have been made to train judges, differences in judge severity [will often] still [be] observable" but fortunately Rasch measurement analysis can "account for these changes and remove their effects from examinee measures so that no examinee is unfairly penalized" (Lunz and Stahl 1990: 442).

5.2.1 Data Preparation

The scores of the ratings were automatically given a value between 0% and 100% by the online scorecard and we converted these percentages to a ten-point scale of zero to nine (a perfect score of nine was given to 95% or above, eight was awarded to 85%-94.9%, etc. and anything less than 15% was converted to a zero) in order to be statistically analysed. The software could only analyse whole numbers, so ratios were converted to a scale from zero to nine according to the software's developer (Linacre 2011).

5.2.2 Facets Analysis

An advantage to running a Many Facets Rasch Analysis is that the facets (in this case translation and rater) can be directly compared to each other using a vertical logit scale. A logit is a unit of measurement that traditionally allows the analyser to examine "candidate ability and item difficulty on the same measurement scale" (NCSBN). However, we are not limited to candidate ability and item difficulty. Rather we can examine any facets, in this case the translation and the rater, on the same scale via a logit scale. (For detailed information on the logit see Institute for Objective Measurements).

The vertical scale can be seen in Figure 3. The logit is the first column, the quality of the translation is the second column, the rater severity is in the third column, and the scale equivalency is in the fourth column. We can see from Figure 3 that the Rater severity ranged from category two to category seven on the equivalent scale. The rater separation reliability, a statistic showing whether the raters are interchangeable in terms of leniency and severity, was reported to be .97. This gives similar information to traditional inter-rater reliability, which also determines whether raters are reliable enough to be interchangeable. However, in a Rasch measurement approach if raters were the facet of interest and scored close to zero, this would indicate that they were indistinguishable from each other and therefore interchangeable but when the rater facet is not close to zero, it is necessary to adjust the score to compensate for the rater bias. In this sense, rater separation reliability is the inverse of classic inter-rater reliability, which takes a score of one to mean that raters are reliably the same. Therefore the raters in this study were reliably different and they could not necessarily be used interchangeably.

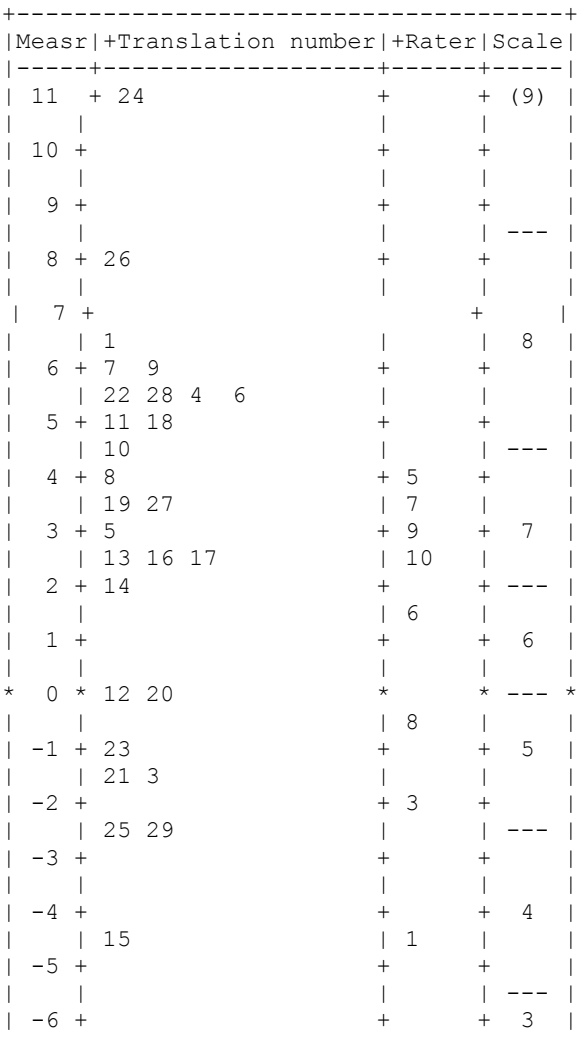

Figure 3: The vertical scale output from the Facets statistics program.

Despite the fact that the raters are not interchangeable due to differing levels of severity, an advantage of using Facets is that "the impact of raters'

severity can be examined and corrected even if the raters rate different items" (Zhu *et al.* 1998: 3). Table 3 shows statistical information about the raters. In the table, the rater is shown first. Next is the observed average, which is the mean of the scores given by the rater. In the third column we see the fair average, which is the score that has been adjusted for the rater's severity. In other words, each translation is given a fair average by the Facets program, which is the score that translation ought to have gotten if it had been rated by the average rater. The fair average score corrects for the impact that the inconsistency between raters may have had on the score.

Despite the fact that Facets can give reliable scores through the fair average, it is also important for the raters to be self-consistent. The fourth column, outfit mean square, shows the self-consistency of the raters — that is, whether the raters are using the scorecard consistently, even if they are not using it in the same way as the other raters. Ideally the outfit mean square should be between 0.5 and 1.5. Rater ten, with the outfit mean square value of 1.76, has poor self-consistency. However, the parameters defined by Linacre show that a value between 1.5 and 2.0 is not ideal but "not degrading" (Linacre n.d.). On the other hand, Rater six, who had an outfit mean square value of .33 and Rater nine with a value of .34 are unusually low. This may be indicative of collusion between the raters. However, we know that the raters did not have contact with each other after the initial training session. In any case, a value of less than 0.5 is also not degrading to the data. Thus, we conclude that this operationalisation of MQM can be applied reliably when the Facets programme is used to ensure raters are self-consistent, as well as to find a fair average score.

Regarding the question on the reliability of novice raters employing this application of the MQM framework, the findings are promising. With less than one hour of training, the raters had acceptable ranges of intra-rater reliability (outfit values between .5 and 1.5). With increased experience, raters should become even more self-consistent. As with most performance judgments, the raters were reliably different from each other. Since they were self-consistent, this means that some were reliably harsher and others were more lenient, so the final score awarded could be affected by who rated the translation. With further training, practice and experience, scoring idiosyncrasies by novices tend to diminish and raters should start to award more similar ratings. But even without being interchangeable, the raters are self-consisted enough to mathematically model a fair average, negating the effects of differences in severity, if the translations are double-rated.

Rater	Observed Average	Fair Average	Outfit MnSq	Measure
2	4.00	4.00	.95	-6.97
1	5.86	4.82	.69	-4.47
3	6.29	6.13	.96	-1.95
8	6.29	6.76	1.04	-0.67
6	6.57	7.54	.33	1.61
10	7.33	7.81	1.76	2.59
9	7.17	7.86	.34	2.82
7	7.33	7.96	.89	3.28
5	8.00	8.06	1.16	3.77

Table 3: Statistical information on the raters related to their reliability.

5.3 Validity

Validity or construct validity is "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (Bachman and Palmer 1996: 21). In other words a test is valid if it "measures what it is supposed to measure" or "you are testing what you want to test" (Shrock and Coscarelli 2007: 20, Koby and Melby 2013: 176). Take the example of shooting at a bulls-eye. A reliable marksman would shoot in the same place every time. However, these shots would not be valid unless they were in the centre of the bulls-eye. Thus, it is evident that "it is not possible for the marksman's shots to be valid without also being reliable" (Shrock and Coscarelli 2007: 21). Validity is a property of interpreting test results and not a property of the test itself. To further the target analogy, a person's ability to hit a bulls-eye would give good evidence of deer hunting in which a hunter would have time to wait and get the animal in their sights. However, it might not be as useful in predicting how well a person could hunt ducks since the target is moving quickly through the air. In that instance a more valid measurement might be shooting clay pigeons. Validity, then is not a binary outcome, but rather reflects the degree to which an instrument is measuring what it claims to measure. Both duck hunting and deer hunting require some degree of aptitude with a gun, but the bulls-eye would be a more valid measure for deer hunting proficiency than would the clay pigeons.

In a study by Christopher Waddington testing the validity of various methods for rating a translation, it was found that analytic, rubric and holistic methods (rubric was not considered a sub-method, as it is in this paper) could all be equally valid if properly constructed (Waddington 2001).

In the present study, validity for this particular operationalisation of MQM was determined via the interpretive argument method as described by Kane (2006). Kane defines validity as the extent to which scores are supported by evidence and states that propositions supporting the validity of a test ought to be made and then confirmed by corresponding evidence (Kane 2006). Our argument is that this operationalisation of the MQM framework will be considered valid if the ratings done by the novice raters agree with

the ratings of recognised experts – ATA certified French to English translators.

5.3.1 Certified Rater Discussion

The two ATA certified translators, who had prior experience grading and were familiar with the ATA grading system, used the traditional ATA grading method when they rated the anchor translation. This involves writing directly on a paper copy of the translation (or leaving comments on a digital copy) in order to identify errors in each sentence. We used the mapping in Table 1, which matches ATA error categories with those on our MQM scorecard, to transfer their ratings to the online scorecard that the novice raters had used. The scores were then automatically calculated by the scorecard.

The scores given by the ATA certified translators were statistically similar. One of them gave the translation a score of 66.9% and the other 58.3%. Although the scores given by these professionals were very similar, the types of errors they identified were not necessarily the same. A table of the errors identified by each ATA certified translator is available in Appendix 2. However, it is necessary to note that this observed inconsistency among these experts is consistent with other research on expert judges, as was found in the study by Lunz and Stahl who extensively discuss studies on this topic (Lunz and Stahl 1990).

5.3.2 Novice rater Discussion

As is stated in our validity argument, for this operationalisation of the MQM framework to be considered valid, the novice raters should be able to use the framework and with minimal training, be able to give similar scores and find the same types of errors as the experts.

When we look at the average score of the novice raters, 61.6%, compared to the average score of the professional raters, 62.6%, we see that they are very similar. We ran a t-test on the data and found that $t(7) = 2.36$, $p = .91$, meaning that there was no significant difference between the experts and the novice raters. However, novice raters have a wider degree of variability than the experts (see Figure 4). As noted in the discussion on reliability, more experience might transition the wide fluctuations of novices to become more stable as the experts were.

Thus, judging by the data we have collected, it would appear that this operationalisation of MQM can be valid as there is no significant statistical difference between the expert raters and the novice raters. Since there is a wider range between novice raters, it would be advisable to use multiple raters for each item to compensate, at least until novice scores become more stable.

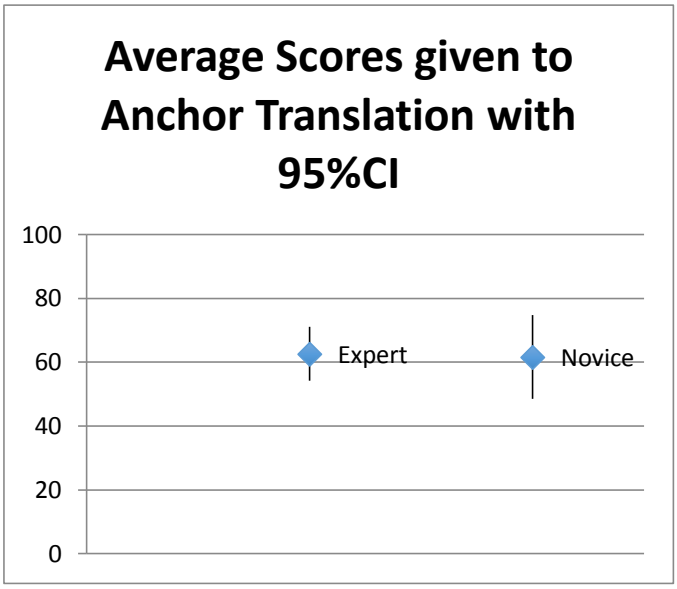

Figure 4: A graph of the average scores, including error bars, given by both the experts and novices to Translation 3, the anchor translation.

6. Discussion

The results of the analysis show that the application of the MQM framework used in this article has potential to be a practical method as far as time is involved, when compared with current translation industry standards. In addition, although raters were not interchangeable, they were self-consistent and therefore the program Facets can negate their inconsistencies and give reliable results when the fair average is used. A comparison to professional ATA certified translators also confirms that this application of MQM is valid. From these results one can conclude that this application of MQM based on the error categories of the ATA translator certification exam, can become a viable option for rating student translations.

7. Future Considerations

To further prove the viability of the MQM framework as an operationalisation of the ATA grading system it would be useful to find a text that is more appropriate to ATA standards, taking into special consideration subject matter and length. In addition, testing the method on different language pairs might be advisable.

When it comes to selecting raters, it may be useful to use expert raters instead of novice raters. It is advisable to train all of the raters in person, or at least by videoconference. It may also be prudent to train all raters at the same time and have a clarifying session, then make appropriate changes to the rating materials and train the raters again. In addition, raters could be asked to rate several anchor translations throughout the rating period, and interventions could be implemented for those raters that

deviated from the group. In this way, raters might become more consistent with each other, increasing inter-rater reliability.

In addition, a threshold for passing scores ought to be established based on pre-determined criteria, such that a translator that receives a passing score can be said to exhibit required competency or traits. This would make the test truly criterion-referenced. Although not explored in this study, the annotation of specific errors using MQM should support formative as well as summative testing. Some feel that the error categories are not sufficient for summative tests in translator training, but addressing this is beyond the scope of this article and calls for further research.

Furthermore, the application of MQM called PIE will be the subject of future study (Kockaert and Segers forthcoming).

Based on the overall positive results of this study, the authors encourage other researchers to apply the MQM framework to their own future projects, with the caution of the need to carefully train their raters or use the Facets program to ensure good reliability. Not only can the MQM framework be personalised to fit many different projects with different specifications, but all studies done using MQM will also be relatively comparable to one another since they will have the common basis of the MQM framework.

Glossary of terms

Analytic method: A way to assess the quality of a translation by looking at segments of the text, such as individual words, sentences or paragraphs and awarding or deducting points to the overall score of the translation based on whether each text unit meets certain criteria, rather than judging the overall text as a whole. In many analytic systems not only are errors counted "but also assessed or characterised, and the two most common criteria for this characterisation are nature and importance" (Conde 2011: 70). The scorecard created using the MQM framework falls under this category because the translation was scrutinised on the sentence level and every single error was counted to deduct points. In addition, as Conde suggests, the errors were characterised based on their type (or nature) and severity (or importance). The ATA grading method also falls under this category for the same reasons.

Holistic method: A means of assessing the quality of a translation by giving a score based on overall impression of the text as a whole. There may be different scoring dimensions used, such as a rubric including grammar, fluency, register, etc., but if each category is given a score based on the text as a whole, then the method is considered to be holistic.

Rubric method: A multi-dimensional way to assess the quality of a translation "[which] evaluates components of quality separately … [and is]

relative to the function and the characteristics of the audience specified for the translated text" by giving a score for each dimension of the translation (Colina 2009: 240). For example, scores for grammar, for fluency, for register, etc., are awarded and then aggregated for a total score. A rubric may be used to rate a translation holistically, in which case the score for each dimension would be given based on overall impression of the text as a whole, or it may be used analytically, where each paragraph, sentence, word, or other text unit is awarded a score in each dimension. This study's MQM scorecard was a type of rubric. Points were deducted from the appropriate rubric dimension when mistakes were detected. It is worth noting that the ATA has published a grading rubric for their translator certification program (ATA Certification program).

Appendix 1

Defining Norm-Referenced Testing versus Criterion-Referenced Testing

Norm-referenced tests

A norm-referenced test is "composed of items that will separate the scores of the test-takers from one another" in order to rank them (Shrock and Coscarelli 2007: 25). In this case, a norm-referenced test would be designed to rank a group of translators from best to worst. The success of an examinee is based on whether he or she performed better than the other examinees. There is often a set limit, determined by the test administrator, on how many people can succeed (for example, the top 50% of test takers).

Rankings are dependent on the test cohort, the group of people taking the test together, because norm-referenced tests "[define] the performance of test-takers in relation to one another" (Shrock and Coscarelli 2007: 26). Therefore it is possible for someone who is a competent translator to rank at the bottom of the cohort, effectively failing the test, if he or she has the misfortune of testing with an exceptionally talented cohort. Similarly, an examinee may not be especially good at translating, but if he or she is better than the rest of their cohort, he or she will still receive a high ranking, and a passing grade, on a norm-referenced test. For this reason, the Centre for the Study of Higher Education asserts that "Norm-referencing, on its own — and if strictly and narrowly implemented — is undoubtedly unfair" as a way for testing students in a regular classroom environment (Centre for the Study of Higher Education 2002).

Norm-referenced tests are often used as entrance exams for schools. Since a school can only accept so many students, administering a norm-referenced test to applicants ensures that the school will admit only the best students who applied that year. In fact, Eyckmans, Anckaert and Segers believe that "the method is only to be promoted for use in summative

contexts (high stakes situations where decisions have to be made)" (Eyckmans, Anckaert and Segers 2009: 87). Norm-referenced tests are more successful when a greater number of examinees participate. This means norm-referenced tests may be a good way to rate translations where multiple people must translate the same article, as in an educational or translation testing environment. However, in a professional production environment, paying multiple translators to translate the same document in order to treat the text as a norm-referenced test for rating purposes would be a colossal waste of resources.

Criterion-referenced tests

Criterion-referenced tests consist of items that are "based on specific objectives, or competency statements" (Shrock and Coscarelli 2007: 25). The items are not designed to distinguish the scores from one another, but instead determine whether or not a person can accomplish a certain objective. Therefore, in a criterion-referenced test, every examinee could get the same score and every examinee could succeed, if each of them were competent enough to perform the required tasks and thus meet the objectives. Unlike norm-referenced tests, criterion-referenced tests "[define] the performance of each test-taker without regard to the performance of the others" (Shrock and Coscarelli 2007: 28).

Criterion-referenced tests are useful for "assessing a person's ability to demonstrate a specific skill" and are therefore often used as licensing or certification exams (Shrock and Coscarelli 2007: 29). This type of test ensures that an examinee is capable of performing a required function, such as translating with an acceptable level of fluency and accuracy, rather than determining that he or she is the best out of a cohort of examinees who may or may not be translating at an acceptable level. Here criterion-referenced tests surpass norm-referenced tests because "without reference to specific competencies, what test-takers can actually do is unverifiable" in a norm-referenced test (Shrock and Coscarelli 2007: 29). Criterion-referenced tests can be used in both a professional production environment as well as a translation-testing environment. In a testing environment, a criterion-referenced test may not determine whether an examinee is better than another, but it does determine whether or not he or she can perform at a certain level or has passed a certain threshold of competency.

Appendix 2

This appendix presents Table 4 which details the errors marked by each ATA certified translator for translation three (the anchor). It is interesting to note that, although the raters highlighted many of the same issues and awarded the translation similar scores their ratings are not interchangeable.

Segment	Expert Rater 1	Expert Rater 2
1	Terminology (6 pts)	Terminology (2 pts)
		Syntax (2 pts)
		Word Form (2 pts)
		Terminology (2 pts)
2	Literalness (1 pt)	Literalness (1 pt)
	Misunderstanding/Terminology (2 pts)	Style (1 pt)
	Grammar (1 pt)	Terminology (2 pts)
	Mistranslation (1pt)	Literalness (1 pt)
3	Misunderstanding (2 pts)	Punctuation (1 pt)
	Punctuation (1 pt)	MU/MT (4 pts)
4	Capitalisation (1 pt)	Literalness/Terminology (2 pts)
	Omission (4pts)	Punctuation (1 pt)
	Misunderstanding (6 pts)	Misunderstanding/Mistranslation (4 pts)
	Spelling (2 pts)	Grammar (1 pt)
	Usage (1 pt)	Spelling (1 pt)
	Faithfulness (2 pts)	
	Mistranslation (2 pts)	
5	Terminology (2 pts)	Misunderstanding/Mistranslation (4 pts)
	Addition (1 pt)	Syntax (2 pts)
		Terminology (2 pts)
		Addition (1 pt)
6	Mistranslation (2 pts)	Misunderstanding/Mistranslation/ Cohesion (8 pts)
	Misunderstanding (4 pts)	Terminology (1 pt)
	Punctuation (1 pt)	Terminology (2 pts)
		Misunderstanding/Mistranslation (4 pts)
7	-	Literalness (1 pt)
		Punctuation (1 pt)

Table 4: Error types identified by the expert raters.

Bibliography

- **Bachman, Lyle F. and Adrian S. Palmer** (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

- **Conde, Tomas** (2011). "Translation Evaluation on the Surface of Texts: A Preliminary Analysis." *The Journal of Specialised Translation* 15, 69-86. http://jostrans.org/issue15/art_conde.php (consulted 16.11.2014)

- **Colina, Sonia** (2009). "Further Evidence for a Functionalist Approach to Translation Quality Evaluation." *Target: International Journal on Translation Studies* 21 (2), 235-64.
  Doherty, Stephen (2013). "Translation Quality Models and Tools – Is There Room for Improvement?". Globalization and Localization Association blog, July 18, http://www.gala-global.org/blog/2013/translation-quality-models-and-tools-is-there-room-for-improvement/ (consulted 16.11.2014)

- **Eckes, Thomas** (2011). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt Am Main: Peter Lang.

- **Evans, Norman W., K. James Hartshorn, Troy L. Cox, and Teresa Martin De Jel** (2014). "Measuring Written Linguistic Accuracy with Weighted Clause Ratios: A Question of Validity." *Journal of Second Language Writing* (24), 33-50.

- **Eyckmans, June, Philippe Anckaert, and Winibert Segers** (2009). "The Perks of Norm-referenced Translation Evaluation." Claudia V. Angelelli and Holly E. Jacobson (eds). *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue Between Research and Practice*. Amsterdam: John Benjamins, 73-93.

- **Kane, M. T.** (2006). "Validation." R. L. Brennan (ed.) *Educational Measurement (4th edition)*. Westport, CT: American Council on Education and Praeger.

- **Koby, Geoffrey S. and Alan K. Melby** (2013). "Certification and Job Task Analysis (JTA): Establishing Validity of Translator Certification Examinations." *The International Journal of Translation and Interpreting Research* (5)1, 174-210.

- **Koby, Geoffrey S., Alan K. Melby, and Arle Lommel** (2013). "Mapping of MQM to ATA." (unpublished, developed through a conference call).

- **Koby, Geoffrey, and Gertrud Champe** (2013). "Welcome to the Real World: Professional-Level Translator Certification." *The International Journal of Translation and Interpreting Research* (5)1, 156-173.

- **Kockaert, Hendrik J. and Winibert Segers** (forthcoming). "Evaluation de la traduction : la méthode PIE (Preselected Items Evaluation)." *Turjuman, Journal of Translation Studies*.

- ***Le Nouvel Observateur*** (2009). "Le Prof "désobéisseur," sanctionné, perd 7.000 Euros." July 24, http://tempsreel.nouvelobs.com/societe/20090724.OBS5350/le-prof-desobeisseur-sanctionne-perd-7-000-euros.html (consulted 14.11.2014).

- **Linacre, John M.** (2013) Facets computer program for many-facet Rasch measurement, version 3.71.2. www.winsteps.com (consulted 16.11.2014).

- — (n.d.). "Misfit Diagnosis: Infit Outfit Mean-square Standardized." Winsteps Rasch Measurement Software (consulted 16.11.2014)

- **Linacre, John M.** (2011) Personal communication (consulted 16.11.2014).

- **Lunz, Mary E.**, and John A. Stahl. (1990) "Judge Consistency and Severity Across Grading Periods." *www.sagepublications.com.* (consulted 15.11.2014).

- **Lommel, Arle** (ed.) (2014). "Multidimensional Quality Metrics (MQM) Definition." http://www.qt21.eu/mqm-definition/definition-2014-08-19.html (consulted 16.11.2014)

- **Nord, Christiane** (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

- **O'Brien, Sharon** (2012). "Towards a Dynamic Quality Evaluation Model for Translation." *The Journal of Specialised Translation* 17, 55-77. http://jostrans.org/issue17/art_obrien.php (consulted 16.11.2014).

- **Shrock, Sharon A. and William C. C. Coscarelli** (2007). *Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training*. San Francisco: Pfeiffer.

- **Snow, Tyler** (forthcoming) "Establishing the Viability of the Multidimensional Quality Metrics Framework." Thesis. Brigham Young University.

- **The Corpus of Contemporary American English: 450 million words, 1990-present**. http://corpus.byu.edu/coca/ (consulted 05.12.2014).

- **The Length of a Logit**. http://www.rasch.org/rmt/rmt32b.htm (consulted 05.12.2014).

- **Vermeer, Hans J.** (1987). "What does it mean to translate?", *Indian Journal of Applied Linguistics* 13(2), 25-33.

- **Waddington, Christopher** (2001). "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta: Translators' Journal* 46(2), 311-25.

- **Zhu, W., C.D. Ennis and A. Chen** (1998). "Many-faceted Rasch modeling expert judgment in test development." *Measurement in Physical Education and Exercise Science* 2(1), 21-39.

Websites

- **ATA, American Translators Association** (2014). *Framework for Standardized Error Marking* http://www.atanet.org/certification/aboutexams_error.php (consulted 16.11.2014)

- **ATA, American Translators Association** (2011). *ATA Certification Program Rubric for Grading*. http://www.atanet.org/certification/aboutexams_rubic.pdf (consulted 16.11.2014)

- **Centre for the Study of Higher Education** (2002). *A Comparison of Norm-referencing and Criterion-referencing Methods for Determining Student Grades in Higher Education*. http://www.cshe.unimelb.edu.au/assessinglearning/05/normvcrit.html (consulted 16.11.2014)

- **Institute for Objective Measurement,** The length of a logit. http://www.rasch.org/rmt/rmt32b.htm (consulted 05.12.2014).

- **MQM** (2014). *Multidimensional Quality Metrics*. http://www.qt21.eu/mqm-definition (consulted 05.12.2014).

- **National Council of State Boards of Nursing (NCSBN)**. "What is a Logit?"https://www.ncsbn.org/What_is_a_Logit.pdf (consulted 16.11.2014)

- **QTLP** (2014). *Quality Translation Launch Pad.* http://www.qt21.eu/launchpad/content/new-goal-quality-translation (consulted 05.12.2014).
  —**,** definition. http://www.qt21.eu/mqm-definition/ (consulted 05.12.2014).
  —**,** section 8. http://www.qt21.eu/mqm-definition/definition-2014-08-19.html#scoring
  —**,** section 10. http://www.qt21.eu/mqm-definition/definition-2014-08-19.html#mappings (consulted 05.12.2014).

—, metric builder, http://scorecard2.gevterm.net/ (consulted 05.12.2014).

Biographies

Alan Melby is an Emeritus Professor of Linguistics at Brigham Young University and a consultant to the QTLP project. He has over 40 years of experience in the translation world, having worked on machine translation, tools for human translators, philosophy of language applied to translation, and translation quality management.

E-mail: akmtrg@byu.edu

Troy Cox has a BA in Linguistics and an MA in Teaching English as a Second Language, and a PhD in Instructional Technology and Technology from Brigham Young University. He has worked at the English Language Center since 1995 where he has taught classes, managed the technology and directed school-wide testing. He recently took a faculty position in the College of Humanities.

E-mail: troy_cox@byu.edu

Valerie Mariana is a Masters Student pursuing her degree in French Studies at Brigham Young University. She currently works as a project manager in the translation industry.

E-mail: valerie.mariana@byu.net